Inhaltsverzeichnis

English version: `https://www.uibk.ac.at/econometrics/einf/kap02_intro_en.pdf`

Kapitel 2

Grundlagen der deskriptiven Regressionsanalyse – OLS Mechanik

"Physics is like sex. Sure, it may give some practical results, but that's not why we do it." (Richard Feynman)

2.1 Vorbemerkungen

Die Statistik beschäftigt sich ganz allgemein mit Methoden zur Erhebung und Auswertung von quantitativen Informationen. Dabei unterscheidet man traditionell zwischen deskriptiver und induktiver Statistik. Während das Ziel der deskriptiven Statistik häufig eine *Informationsverdichtung* gegebener Daten ist, beschäftigt sich die induktive Statistik hauptsächlich mit möglichen Schlussfolgerungen von einer beobachteten Stichprobe auf eine nicht beobachtbare Grundgesamtheit.

Auch die Regressionsanalyse kann für beide Zwecke eingesetzt werden. Obwohl sie in der Ökonometrie fast ausschließlich im Sinne der induktiven Statistik verwendet wird, beginnen wir hier mit der deskriptiven Regressionsanalyse. Der Grund dafür ist vor allem didaktischer Natur, dies erlaubt uns die eher technischen Aspekte von den etwas abstrakteren Konzepten der stochastischen Regressionsanalyse zu trennen; dies soll einen möglichst einfachen Einstieg in die Materie ermöglichen.

Wir werden argumentieren, dass die deskriptive Regressionsanalyse mehr oder weniger als eine Verallgemeinerung der Methode zur Berechnung einfacher Mittelwerte angesehen werden kann. Darüber hinaus erlaubt uns die Regressionsanalyse aber zusätzlich den Zusammenhang zwischen zwei oder mehreren Variablen kompakt darzustellen.

Genau darum wird es in diesem Kapitel gehen, nach ein paar allgemeinen Überlegungen werden wir die Technik kennen lernen, mit deren Hilfe wir die Koeffizienten einer linearen Regression berechnen können. Darauf aufbauend werden wir uns mit der Interpretation der Ergebnisse befassen, bevor wir die Technik auf mehr als zwei Variablen verallgemeinern und ein paar wichtige Spezialfälle untersuchen.

Alle späteren Kapitel bauen unmittelbar auf diesen einfachen Konzepten auf, deshalb lohnt es sich diese Grundlagen etwas genauer anzuschauen.

[0]

2.2 Lineare Zusammenhänge

> *"Von nichts sind wir stärker überzeugt als von dem, worüber wir am wenigsten Bescheid wissen"*
> (Michel de Montaigne, 1533–1592)

Eine der zentralen Aufgaben der Ökonometrie besteht in der *'Messung von Zusammenhängen'*. Dazu müssen die interessierenden Zusammenhänge zuerst formal dargestellt werden. Dies geschieht mit Hilfe von mathematischen Funktionen.

Eine *Funktion* $y = f(x)$ ist im wesentlichen eine 'Input-Output' Beziehung, sie liefert den Wert einer *abhängigen* Variable y für gegebene Werte der erklärenden Variable x, oder im Fall mehrerer erklärender Variablen $y = f(x_1, x_2, \ldots, x_k)$, wobei f die Funktionsform und k die Anzahl erklärender Variablen bezeichnet.

Wir werden uns vorerst auf den allereinfachsten Fall beschränken, auf lineare Funktionen mit nur einer erklärenden Variable x.

$$y = b_1 + b_2 x$$

Dabei stehen b_1 und b_2 für einfache Zahlen, die den linearen Zusammenhang zwischen den Variablen x und y beschreiben.

Wenn wir diese Funktion in ein Koordinatensystem einzeichnen erhält man eine gerade Linie. Das *Interzept* b_1 gibt dabei den Schnittpunkt mit der vertikalen y-Achse (Ordinate) an, d.h. es misst den Wert von y an der Stelle $x = 0$. Der Koeffizient b_2 der erklärenden x Variable misst die Steigung der Geraden, und wird deshalb wenig überraschend *Steigungskoeffizient* (*'slope'*) genannt. Für lineare Funktionen ist der Steigungskoeffizient b_2 gleich der Ableitung

$$\frac{dy}{dx} = b_2$$

und gibt an, um wie viele Einheiten sich y ändert, wenn x um eine Einheit zunimmt. Deshalb misst b_2 den *marginalen Effekt* einer Änderung von x auf y.

Aufgrund der linearen Funktionsform gilt dies hier nicht nur für infinitesimale Änderungen, sondern auch für diskrete Änderungen, d.h. wenn x um eine Einheit zunimmt ($\Delta x = 1$), dann *ändert* sich y um b_2 Einheiten ($\Delta y = b_2$, bzw., $\Delta y / \Delta x = b_2$). Ein wesentlicher Vorteil der linearen Funktionsform ist, dass der *marginale Effekt* b_2 nicht von x abhängt, d.h., dass er für *alle* Ausprägungen von x den gleichen Wert annimmt.

2.2.1 Exakte und 'ungefähre' Zusammenhänge

Auch wenn derart einfache lineare Zusammenhänge zunächst wie eine Karikatur einer komplexen Realität anmuten, kommen diese im täglichen Leben häufig vor.

Wenn wir zum Beispiel ein Auto tanken wissen wir, dass sich der zu bezahlende Betrag als Produkt von Preis und der Anzahl der getankten Liter ergibt. Wenn wir den zu bezahlenden Betrag mit y und die Anzahl der getankten Liter mit x

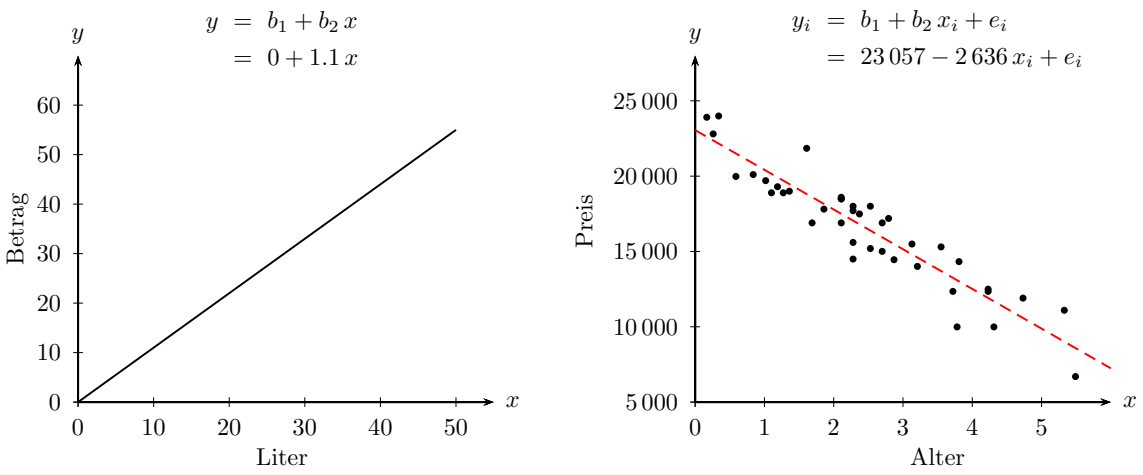


Abbildung 2.1: Linkes Panel: ein exakter Zusammenhang zwischen getankten Litern und zu bezahlendem Betrag für einen Preis $b_2 = 1.1$ Euro. Rechtes Panel: ein 'ungefährer' Zusammenhang zwischen dem Alter von Gebrauchtautos und deren Preis.

bezeichnen wird der Zusammenhang zwischen x und y durch die Funktion $y = b_1 + b_2\, x$ (für $x \geq 0$) exakt beschrieben.

Dabei bezeichnet der Steigungskoeffizient b_2 den Preis, das heißt, wenn wir einen *zusätzlichen* Liter tanken steigt der zu bezahlende Betrag um b_2 Euro. Vom Interzept b_1 wissen wir, dass es in diesem Beispiel gleich Null sein muss, denn wenn wir Null Liter tanken ($x = 0$) müssen wir auch nichts bezahlen ($y = 0$), die Funktion beginnt also im Nullpunkt. Diese Funktion ist im linken Panel von Abbildung 2.1 für einen Preis $b_2 = 1.1$ grafisch dargestellt.

Das rechte Panel von Abbildung 2.1 zeigt einen anderen Zusammenhang, den Zusammenhang zwischen dem Alter von Gebrauchtautos einer bestimmten Type und deren Preis. Jeder Punkt zeigt Alter und Preis für ein spezifisches Gebrauchtauto, insgesamt stellen die 40 Punkte Alter und Preise von 40 verschiedenen Autos dar (die zugrunde liegenden Daten sind in Tabelle 2.1 widergegeben). Offensichtlich sinkt der 'durchschnittliche' Preis mit dem Alter, aber der Zusammenhang gilt nicht länger exakt.

Dies hat verschiedene Ursachen, zum einen unterscheiden sich die Autos in anderen hier nicht dargestellten Charakteristika (Kilometerstand, Ausrüstung, Farbe, ...), aber auch Verkäufer und deren Motive, der Ort und vieles mehr unterscheidet sich von Beobachtung zu Beobachtung.

Trotzdem ist klar erkennbar, dass ältere Autos *'im Durchschnitt'* billiger sind, und dass dieser Zusammenhang durch die strichliert eingezeichnete Gerade relativ gut *approximiert* werden kann.

Wie können wir solche 'approximative' Zusammenhänge allgemein anschreiben? Wir könnten unter Verwendung des '$\approx$' Zeichens (*'ist ungefähr'*) schreiben $y \approx b_1 + b_2 x$, aber mit '$\approx$' ist schlecht Rechnen. Deshalb benötigen wir eine geeignetere Darstellungsform. Die Lösung ist einfach, wir führen einen 'Rest' ein, sogenannte *'Residuen'* (*'residuals'*), die alle anderen (unbeobachteten) Einflussfaktoren erfassen sollen. Für diese Residuen verwenden wir das Symbol e.

Tabelle 2.1: Preise (in Euro) und Alter (in Jahren) von 40 Gebrauchtautos (AlterJ ist das Alter gerundet auf ganze Jahre);
`http://www.hsto.info/econometrics/data/auto40.csv`

Obs.	Preis	Alter	AlterJ	km
1	10000	3.78	4	188000
2	21850	1.61	2	25900
3	14500	2.28	2	83300
4	11100	5.33	5	120300
5	6700	5.49	5	142000
6	24000	0.34	0	5500
7	10000	4.31	4	100500
8	16900	1.69	2	31000
9	18000	2.53	3	23000
10	15300	3.55	4	73000
11	19980	0.59	1	1500
12	15600	2.28	2	21700
13	17200	2.79	3	27570
14	18890	1.10	1	13181
15	23900	0.17	0	1800
16	14320	3.81	4	67210
17	11900	4.73	5	73900
18	15200	2.53	3	27000
19	14450	2.87	3	90000
20	18600	2.11	2	27000

Obs.	Preis	Alter	AlterJ	km
21	15000	2.70	3	51500
22	18500	2.11	2	25880
23	18500	2.11	2	19230
24	12350	3.72	4	75000
25	16900	2.70	3	22000
26	18000	2.28	2	35000
27	18890	1.27	1	22500
28	20100	0.84	1	18000
29	19700	1.02	1	12600
30	17500	2.37	2	35900
31	19300	1.19	1	5000
32	15500	3.13	3	39000
33	14000	3.21	3	56400
34	16900	2.11	2	55000
35	17700	2.28	2	25100
36	12500	4.23	4	59200
37	19000	1.36	1	19000
38	22800	0.26	0	5000
39	12350	4.23	4	73000
40	17800	1.86	2	35000

Diese Residuen e werden sich natürlich von Beobachtung zu Beobachtung (d.h. hier von Auto zu Auto) unterscheiden, deshalb benötigen wir für jede Beobachtung eine eigene Gleichung

$$\begin{aligned} y_1 &= b_1 + b_2 x_1 + e_1 \\ y_2 &= b_1 + b_2 x_2 + e_2 \\ \vdots & \qquad \vdots \\ y_n &= b_1 + b_2 x_n + e_n \end{aligned}$$

wobei n die Anzahl der Beobachtungen bezeichnet.

Da dies etwas umständlich zu schreiben wäre wird dies meist in der folgenden Form kürzer notiert

$$y_i = b_1 + b_2 x_i + e_i, \qquad \text{mit } i = 1, 2, \ldots, n \tag{2.1}$$

wobei i den Laufindex und n die Anzahl der Beobachtungen bezeichnet. Manchmal schreibt man auch $i \in \mathbb{N}$, d.h., der Index i ist ein Element der natürlichen Zahlen $\mathbb{N}$.

Das Residuum e_i nimmt dabei jeweils den Wert an, der notwendig ist, damit Gleichung i exakt erfüllt ist. Wenn man obige Gleichung umschreibt zu $e_i = y_i - b_1 - b_2 x_i$ erkennt man, dass es einen unmittelbaren Zusammenhang zwischen den Residuen e_i und den Koeffizienten b_1 und b_2 gibt.

An dieser Stelle sind zwei wichtige Hinweise angebracht:

1. nur die Ausprägungen der Variablen y_i und x_i sind beobachtbar (in unserem Beispiel also Preis und Alter der Gebrauchtautos), die Koeffizienten b_1 und b_2 sowie die Residuen e_i sind *nicht* direkt beobachtbar.

2. nur die Ausprägungen der Variablen y_i, x_i sowie der Residuen e_i unterscheiden sich zwischen den einzelnen Beobachtungen, die Koeffizienten b_1 und b_2 sollen für alle Beobachtungen gelten, sie sind also *nicht* beobachtungsspezifisch, oder in anderen Worten, wir nehmen sie als *konstant* an. Wir können uns vorstellen, dass die Koeffizienten b_1 und b_2 der linearen Funktion gewissermaßen den hinter den Daten liegenden Zusammenhang beschreiben. Ob ein Wert beobachtungsspezifisch ist oder nicht kann man häufig am Subindex i erkennen, nur beobachtungsspezifische Werte weisen einen Subindex i auf.[1]

   Man beachte, dass wir nicht behauptet haben die Koeffizienten b_1 und b_2 seien *'in Wirklichkeit'* konstant, dabei handelt es sich um eine *Annahme*, die uns überhaupt erst die Berechnung der beiden unbekannten Koeffizienten aus den Daten erlauben wird.

Man beachte, dass die obigen Gleichungen ziemlich inhaltslos wären, wenn man keine Restriktionen für die Residuen festlegt. Ohne diese Restriktionen wäre jeder Wert von b_1 und b_2 mit den beobachteten Werten von y_i und x_i kompatibel! Daher spielen

[1] Vorsicht, die Indizes 1 und 2 der Koeffizienten b_1 und b_2 haben eine andere Bedeutung als der Index i.

die Restriktionen für die Residuen, die notwendig sind, um die 'bestmögliche' Anpassung zu erreichen, eine wesentliche Rolle in allen folgenden Ausführungen und in der Ökonometrie im Allgemeinen. Wir werden zeigen, dass vor allem Restriktionen erforderlich sind, dass der Mittelwert über alle Residuen Null ist, und dass die Korrelation zwischen den Residuen und der x-Variable Null ist. Dazu später mehr.

Im Folgenden wird es darum gehen, wie wir aus den beobachteten Daten y_i und x_i mit $i = 1, \ldots, n$ die beiden Koeffizienten b_1 und b_2 der linearen Funktion $y_i = b_1 + b_2 x_i + e_i$ berechnen können, weil uns dies eine sehr kompakte Beschreibung der Daten im Sinne der deskriptiven Statistik ermöglicht, ähnlich wie der Mittelwert eine kompakte Zusammenfassung einer einzelnen Datenreihe liefert.

Im Autobeispiel approximiert die Geradengleichung die Beobachtungen relativ gut, aber es ist auch klar, dass diese Approximation nur für einen bestimmten Bereich der x zufriedenstellende Resultate liefert. Für ein 10 Jahre altes Autos würde die Regressionsgerade z.B. einen negativen Preis liefern. Preissteigerungen für Oldtimer können durch diese Gerade selbstverständlich überhaupt nicht abgebildet werden. Das bedeutet, dass der Zusammenhang zwischen Alter und Preis eigentlich nicht linear ist.

Aber wie dieses Beispiel zeigt können selbst nicht lineare Zusammenhänge oft *über einen begrenzten Bereich* der Variablen durch eine lineare Funktion manchmal relativ gut approximiert werden.

Interzept und Regressionskonstante Wir haben bisher sowohl b_1 als auch b_2 als Koeffizienten bezeichnet, obwohl b_1 zumindest nicht 'sichtbar' mit einer Variablen multipliziert wird. Wir können uns aber vorstellen, dass b_1 mit einem Einsenvektor multipliziert wird, wie dies in der folgenden Vektordarstellung deutlich wird

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = b_1 \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} + b_2 \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

Der Einsenvektor wird in diesem Zusammenhang häufig 'Regressionskonstante' genannt, und das Interzept b_1 ist einfach der Koeffizient der Regressionskonstanten.[2]

Alternative Bezeichnungen für y und x Wenn man eine Regressionsgleichung $y_i = b_1 + b_2 x_i + e_i$ schätzt sagt man auch, y wird auf x regressiert. Für die Variablen y und x haben sich in der Literatur eine ganze Reihe verschiedener Bezeichnungen eingebürgert, einige davon sind in Tabelle 2.2 zusammengefasst.

Wir werden im Folgenden y meist als *abhängige Variable* und x als *erklärende Variable* bezeichnen. Man sollte dabei den Begriff 'erklärend' dabei nicht allzu wörtlich nehmen, denn dies muss nicht bedeuten, dass y durch x 'erklärt' wird; mit dieser Methode können wir bestenfalls zeigen, dass zwischen y und x ein linearer Zusammenhang besteht, aber die Methode alleine liefert uns keinesfalls eine inhaltliche

[2]Die Literatur ist in dieser Hinsicht leider manchmal etwas verwirrend, in manchen älteren Lehrbüchern werden die Bezeichnungen 'Interzept' und 'Regressionskonstante' auch synonym verwendet.

Tabelle 2.2: Alternative Bezeichnungen für y und x der Funktion $y = b_1 + b_2 x$

y	x
– links-stehende Variable (*'left-hand side variable'*)	rechts-stehende Variable (*'right-hand side variable'*)
– abhängige Variable (*'dependent variable'*)	[unabhängige Variable] (*'independent variable'*)
– erklärte Variable (*'explained variable'*)	erklärende Variable (*'explanatory variable'*)
– Regressand (*'regressand'*)	Regressor (*'regressor'*)
– Antwortvariable (*'response variable'*)	Kovariable (*'covariate'*)
– Effektvariable (*'effect variable'*)	Kontrollvariable (*'control variable'*)

'Erklärung' für diesen Zusammenhang, und natürlich erst recht keine Hinweise auf eine mögliche Kausalbeziehung zwischen y und x. Wir werden im Folgenden aber trotzdem bei den Bezeichnungen *abhängige* und *erklärende* Variable bleiben, weil sie sich in der Literatur eingebürgert haben.

Die erklärenden x Variablen werden häufig auch Regressoren genannt, während die Bezeichnung Regressand für y nicht ganz so gebräuchlich ist.

Vor allem in der Statistik werden die erklärenden Variablen häufig *Kovariate* genannt, in eher technischen Zusammenhängen ist auch die Bezeichnung *Kontrollvariablen* für die x Variablen gebräuchlich.

In älteren Lehrbüchern findet sich für die x Variable auch noch öfter die Bezeichnung 'unabhängige Variable' (*'independent variable'*). Während die Bezeichnung 'abhängige Variable' für y durchaus zutreffend und üblich ist, kann die Bezeichnung 'unabhängige Variable' für x irreführend sein, da dies mit 'statistischer Unabhängigkeit' verwechselt werden könnte, was ein völlig anders Konzept ist. Deshalb wird generell von der Bezeichnung von x als unabhängige Variable abgeraten.

Im nächsten Abschnitt werden wir nun eine Methode kennen lernen, die es uns erlaubt aus den beobachteten Werten der Variablen x und y die Koeffizienten b_1 und b_2 derart zu berechnen, dass der Zusammenhang zwischen x und y 'möglichst gut' beschrieben wird.

2.3 Die OLS Methode

"Wer hohe Türme bauen will, muß lange beim Fundament verweilen."
(Anton Bruckner, 1824–1896)

Die Bezeichnung OLS steht für *'Ordinary Least Squares'*, auf deutsch **Methode der (Gewöhnlichen) Kleinsten Quadrate**. Wir werden hier meist das englischen

Akronym OLS verwenden, da sich dies mittlerweile auch in der deutschsprachigen Literatur eingebürgert hat.

Unser konkretes Anliegen in diesem Abschnitt ist es eine Formel zu finden, in die wir die beobachteten Daten y und x einsetzen können, und die uns als Resultat *'bestmögliche'* Zahlenwerte für die nicht direkt beobachtbaren Koeffizienten b_1 und b_2 einer Geradengleichung $y_i = b_1 + b_2 x_i + e_i$ liefert. Was genau unter 'bestmöglich' zu verstehen ist werden wir später erläutern, aber wir werden sehen, dass die OLS Methode genau dieses Problem löst.

Wir beginnen unsere Überlegungen mit einer gedanklichen Zerlegung der abhängigen Variable y_i in zwei Teile, in eine *systematische Komponente* $b_1 + b_2 x_i$, in der die den Daten zugrunde liegende Zusammenhang in Form einer Geradengleichung zum Ausdruck kommt, und in den Rest, d.h. die *Residuen* e_i

$$y_i = \underbrace{b_1 + b_2 x_i}_{\substack{\text{systematische} \\ \text{Komponente } \widehat{y}_i}} + \underbrace{e_i}_{\substack{\text{Resi-} \\ \text{duen}}}$$

Wir wollen uns diese Zerlegung anhand von Abbildung 2.2 veranschaulichen. Das obere Panel zeigt 5 Datenpunkte und eine gedachte Gerade, die sich an diese Beobachtungspunkte 'bestmöglich' anpasst. Diese Gerade werden wir in Zukunft *'Regressionsgerade'* nennen. Angenommen, wir hätten diese Regressionsgerade bereits, dann könnten wir diese nützen, um jedes beobachtete y_i in zwei Teile zu zerlegen, in einen Wert, der exakt *auf* der Regressionsgeraden liegt, $\widehat{y}_i$ (gesprochen y_i Dach), und in die Differenz zwischen diesem auf der Regressionsgerade liegenden $\widehat{y}_i$ und dem tatsächlich beobachteten Wert y_i. Diese Differenz ist natürlich das Residuum e_i, also $y_i = \widehat{y}_i + e_i$ (mit $\widehat{y}_i = b_1 + b_2 x_i$) für $i = 1, \ldots, n$. Das untere Panel in Abbildung 2.2 zeigt diese Zerlegung.

Die exakt *auf* der Regressionsgerade liegenden 'gefitteten' Werte $\widehat{y}_i$ nennen wir *systematische Komponente*, das heißt, die durch die Variable x und die Koeffizienten b_1 und b_2 beschriebene Komponente.

Aber für die Berechnung dieser 'gefitteten' Werte $\widehat{y}_i$ benötigen wir neben der x Variable die (vorerst noch) unbekannten Koeffizienten b_1 und b_2

$$\widehat{y}_i = b_1 + b_2 x_i$$

Eine 'gute' Regressionsgerade sollte zwei Bedingungen erfüllen:

1. der Anteil der 'systematischen' Komponente sollte möglichst groß sein, was impliziert, dass die Residuen einen möglichst kleinen Erklärungsbeitrag liefern sollten;

2. dies erfordert, dass die Korrelation zwischen 'systematischer' Komponente und den Residuen möglichst klein sein muss. Wir werden etwas später zeigen, dass uns die OLS Methode genau solche Werte für b_1 und b_2 liefert, die garantieren, sodass die Korrelation zwischen der 'systematischen' Komponente und den Residuen exakt gleich Null ist.

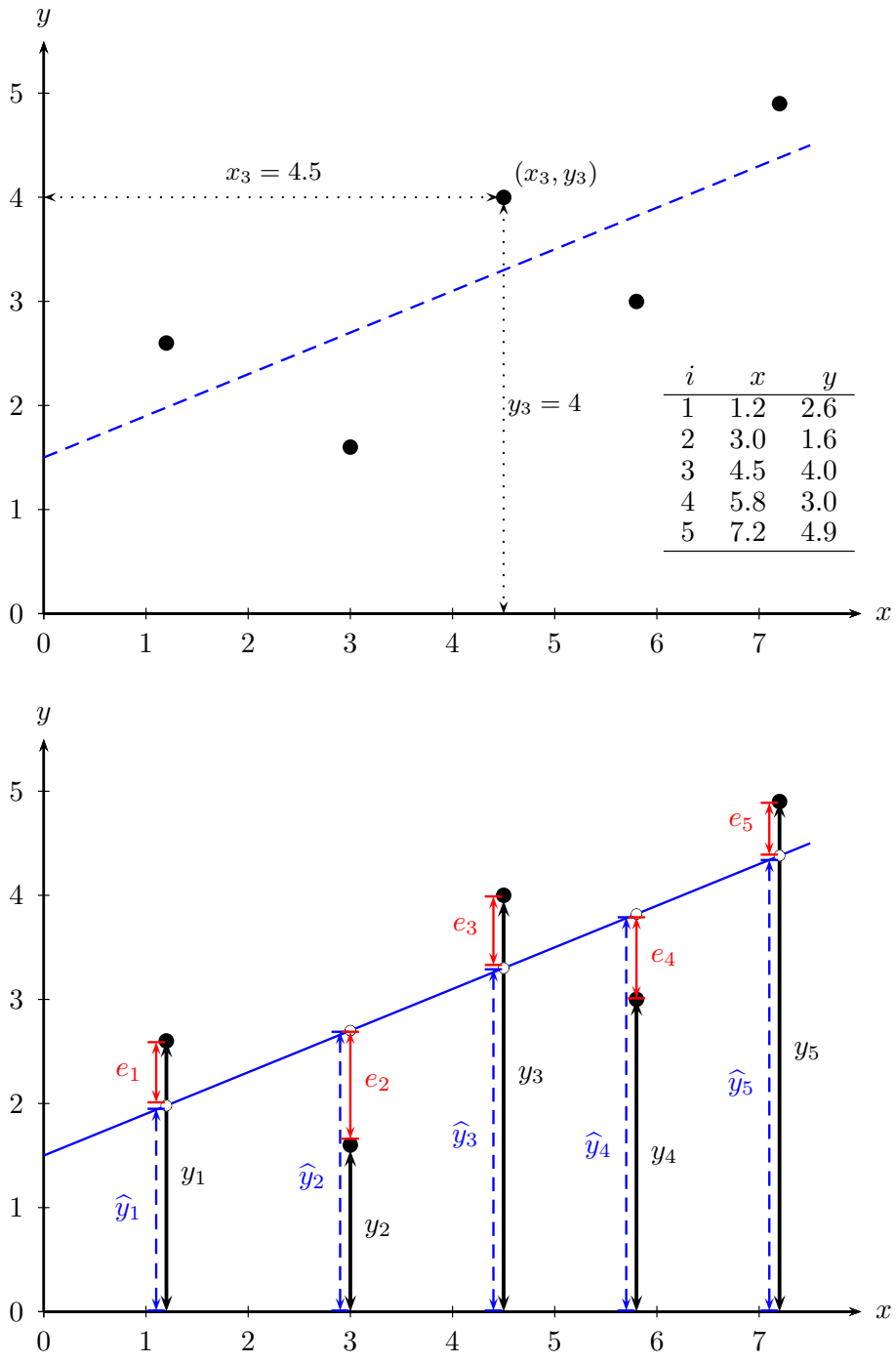


Abbildung 2.2: Zerlegung von y_i in eine systematische Komponente $\widehat{y}_i$ und in ein unsystematisches Residuum e_i (für $i = 1, \ldots, 5$). [local, www]

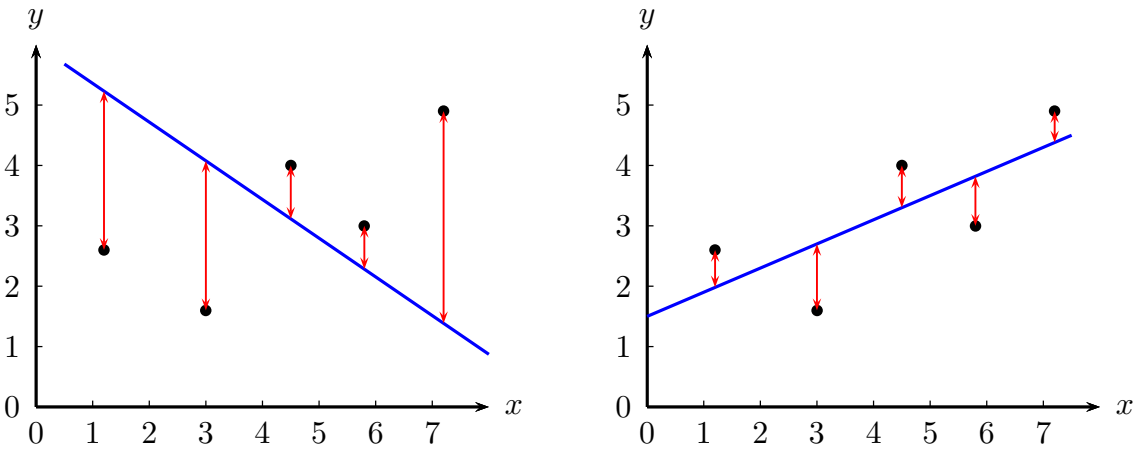


Abbildung 2.3: Die Summe der Abweichungen $\sum_i e_i = \sum_i (y_i - \widehat{y}_i)$ hat in beiden Abbildungen den gleichen Wert, da sich positive und negative Werte aufheben.

Zur tatsächlichen Berechnung der Koeffizienten könnte man auf die Idee kommen die Werte b_1 und b_2 derart zu wählen, dass die Summe aller Residuen $\sum_i e_i$ möglichst klein wird.

Dies würde allerdings dazu führen, dass sich positive und negative Abweichungen beim Summieren aufheben. Man kann sogar einfach zeigen, dass die Summe der Residuen für jede Gerade Null ist, die durch die Mittelwerte von x und y gelegt wird. Deshalb ist diese Methode ungeeignet um eine gute Approximation zu erhalten.

Abbildung 2.3 veranschaulicht das Problem: die *Summe der Abweichungen* $\sum_i e_i$ hat in der linken und rechten Grafik den gleichen Wert, obwohl die Gerade in der rechten Grafik die Punkte offensichtlich weit besser approximiert.

Dieses Problem könnte man vermeiden, wenn man den absoluten Wert der Abweichungen minimiert. Dies wirft jedoch zwei Probleme auf: Zum einen ist dieses Problem numerisch deutlich schwieriger zu lösen, zum anderen werden damit große Abweichungen nicht überproportional stärker gewichtet als kleine Abweichungen. Tatsächlich sind Menschen häufig risikoavers und werden große Fehler lieber überproportional stärker 'bestraft' sehen als kleine Fehler.

Die einfachste Lösung für diese Probleme besteht darin, die Koeffizienten b_1 und b_2 derart zu wählen, dass die *Summe der quadrierten Abweichungen* (d.h. $\sum_i e_i^2$) minimiert wird. Genau dies ist das Prinzip der OLS Methode.

Daraus erklärt sich auch der Name **Methode der (Gewöhnlichen) Kleinsten Quadrate** (*'Ordinary Least Squares'*, OLS).

Diese ziemlich einfache Grundidee der OLS Methode kann mit Hilfe von Abbildung 2.4 einfach erklärt werden. Man beachte, dass die Funktion $y_i = b_1 + b_2 x_i + e_i := \widehat{y}_i + e_i$ umgeschrieben werden kann zu $e_i = y_i - \widehat{y}_i$. In Abbildung 2.4 sind die Quadrate der Residuen $e_i^2 = (y_i - \widehat{y}_i)^2 := (y_i - b_1 - b_2 x_i)^2$ eingezeichnet. In einem Gedankenexperiment können wir die Gerade dieser Abbildung solange drehen und verschieben, dass heißt die Werte von b_1 und b_2 verändern, bis die *Summe* der eingezeichneten Quadratflächen so klein wie möglich wird. Die Werte von b_1 und b_2,

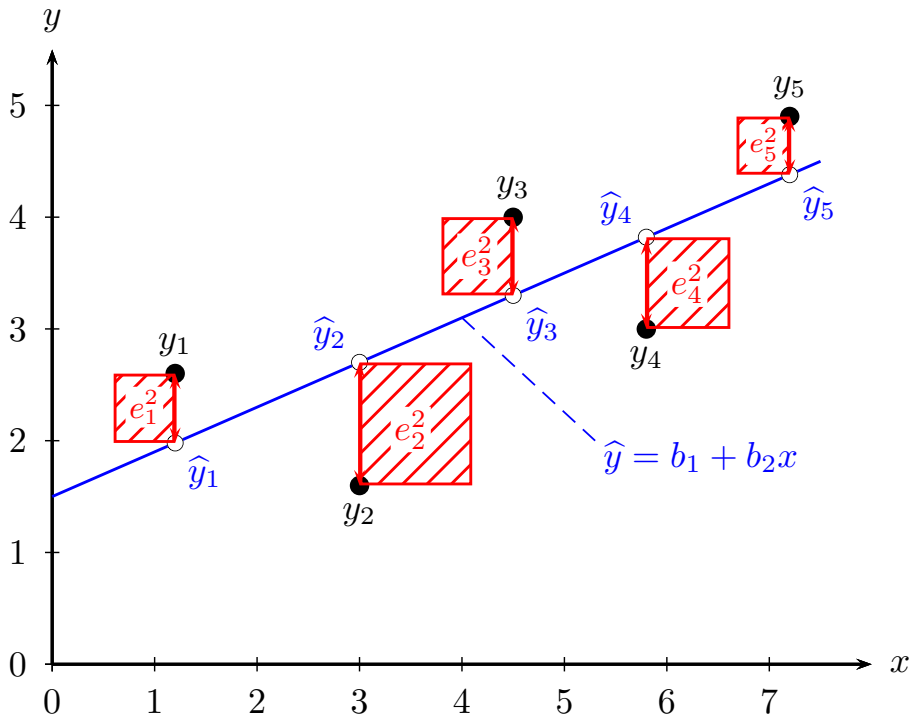


Abbildung 2.4: Nach der OLS Methode werden b_1 und b_2 derart gewählt, dass die *Summe der quadrierten Abweichungen* möglichst klein wird, d.h., die Gesamtfläche der schraffierten Quadrate wird minimiert. [local,www]

die die *kleinst mögliche Summe der Quadratflächen* liefern, sind die gesuchten OLS Koeffizienten.

Dieses Gedankenexperiment liefert eine gute Intuition, aber diese Vorgangsweise eignet sich kaum für das praktische Arbeiten. Wir benötigen eine allgemeine Methode, die uns erlaubt die unbeobachtbaren Koeffizienten b_1 und b_2 aus den beobachtbaren Daten x und y zu berechnen, und eine solche Formel werden wir nun herleiten.

Bevor wir damit beginnen noch eine kurze Anmerkung. Sie werden sich vielleicht fragen, wozu diese ganze nun folgende 'Rechnerei' gut sein soll, wenn die fertigen Formeln selbst in Excel bereits fix und fertig implementiert und denkbar einfach anzuwenden sind. Nun, wir werden in den folgenden Kapiteln sehen, dass die Anwendung dieser Formel nur unter ganz bestimmten Voraussetzungen zu den gewünschten Ergebnissen führt. Ein Verständnis der Mechanik der OLS-Methode wird es uns erlauben auch die Grenzen dieses Ansatzes zu verstehen, und in einem weiteren Schritt geeignete Maßnahmen zu ergreifen, wenn die Annahmen verletzt sind. Eine naive Anwendung dieser Methoden führt häufig zu irreführenden oder zumindest unnotwendig ungenauen Ergebnissen. Um solche Fehler zu vermeiden ist ein fundiertes Verständnis der Grundlagen erforderlich, und für ein solches Verständnis ist ein bisschen Rechnerei manchmal erstaunlich nützlich.

Den Zusammenhang zwischen der Fläche eines Quadrates und den beiden Koeffizienten b_1 und b_2 können wir durch umschreiben von $y_i = b_1 + b_2 x_i + e_i$ einfach darstellen

$$e_i = y_i - b_1 - b_2 x_i$$

Die Fläche eines einzelnen schraffierten Quadrates in Abbildung 2.4 ist $e_i^2 = (y_i - b_1 - b_2 x_i)^2$, und die Fläche *aller* Quadrate ist einfach die Summe über $i = 1, \ldots, n$

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_1 - b_2 x_i)^2$$

Gesucht sind die Werte von b_1 und b_2, für die die *Summe aller Flächen* – also die Quadratsumme der Residuen $\sum_i e_i^2$ – minimal ist, das Minimierungsproblem lautet also

$$\min_{b_1, b_2} \sum_{i=1}^{n} e_i^2 = \min_{b_1, b_2} \sum_{i=1}^{n} (y_i - b_1 - b_2 x_i)^2$$

wobei das b_1 und b_2 unter der 'min' Anweisung darauf hinweisen sollen, dass dies die zwei gesuchten Größen sind.

Der Rest ist simple Rechnerei. Wir leiten partiell nach den unbekannten Koeffizienten b_1 und b_2 ab, setzen diese beiden Ableitungen gleich Null. Dies liefert die Bedingungen erster Ordnung, bzw. notwendige Bedingungen für ein Minimum.[3] Die

[3] Man kann zeigen, dass die Bedingungen zweiter Ordnung, d.h. die hinreichenden Bedingungen, ebenfalls erfüllt sind.

Ableitungen sind[4]

$$\frac{\partial \sum_i e_i^2}{\partial b_1} = 2\sum_i \underbrace{(y_i - b_1 - b_2 x_i)}_{e_i}(-1) = -2\sum_i e_i = 0 \tag{2.2}$$

$$\frac{\partial \sum_i e_i^2}{\partial b_2} = 2\sum_i \underbrace{(y_i - b_1 - b_2 x_i)}_{e_i}(-x_i) = -2\sum_i x_i e_i = 0 \tag{2.3}$$

Wie man sieht implizieren diese Bedingungen erster Ordnung (*'first order conditions'*, FOC)

$$\boxed{\begin{aligned} \sum_i e_i &= 0 \\ \sum_i x_i e_i &= 0 \end{aligned}} \quad \Rightarrow b_1, b_2$$

Diese zwei Bedingungen sind von größter Bedeutung, die Lösungen definieren nicht nur die beiden gesuchten Koeffizienten, sie werden uns später immer wieder begegnen, denn aus diesen beiden Bedingungen folgen die wesentlichen Eigenschaften der OLS Methode!

Die erste dieser Bedingungen erster Ordnung, $\sum_i e_i = 0$, folgt aus der Ableitung nach dem Interzept b_1, d.h. Sie gilt nur, wenn die Regressionsgleichung ein Interzept enthält. Die zweite Bedingung folgt aus der Ableitung nach dem Steigungskoeffizienten b_2 und stellt – gemeinsam mit der ersten Bedingung – sicher, dass die Kovarianz zwischen x und e Null ist.[5]

Die gesuchten Koeffizienten b_1 und b_2 sind die Lösungen des Minimierungsproblems und garantieren deshalb, dass diese zwei Bedingungen erster Ordnung erfüllt sind! Die einfache Struktur – es wird lediglich das Minimum einer quadratischen Funktion bestimmt – stellt sicher, dass die Lösung eindeutig ist.

Nun wollen wir endlich die beiden unbekannten Koeffizienten b_1 und b_2 aus den beiden Bedingungen erster Ordnung berechnen. Dazu formen wir diese etwas um, wobei wir beachten, dass wir 'Alles ohne Subindex i' vor das Summenzeichen ziehen können, und dass $\sum_i b_1 = nb_1$, weil b_1 eine Konstante ist

$$\sum_{i=1}^n y_i = nb_1 + b_2 \sum_{i=1}^n x_i \tag{2.4}$$

$$\sum_{i=1}^n y_i x_i = b_1 \sum_{i=1}^n x_i + b_2 \sum_{i=1}^n x_i^2 \tag{2.5}$$

[4]Für die Ableitungen benötigen wir die Kettenregel, d.h. wenn $y = f(z)$ und $z = g(x)$ folgt $y = f[g(x)]$ und die Ableitung ist

$$\frac{dy}{dx} = \frac{dy}{dz}\,\frac{dz}{dx}$$

[5]$\sum_i x_i e_i = \sum_i e_i(x_i - \bar{x} + \bar{x}) = \sum_i e_i(x_i - \bar{x}) + \bar{x}\sum_i e_i = \sum_i e_i(x_i - \bar{x}) = \sum_i (e_i - \bar{e})(x_i - \bar{x}) = n\,\mathrm{cov}(e, x) = 0$ da $\sum_i e_i = 0$ und $\bar{e} = 0$ wenn die Regression ein Interzept enthält (im ersten Schritt wird lediglich die Konstante $\bar{x}$ subtrahiert und addiert).

Dies sind die sogenannten Normalgleichungen, die wir nach den gesuchten Koeffizienmten b_1 und b_2 lösen.

Dazu multiplizieren wir die erste Gleichung mit $\sum x_i$ und die zweiten Gleichung mit n (man beachte, dass $\sum x_i$ eine einfache Zahl ist, mit der ganz normal gerechnet werden kann)

$$\begin{aligned} \sum_i x_i \sum_i y_i &= nb_1 \sum_i x_i + b_2 \left(\sum_i x_i \right)^2 \\ n \sum_i y_i x_i &= nb_1 \sum_i x_i + b_2 n \sum_i x_i^2 \end{aligned}$$

und subtrahieren die erste Gleichung von der zweiten

$$n \sum_i y_i x_i - \sum_i x_i \sum_i y_i = b_2 \left[n \sum_i x_i^2 - \left(\sum_i x_i \right)^2 \right]$$

woraus folgt

$$b_2 = \frac{n \sum_i y_i x_i - \sum_i x_i \sum_i y_i}{n \sum_i x_i^2 - \left(\sum_i x_i \right)^2} \tag{2.6}$$

Dies ist genau die Funktion, die wir suchen. Auf der rechten Seite kommen nur noch die beobachtbaren x_i und y_i vor. Wenn wir die Beobachtungen in diese Formel einsetzen erhalten wir als Resultat den Wert des Steigungsgkoeffizienten b_2, der die Quadratsumme der Residuen minimiert!

Sobald b_2 berechnet ist kann das Interzept b_1 einfach berechnet werden, wir dividieren beide Seiten der Normalgleichung (2.4) durch n und erhalten

$$\frac{1}{n} \sum_i y_i = b_1 + b_2 \frac{1}{n} \sum_i x_i$$

Es ist üblich den Mittelwert einer Variable mit einem Querstrich über dem Variablennamen zu bezeichnen, also z.B. $\bar{y}$ (gesprochen y quer) für den Mittelwert von y. Natürlich ist $\bar{y} := \frac{1}{n} \sum_{i=1}^{n} y_i$, wobei das Symbol ':=' als 'ist definiert' (bzw. 'definitorisch identisch') gelesen wird. Man beachte, dass die Mittelwerte nicht beobachtungsspezifisch sind, und deshalb keinen Subindex i haben.

Unter Verwendung dieser Schreibweise für die Mittelwerte erhalten wir für das Interzept

$$b_1 = \bar{y} - b_2 \bar{x} \tag{2.7}$$

Diese beiden obigen OLS-Formeln lösen unser Problem bereits, aber insbesondere die Formel für den Steigungskoeffizienten (2.6) sieht etwas 'unappetitlich' aus. Glücklicherweise kann diese Formel mit Hilfe von Varianzen und Kovarianzen deutlich einfacher dargestellt werden.

Wir erinnern uns, dass die *empirische Varianz* – ein deskriptives Streuungsmaß für gegebene Beobachtungen – sowie die *empirische Kovarianz* – ein deskriptives Maß

für den Zusammenhang zwischen zwei Variablen – definiert sind als[6]

$$\mathrm{var}^p(x) := \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad \mathrm{cov}^p(x,y) := \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Mit Hilfe dieser Definitionen können die OLS-Koeffizienten einfacher geschrieben als

$$\boxed{\begin{aligned} b_2 &= \frac{\mathrm{cov}(x,y)}{\mathrm{var}(x)} \\ b_1 &= \bar{y} - b_2\bar{x} \end{aligned}}$$

wobei die Gleichung für das Interzept aus Gleichung (2.7) übernommen wurde. Man beachte, dass dies nur für Regressionen mit Interzept gilt!

Beweis:* Um zu zeigen, dass

$$b_2 = \frac{n\sum y_i x_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} = \frac{\mathrm{cov}(x,y)}{\mathrm{var}(x)}$$

dividieren wir Zähler und Nenner des mittleren Ausdrucks von Gleichung (2.6) durch n und erhalten

$$b_2 = \frac{\sum y_i x_i - \frac{1}{n}\sum x_i \sum y_i}{\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2} = \frac{\sum y_i x_i - n\left(\frac{1}{n}\sum x_i\right)\left(\frac{1}{n}\sum y_i\right)}{\sum x_i^2 - n\left(\frac{1}{n^2}\left(\sum x_i\right)^2\right)}$$

und berücksichtigen, dass der Mittelwert von x bzw. y definiert ist als $\bar{x} := \frac{1}{n}\sum_i x_i$ bzw. $\bar{y} := \frac{1}{n}\sum_i y_i$.
Damit kann der obige Ausdruck geschrieben werden als

$$b_2 = \frac{\sum_i y_i x_i - n\bar{x}\bar{y}}{\sum_i x_i^2 - n\bar{x}^2}$$

Anschließend addieren und subtrahieren wir vom Zähler $n\bar{x}\bar{y}$ und vom Nenner $n\bar{x}^2$. Dies ergibt

$$b_2 = \frac{\sum_i y_i x_i - n\bar{x}\bar{y} - n\bar{x}\bar{y} + n\bar{x}\bar{y}}{\sum_i x_i^2 - n\bar{x}^2 - n\bar{x}^2 + n\bar{x}^2}$$

Als nächstes schreiben wir die Definition der Mittelwerte etwas um, aus $\bar{x} = \frac{1}{n}\sum_i x_i$ folgt $n\bar{x} = \sum_i x_i$ bzw. $n\bar{y} = \sum_i y_i$, und setzen dies ein

$$b_2 = \frac{\sum_i y_i x_i - \bar{x}\sum_i y_i - \bar{y}\sum_i x_i + n\bar{x}\bar{y}}{\sum_i x_i^2 - 2\bar{x}\sum_i x_i + n\bar{x}^2}$$

[6]Man beachte, dass dies die Populations-Varianz var^p ist. Dagegen ist die Stichproben-Varianz definiert als $\mathrm{var}(x) := \frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2$. Der folgende Zusammenhang gilt für beide Definitionen.

ziehen das Summenzeichen heraus

$$b_2 = \frac{\sum_i (y_i x_i - \bar{x} y_i - \bar{y} x_i + \bar{x}\bar{y})}{\sum_i (x_i^2 - 2\bar{x} x_i + \bar{x}^2)}$$

und Faktorisieren

$$b_2 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \tag{2.8}$$

Dies sieht schon deutlich einfacher aus! Noch einfacher zu merken ist die Formel, wenn wir Zähler und Nenner durch n (oder $n-1$) dividieren, denn dann erkennt man, dass Gleichung (2.6) einfacher als Verhältnis von empirischer Kovarianz zu empirischer Varianz geschrieben werden kann

$$b_2 = \frac{\frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_i (x_i - \bar{x})^2} = \frac{\mathrm{cov}(y,x)}{\mathrm{var}(x)} \tag{2.9}$$

□

Achtung: Wenn die x konstant sind (d.h. alle x_i den gleichen Zahlenwert aufweisen) ist $b_2 = \mathrm{cov}(y,x)/\mathrm{var}(x)$ nicht definiert, da für konstante x $\mathrm{var}(x) = 0$! Wir werden später sehen, dass dies ein Spezialfall von *perfekter Multikollinearität* ist (d.h. x ist ein Vielfaches der Regressionskonstante).

Rechenbeispiele

Beispiel 1: Den Abbildungen 2.2 bis 2.4 liegen folgende Daten zugrunde:

i	x	y
1	1.2	2.6
2	3.0	1.6
3	4.5	4.0
4	5.8	3.0
5	7.2	4.9

Mit Hilfe der vorhin gefundenen OLS-Formeln können wir nun die Koeffizienten b_1 und b_2 berechnen, die die Quadratsumme der Residuen minimieren.

Dazu erweitern wir die Tabelle um die Spalten xy und x^2 und bilden die jeweiligen Summen:

i	x	y	$x\,y$	x^2
1	1.2	2.6	3.1	1.4
2	3.0	1.6	4.8	9.0
3	4.5	4.0	18.0	20.3
4	5.8	3.0	17.4	33.6
5	7.2	4.9	35.3	51.8
$\sum$	21.7	16.1	78.6	116.2

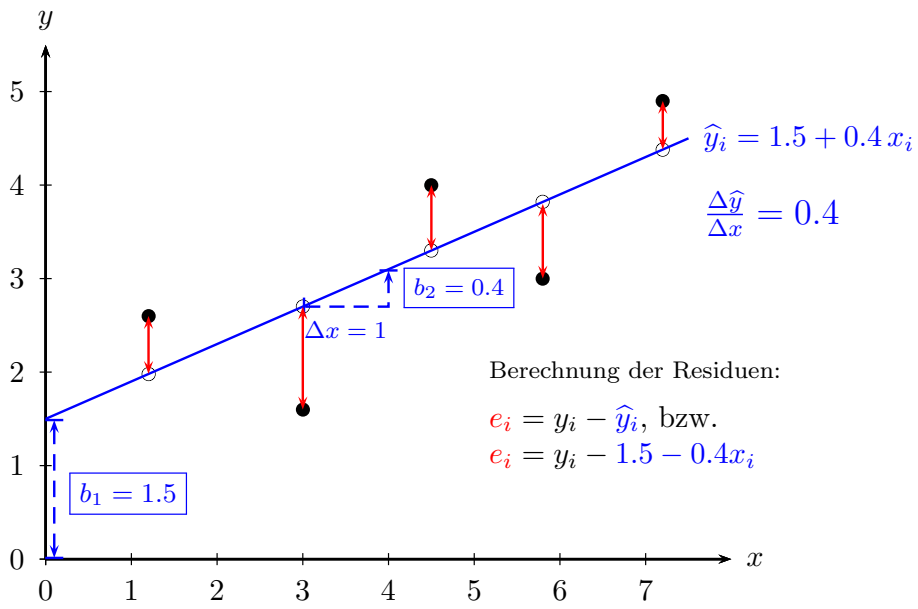


Abbildung 2.5: Beispiel
[local,www]

Wenn wir in Gleichungen (2.6) und (2.7) einsetzen erhalten wir

$$b_2 = \frac{n\sum y_i x_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} = \frac{5 \times 78.6 - 21.7 \times 16.1}{5 \times 116.2 - (21.7)^2} = 0.4$$

$$b_1 = \overline{y} - b_2\overline{x} = 16.1/5 - 0.4 \times 21.7/5 = 1.5$$

Die in Abbildung 2.5 eingezeichnete Regressionsgleichung ist also

$$\widehat{y}_i = 1.5 + 0.4x_i$$

bzw. unter Verwendung der alternativen Formel (2.8) für mittelwerttransformierte Daten

i	$x_i - \overline{x}$	$y_i - \overline{y}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	−3.1	−0.6	9.9	1.9
2	−1.3	−1.6	1.8	2.2
3	0.2	0.8	0.0	0.1
4	1.5	−0.2	2.1	−0.3
5	2.9	1.7	8.2	4.8
$\sum_i$	0.0	0.0	22.0	8.7

$$b_2 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{\mathrm{cov}(y, x)}{\mathrm{var}(x)} = \frac{8.7}{22} = 0.4$$

Beispiel 2: In diesem Beispiel zeigen wir, dass der übliche Mittelwert auch mit Hilfe der OLS-Methode berechnet werden kann, nämlich durch eine Regression auf die Regressionskonstante.

Sei

$$y_i = b_1 + e_i$$

Die Residuen sind in diesem Fall $e_i = y_i - b_1$. Die OLS-Methode beruht auf der Minimierung der Quadratsumme der Residuen, d.h.

$$\min_{b_1} \sum_i e_i^2 = \min_{b_1} \sum_i (y_i - b_1)^2$$

Ableiten nach dem unbekannten Koeffizienten b_1 und diese Ableitung Null setzen gibt den Wert von b_1, der die Quadratsumme der Residuen minimiert

$$\begin{aligned} \frac{\partial \sum_i e_i^2}{\partial b_1} &= 2\sum_i (y_i - b_1)(-1) = 0 \\ &= \sum_i y_i - \sum_i b_1 = \sum_i y_i - nb_1 = 0 \end{aligned}$$

woraus folgt

$$b_1 = \frac{1}{n}\sum_i y_i := \bar{y}$$

Eine OLS-Regression auf die Regressionskonstante liefert also tatsächlich das arithmetische Mittel, man kann also den Mittelwert als Spezialfall eines OLS-Schätzers betrachten!

Beispiel 3: Wir haben verschiedentlich angedeutet, dass die OLS Methode in einem gewissen Sinne 'optimal' ist, ohne genauer zu spezifizieren, worauf sich diese Optimalität bezieht. In diesem Übungsbeispiel werden wir zeigen, dass die nach der OLS Methode berechneten gefitteten Werte $\widehat{y}_i$ eine ganz besondere Eigenschaft haben, dass nämlich die Streuung um diese OLS gefitteten $\widehat{y}_i$ kleiner ist als die Streuung um alle anderen Werte $\widetilde{y}_i$, die mit einer beliebigen anderen *linearen* Funktion berechneten wurden.

Dies ist analog zum Mittelwert einer Variable, denn vom Mittelwert $\bar{x}$ wissen wir, dass er die Summe der quadrierten Abweichungen (bzw. die empirische Varianz) minimiert, d.h. für jede beliebige Zahl z gilt

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 < \frac{1}{n}\sum_{i=1}^{n}(x_i - z)^2 \qquad \text{für } \bar{x} \neq z$$

Warum?

$$\begin{aligned} \sum_i (x_i - z)^2 &= \sum_i (x_i - \bar{x} + \bar{x} - z)^2 \\ &= \sum_i (x_i - \bar{x})^2 + 2(\bar{x} - z)\sum_i (x_i - \bar{x}) + \sum_i (\bar{x} - z)^2 \\ &= \sum_i (x_i - \bar{x})^2 + \sum_i (\bar{x} - z)^2 \end{aligned}$$

da $\sum_i (x_i - \bar{x}) = \sum_i x_i - n\bar{x} = n\bar{x} - n\bar{x} = 0$ (beachte $\bar{x} := \frac{1}{n}\sum_i x_i \Rightarrow \sum_i x_i = n\bar{x}$).
Weil $\sum_i (\bar{x} - z)^2 > 0$ für $\bar{x} \neq z$ muss gelten $\sum_i (x_i - \bar{x})^2 < \sum_i (x_i - z)^2$.

Zeigen Sie, dass auch die nach der OLS Methode berechneten gefitteten Werte $\widehat{y}_i$ diese Eigenschaft besitzen.
Vergleichen Sie dazu die mit den OLS Koeffizienten b_1 und b_2 berechneten $\widehat{y}_i = b_1 + b_2 x_i$ mit den gefitteten Werten einer beliebigen anderen linearen Funktion $\widetilde{y}_i = c_1 + c_2 x_i$ und beweisen Sie, dass

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 < \frac{1}{n}\sum_{i=1}^{n}(y_i - \widetilde{y}_i)^2$$

Lösung: Um dies zu zeigen gehen wir analog wie oben vor

$$\begin{aligned}\sum_i (y_i - \widetilde{y}_i)^2 &= \sum_i (y_i - \widehat{y}_i + \widehat{y}_i - \widetilde{y}_i)^2 \\ &= \sum_i (y_i - \widehat{y}_i)^2 + \sum_i (\widehat{y}_i - \widetilde{y}_i)^2 + 2\sum_i (y_i - \widehat{y}_i)(\widehat{y}_i - \widetilde{y}_i)\end{aligned}$$

Die ersten beiden Terme auf der rechten Seite sind quadratisch und können deshalb nie negativ werden. Sehen wir uns deshalb zuerst den dritten Term $2\sum_i (y_i - \widehat{y}_i)(\widehat{y}_i - \widetilde{y}_i)$ an, wobei wir berücksichtigen, dass $y_i - \widehat{y}_i := e_i$ die OLS Residuen sind.
Also

$$\begin{aligned}\sum_i (y_i - \widehat{y}_i)(\widehat{y}_i - \widetilde{y}_i) &= \sum_i e_i(\widehat{y}_i - \widetilde{y}_i) \\ &= \sum_i e_i\left[(b_1 + b_2 x_i) - (c_1 + c_2 x_i)\right] \\ &= \sum_i \left[(b_1 - c_1) + (b_2 - c_2)x_i\right] e_i \\ &= (b_1 - c_1)\underbrace{\sum_i e_i}_{=0} + (b_2 - c_2)\underbrace{\sum_i x_i e_i}_{=0} \\ &= 0\end{aligned}$$

da für die OLS Residuen die beiden Bedingungen erster Ordnung $\sum_i e_i = 0$ und $\sum_i x_i e_i = 0$ gelten (siehe Gleichungen (2.2) und (2.3), Seite 15).
Es folgt also

$$\sum_i (y_i - \widetilde{y}_i)^2 = \sum_i (y_i - \widehat{y}_i)^2 + \underbrace{\sum_i (\widehat{y}_i - \widetilde{y}_i)^2}_{>0} \qquad \text{oder}$$

$$\sum_i (y_i - \widehat{y}_i)^2 < \sum_i (y_i - \widetilde{y}_i)^2 \qquad \text{wenn } b_h \neq c_h \text{ mit } h = 1, 2$$

Dies ist natürlich nicht weiter überraschend, denn schließlich haben wir die OLS Koeffizienten ja hergeleitet, indem wir die Quadratsumme der Residuen minimiert haben ;-) ■

Weitere Übungsbeispiele:

1. Berechnen Sie die OLS-Formel für eine Regression ohne Interzept, d.h. für das Modell $y_i = bx_i + e_i$.
2. Zeigen Sie, dass $\sum_i (x_i - \bar{x}) = 0$.
3. Zeigen Sie, dass $\sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_i (x_i - \bar{x}) y_i$.

2.4 Bedingte Mittelwerte

> "Without data you're just another person with an opinion." (W. Deming)

Wir haben nun eine Methode kennen gelernt, mit deren Hilfe wir aus beobachteten Daten die zwei nicht direkt beobachtbaren Koeffizienten b_1 und b_2 berechnen können, ohne wirklich zu begründen, wozu wir diese benötigen. In diesem Abschnitt werden wir dies nachholen und eine eher intuitive Einsicht vermitteln, wie wir die gefitteten Werte $\widehat{y}$ und die Koeffizienten interpretieren können. Diese Einsichten werden wir im nächsten Abschnitt über das multiple Regressionsmodell erweitern und im Abschnitt zu den Dummy Variablen vertiefen, und sie liefern uns die Grundlagen für das Verständnis des stochastischen Regressionsmodells im nächsten Kapitel.

Erinnern wir uns, dass die OLS Methode in erster Linie eine *Zerlegungsmethode* ist, eine interessierende Variable y wird in eine systematische Komponente $\widehat{y}$ und in eine nicht-systematische Komponente – die Residuen e – zerlegt.

Für die Interpretation interessieren wir uns ausschließlich für die *systematische* Komponente, da wir bei einer richtigen Spezifikation aus den Residuen wenig lernen können.

Die *systematische* Komponente ist

$$\widehat{y}_i = b_1 + b_2 x_i$$

oder für das frühere Gebrauchtautobeispiel $\widehat{\text{Preis}}_i = 23\,057 - 2\,636\,\text{Alter}_i$ (siehe Abbildung 2.1, Seite 5), wobei der Preis hier in Euro und das Alter in Jahren gemessen wurde.

Die *systematische Komponente* ist einfach der gefittete Preis, und dieser wird durch eine *lineare Funktion* in Abhängigkeit vom Alter beschrieben (bzw. im Rahmen eines Modells 'erklärt').

Für ein tieferes Verständnis werden wir nun auf zwei Fragen etwas näher eingehen, nämlich

1. was können wir uns unter der systematischen Komponente $\widehat{y}_i$ intuitiv vorstellen, und
2. welche Bedeutung kommt der (linearen) Funktionsform zu?

Wir werden im Folgenden argumentieren, dass wir die lineare Regression einfach als als *lineare Approximation an die bedingten Mittelwerte* interpretieren können.

Dazu kommen wir nochmals auf das Beispiel mit den Gebrauchtautos zurück, aber wir wenden vorerst einen kleinen Trick an: wir runden die erklärende Variable 'Alter' auf ganze Jahre, um für jedes Jahr mehrere Beobachtungen zu erhalten. Damit wird aus der stetigen Variable 'Alter' eine diskrete Variable (in diesem Fall ein Integer), die wir 'AlterJ' nennen; in diesem Beispiel nimmt die Variable 'AlterJ' einen ganzzahligen Wert zwischen 0 und 5 an, d.h. AlterJ $\in \{0, 1, 2, \ldots, 5\}$ (siehe Tabelle 2.1, Seite 6).[7]

Tabelle 2.3 zeigt die gleichen Beobachtungen wie Tabelle 2.1, allerdings anders angeordnet, gruppiert nach dem gerundeten Alter (AlterJ). Für AlterJ = 0 (d.h. $0 <$ Alter ≤ 0.5) liegen zum Beispiel drei Beobachtungen vor. Da wir nun für jedes gerundete Alter mehrere Beobachtungen haben, können wir *für jede Altersstufe* die Mittelwerte berechnen; der Durchschnittspreis für die drei Autos mit AlterJ = 0 beträgt z.B. 23 567 Euro.

Tabelle 2.3: Autopreise nach gerundetem Alter. $\bar{y}$ bezeichnet das arithmetische Mittel nach Altersklassen und $\widehat{y}$ die gefitteten Werte der Regression $\widehat{y}_i = 22\,709 - 2\,517x_i$. (Die Zahlen stammen aus Tabelle 2.1 (Seite 6), sie sind hier nur anders angeordnet.)

	AlterJ = 0	AlterJ = 1	AlterJ = 2	AlterJ = 3	AlterJ = 4	AlterJ = 5
	24000	19980	21850	18000	10000	11100
	23900	18890	14500	17200	10000	6700
P	22800	18890	16900	15200	15300	11900
r		20100	15600	14450	14320	
e		19700	18600	15000	12350	
i		19300	18500	16900	12500	
s		19000	18500	15500	12350	
e			18000	14000		
			17500			
			16900			
			17700			
			17800			
n	3	7	12	8	7	3
$\overline{y}$	23567	19409	17696	15781	12403	9900
$\Delta\overline{y}$		−4158	−1713	−1915	−3378	−2503
$\widehat{y}$	22709	20192	17675	15158	12641	10124
$\Delta\widehat{y}$		−2517	−2517	−2517	−2517	−2517

Den Mittelwert für eine Altersstufe nennen wir im Folgenden einen *bedingten* Mittelwert, wir schreiben

$$(\overline{\text{Preis}}|\text{AlterJ} = 0) = 23\,567$$

[7] In diesem Fall verwenden wir die Variable 'AlterJ' um *Alterskategorien* zu bilden, solche Variablen werden deshalb *kategoriale* Variablen genannt; wir werden diese im Abschnitt zu Dummy Variablen ausführlicher diskutieren.

und lesen dies als: Mittelwert des Preises, *gegeben* das Alter ist Null Jahre.

Wenn wir dies für alle Altersstufen machen erhalten wir die *bedingte Mittelwertfunktion*, jeder Altersstufe 'AlterJ' wird ein bedingter Mittelwert zugeordnet

$$(\overline{\text{Preis}}|\text{AlterJ}) = \begin{cases} 23567 & \text{für AlterJ} = 0 \\ 19409 & \text{für AlterJ} = 1 \\ 17696 & \text{für AlterJ} = 2 \\ 15781 & \text{für AlterJ} = 3 \\ 12403 & \text{für AlterJ} = 4 \\ 9900 & \text{für AlterJ} = 5 \end{cases}$$

vergleiche Tabelle 2.3 Zeile $\overline{y}$.

Dies ermöglicht – im Sinne der deskriptiven Statistik – eine 'Verdichtung' der Information aus Tabelle 2.3, anstelle der 40 Beobachtungen haben wir nur noch 6 Mittelwerte, je einen für jede Alterkategorie.

Mit Hilfe dieser bedingten Mittelwertfunktion können wir einfach erkennen, dass die Durchschnittspreise mit dem Alter fallen, im ersten Jahr z.B. um 4158 Euro, im zweiten Jahr um 1713 Euro, usw., siehe Zeile $\Delta\overline{y}$ $(:= \overline{y}_j - \overline{y}_{j-1}$, mit $j = 1, \ldots, 5)$ in Tabelle 2.3.

Eine noch größere 'Informationsverdichtung' erreichen wir, wenn wir auf die 40 Beobachtungen aus Tabelle 2.3 die OLS Methode anwenden.

Für die gerundete erklärende Variable 'AlterJ' erhalten wir

$$\widehat{\text{Preis}}_i = 22\,709 - 2\,517\text{AlterJ}_i$$

Für Autos mit AlterJ = 4 erhalten wir z.B. den gefitteten Wert $\widehat{\text{Preis}}|(\text{Alter} = 4) = 22\,709 - 2\,517 * 4 \approx 12641$, und analog die gefitteten Werte für die anderen Altersklassen (gerundet), siehe auch Zeile $\widehat{y}$ in Tabelle 2.3

$$(\widehat{\text{Preis}}|\text{AlterJ}) = \begin{cases} 22709 & \text{für AlterJ} = 0 \\ 20192 & \text{für AlterJ} = 1 \\ 17675 & \text{für AlterJ} = 2 \\ 15158 & \text{für AlterJ} = 3 \\ 12641 & \text{für AlterJ} = 4 \\ 10124 & \text{für AlterJ} = 5 \end{cases}$$

vergleiche Tabelle 2.3 Zeile $\widehat{y}$.

Man beachte, dass aufgrund der linearen Funktionsform die *Änderung* $\Delta(\widehat{\text{Preis}}|\text{AlterJ}) = b_2 = 2\,517$ konstant ist.

Für die Berechnung dieser Werte benötigen wir lediglich die zwei OLS Koeffizienten b_1 und b_2, wir erreichen also eine noch größere 'Informationsverdichtung', die allerdings auf Kosten der Genauigkeit geht. Dies ist der übliche *'trade off'* zwischen Informationsverdichtung und Genauigkeit, der uns noch öfter begegnen wird.

Abbildung 2.6 zeigt die zugrunde liegenden Daten, die bedingten Mittelwerte sowie die mit der OLS Methode gefitteten Werte.

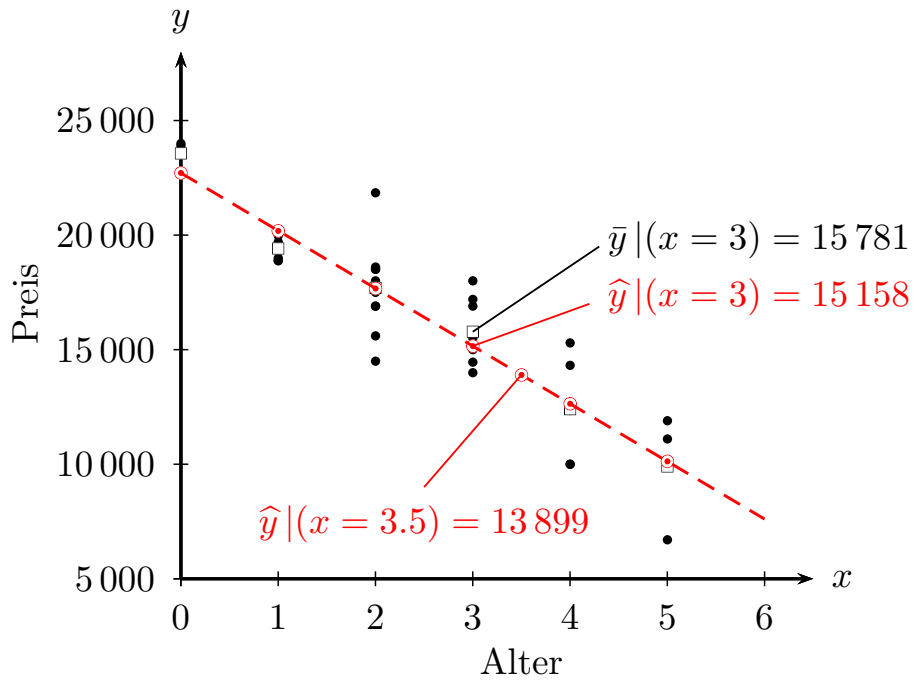


Abbildung 2.6: Deskriptive Regression als lineare Approximation an die 'bedingte Mittelwertfunktion'. (• Beobachtungen; □ bedingte Mittelwerte; ⊙ lineare Approximation).

Offensichtlich liegen die bedingten Mittelwerte (d.h. Mittelwerte nach Alterskategorie) und die OLS-gefitteten Werte sehr nahe beieinander, teilweise so nahe, dass sie sich in der Abbildung teilweise überdecken.[8]

Intuitiv können wir uns die auf der Regressionsgerade liegenden gefitteten Werte $\widehat{y}$ als *lineare Approximation an die bedingten Mittelwerte* vorstellen. Wir werden diese Interpretation später weiter vertiefen, wenn wir Dummy Variablen diskutieren. Hier dient sie v.a. als Vorbereitung auf die stochastische Regressionsanalyse, in deren Rahmen wir die $\widehat{y}$ ganz ähnlich als lineare Approximation an die *bedingten Erwartungswerte* interpretieren werden.

Als nächstes wenden wir uns der linearen Funktionsform zu. Aufgrund der unterstellten linearen Funktionsform $\widehat{y} = b_1 + b_2 x$ können wir $\widehat{y}$ für beliebige x berechnen, in unserem Beispiel können wir z.B. den gefitteten Preis $\widehat{y}_i$ für ein Auto mit einem Alter von 3.5 Jahren berechnen: $(\widehat{y}|x = 3.5) = 22\,709 - 2\,517 \times 3.5 \approx 13\,899$, obwohl in diesem Datensatz kein einziges Auto mit einem Alter von 3.5 Jahren existiert. Trotzdem können wir uns $(\widehat{y}|x = 3.5) = 13\,899$ als eine lineare Approximation an den (hypothetischen) Durchschnittspreis von Autos mit einem Alter von 3.5 Jahren vorstellen. Man beachte aber, dass diese Interpretation auf der *angenommenen* linearen Funktionsform beruht, die diese Interpolation ermöglichte.

Diese Intuition bleibt auch dann gültig, wenn wir keine wiederholten y-Beobachtungen für Ausprägungen der x-Variable haben, wie z.B. im ursprünglichen Beispiel aus Abbildung 2.1 (Seite 5).

[8] Dies muss natürlich nicht immer so sein, aber in diesem Beispiel werden die Beobachtungen durch eine lineare Funktion relativ gut approximiert.

In diesem Sinne werden wir in der deskriptiven Regressionsanalyse die gefitteten Werte $(\widehat{y}|x = \underline{x})$ generell als lineare Approximation an die bedingten Mittelwerte für x vorstellen, wobei $\underline{x}$ eine konkrete mögliche Ausprägung von x bezeichnet (z.B. AlterJ $= 3.5 = \underline{x}$)

$$\widehat{y}|(x = \underline{x}) \overset{\text{lin}}{\approx} \bar{y}|(x = \underline{x})$$

wobei hier $\overset{\text{lin}}{\approx}$ für 'lineare Approximation' steht.

Nachdem es extrem umständlich wäre, jedes Mal von einer 'linearen Approximation an den bedingten Mittelwert' zu sprechen, wollen wir in Zukunft einfach von einem 'mittleren' Preis oder Durchschnittspreis sprechen, aber es ist für die späteren Ausführungen wichtig festzuhalten, dass wir die gefitteten Werten $\widehat{y}_i$ als lineare Approximation an die bedingten Mittelwerte interpretieren.

In den meisten Fällen interessieren wir uns dafür, wie sich eine Änderung von x 'im Durchschnitt' auf y auswirkt, zum Beispiel, um wie viele Euro der 'durchschnittliche' Preis von Gebrauchtautos sinkt, wenn das Alter um ein Jahr zunimmt.

Mit Hilfe der OLS Methode können wir diese Frage zumindest für eine lineare Approximation an die bedingten Mittelwerte von y beantworten, denn die erste Ableitung (d.h. der Differentialquotient $d\widehat{y}/dx$) der Regressionsfunktion[9] liefert uns die gewünschte Antwort, den Steigungskoeffizienten b_2

$$\widehat{y} = b_1 + b_2 x \qquad \rightarrow \qquad \frac{d\,\widehat{y}}{dx} = b_2$$

Diese erste Ableitung wird häufig als *'marginaler Effekt'* bezeichnet, wobei der Begriff 'marginal' auf eine infinitesimal kleine Änderung von x hinweist.

Für lineare Funktionen spielt es allerdings keine Rolle, ob wir infinitesimal kleine oder diskrete Änderungen betrachten, der *marginale Effekt* ist in diesem Fall gleich dem Steigungskoeffizienten b_2, und somit über den gesamten Funktionsverlauf konstant

$$\frac{d\widehat{y}}{dx} = \frac{\Delta\widehat{y}}{\Delta x} = b_2$$

aber dies gilt natürlich nur für lineare Funktionsformen (vgl. Abbildung 2.7).

Der Steigungskoeffizient b_2 sagt uns also, dass eine Zunahme von x um eine Einheit mit einer Änderung von $\widehat{y}$ um b_2 Einheiten einher geht, wobei wir $\widehat{y}$ in der deskriptiven Regressionsanalyse als lineare Approximation an den bedingten Mittelwert interpretieren können.

Dazu muss natürlich auch bekannt sein, in welchen Einheiten x und $\widehat{y}$ gemessen wurden. Im Beispiel mit den Gebrauchtautos sagt uns b_2, um wie viele *Euro* sich die lineare Approximation an den bedingten Durchschnittspreis ändert, wenn das Alter um ein *Jahr* zunimmt, nämlich um 2 517 Euro.

$$\widehat{\text{Preis}} = 22\,709 - 2\,517\,\text{AlterJ} \qquad \rightarrow \qquad \frac{d\,\widehat{\text{Preis}}}{d\,\text{AlterJ}} = 2\,517$$

[9] Wir lassen hier den Subindex i weg, da die lineare Approximation nicht nur für die beobachteten x_i gilt, sondern weil wir zumindest prinzipiell für jedes x ein dazugehöriges $\widehat{y}$ berechnen können; natürlich wird dies meist nur für $x_{\min} \leq x \leq x_{\max}$ Sinn machen.

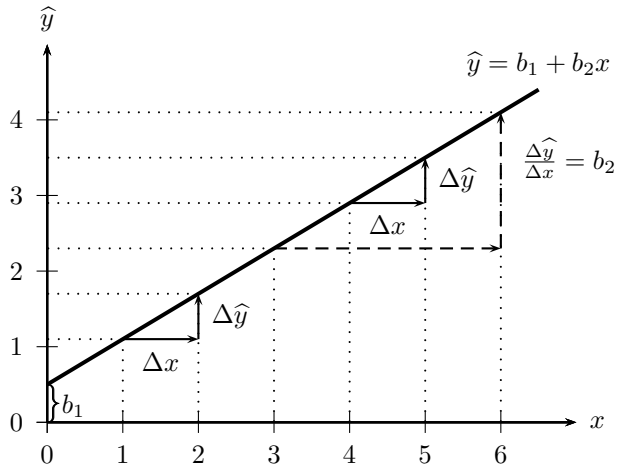


Abbildung 2.7: Lineare Funktion $\widehat{y} = b_1 + b_2 x = 0.5 + 0.6x$. Eine Zunahme von x um eine Einheit geht einher mit einer Änderung von $\widehat{y}$ um $+0.6$ Einheiten.

Für eine korrekte Interpretation dieser Regressionskoeffizienten sind unbedingt die Dimensionen von y und x erforderlich, in diesem Beispiel wird $\widehat{y}$ in Euro und x in Jahren gemessen, d.h. wenn das Alter *um ein Jahr* zunimmt sinkt der durchschnittliche Preis $\widehat{y}$ um 2 517 *Euro*.

Wichtig ist auch zu betonen, dass uns eine solche Regression per se nichts über eine mögliche Kausalbeziehung verrät, sie beschreibt lediglich eine Assoziation. Die bloßen Daten sagen uns nichts über eine mögliche Ursachen-Wirkungsbeziehung, dies wäre eine weit über die reine Beschreibung hinausgehende Interpretation. In einem späteren Kapitel über *Endogenität* werden wir die Möglichkeit von Kausalaussagen ausführlicher diskutieren, und wir werden sehen, dass Kausalaussagen immer einer besonderen Begründung bedürfen.

Man beachte auch, dass wir mit der OLS Methode *von vornherein* eine lineare Funktionsform unterstellt haben, und dass die Interpretation der Koeffizienten unmittelbar aus dieser angenommenen Funktionsform folgt.

In Beispiel mit den Gebrauchtautos wurden die bedingten Mittelwerte durch eine lineare Funktion sehr gut approximiert, aber dies muss natürlich nicht immer der Fall sein.

Abbildung 2.8 zeigt Datenpunkte, die durch eine nicht-lineare Funktion offensichtlich deutlich besser beschrieben werden als durch die strichliert eingezeichnete einfache Regressionsgerade.

In diesem sehr speziellen Fall können die Punkte durch eine quadratische Funktion $\widehat{y} = b_1 + b_2 x + b_3 x^2$ gut beschrieben werden, und wir werden später sehen, dass auch solche Funktionen einfach mit der OLS Methode berechnet werden können. Allerdings ist selbst in diesem einfachen Fall der marginale Effekt nicht mehr konstant, sondern ändert sich mit x; wenn wir die quadratische Funktion nach x ableiten erhalten wir

$$\text{Marg. Effekt für } \widehat{y} = b_1 + b_2 x + b_3 x^2 \quad \rightarrow \quad \frac{d\,\widehat{y}}{dx} = b_2 + 2b_3 x$$

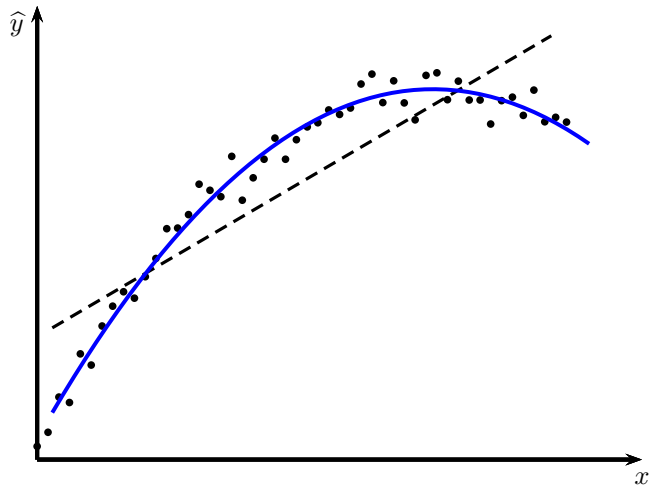


Abbildung 2.8: Eine lineare Funktion $\widehat{y} = b_1 + b_2 x$ kann einen sehr schlechten Fit liefern, wenn der tatsächliche Zusammenhang nicht-linear ist. Offensichtlich würde in diesem Fall eine nicht-lineare Funktion wie die blaue Linie einen deutlich besseren Fit liefern, aber für nicht-lineare Funktionen ist der marginale Effekt (Steigung der Tangente) nicht konstant, d.h. für jedes x unterschiedlich.

d.h., der marginale Effekt (die Steigung der Tangente) ist in diesem Beispiel für jedes x unterschiedlich groß.

Darüber hinaus gibt es Schätzverfahren für komplexere Formen von Nicht-Linearitäten, z.B. Spline Funktionen. Abbildung 2.9 zeigt eine solche nicht-lineare Schätzung für das Autobeispiel.

Offensichtlich kann diese Funktion die Daten 'genauer' abbilden, man erkennt z.B., dass der 'bedingte mittlere Preis' im ersten Jahr stärker fällt als in den späteren Jahren. Allerding hat diese 'genauere' Beschreibung auch Kosten, die 'Informationsverdichtung' ist deutlich kleiner, auch die marginalen Effekte unterscheiden sich für alle Ausprägungen von x, und können deshalb nicht mehr einfach angegeben werden.

Hier wird wieder ein allgemeineres Prinzip sichtbar, es gibt einen *'trade-off'* zwischen der Genauigkeit der Beschreibung und der 'Informationsverdichtung', bzw. Einfachheit.

Die größere Einfachheit wird häufig durch restriktivere Annahmen erreicht (z.B. die Linearität der Funktionsform). Diese Einfachheit hat in den meisten Fällen den Vorteil einer besseren Interpretierbarkeit der Ergebnisse, aber dieser Vorteil bringt meistens Kosten in Bezug auf die Genauigkeit mit sich.

Das optimale Ausmaß an Datenverdichtung hängt vom Verwendungszweck ab, wenn man mit dem Auto von Rom nach Hamburg fahren möchte benötigt man eine Strassenkarte mit einem anderen Maßstab, als wenn man in Rom das nächstgelegene Hotel sucht.

Generell können wir festhalten

$$\text{Daten} + \text{Theorie(Annahmen)} \rightarrow \text{Schlussfolgerungen}$$

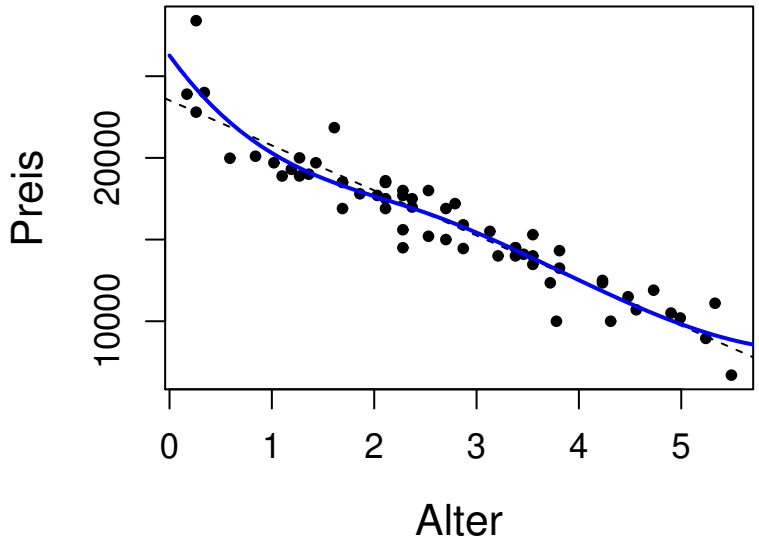


Abbildung 2.9: Spline-Funktion für die Preise von Gebrauchtautos

Es gibt keine Datenanalyse, die völlig ohne Theorie und den der Theorie zugrunde liegenden Annahmen auskommt. Selbst für die Berechnung eines einfachen Mittelwerts muss vorher geklärt werden, 'was' gezählt werden soll, oder in anderen Worten, eine Klassifizierung vorgenommen werden. In der Regel erlauben stärkere Annahmen weiterreichende Schlussfolgerungen, aber inwieweit diese dann auch zutreffend sind hängt weitgehend davon ab, inwieweit die Annahmen korrekt waren. Deshalb sollten wir uns jeweils sehr genau bewusst sein, welche Annahmen unserer Analyse zugrunde liegen, und welche Konsequenzen zu befürchten sind, wenn die Annahmen verletzt sind.

Im Beispiel mit den Gebrauchtautos ist die Annahme der linearen Funktionsform für die Altersklassen $0 - 5$ offensichtlich ziemlich gut erfüllt, aber die gleiche Annahme würde für 10 Jahre alte Gebrauchtautos offensichtlich ziemlich unsinnige gefittete Preise liefern.

Darstellung von Regressionsgleichungen und Standardfehler

In Lehrbüchern werden Regressionsgleichungen häufig in folgender Form wiedergegeben

$$\begin{array}{rcccc} \text{Preis} & = & 22\,709.303 & - & 2\,517.267\ \text{AlterJ} \\ & & (532.689)^{***} & & (190.125)^{***} \end{array}$$

$$R^2 = 0.822, \qquad n = 40$$

	Dependent variable:	
	Preis	
	(1)	(2)
Constant	$22,709.300^{***}$	$23,056.710^{***}$
	(532.689)	(468.871)
AlterJ	$-2,517.267^{***}$	
	(190.125)	
Alter		$-2,635.669^{***}$
		(166.935)
Observations	40	40
R^2	0.822	0.868
Note:	*p<0.1; **p<0.05; ***p<0.01	

Tabelle 2.4: Darstellung von Regressionen in Tabellenform, hier in Spalte (1) mit dem auf ganze Jahre gerundeten 'AlterJ' (vgl. obige Regression in Zeilenform) , und in Spalte (2) mit nicht gerundetem Alter. Diese Darstellung wurde mit Hilfe des R-packages `stargazer` erzeugt.

Unter den Koeffizienten in Klammern sind die *Standardfehler der Koeffizienten* angegeben. Deren Bedeutung wird sich erst in der stochastischen Regressionsanalyse voll erschließen, und im Kapitel zu den Eigenschaften und von OLS Schätzfunktionen werden wir deren Berechnung und genaue Interpretation ausführlich diskutieren.

Die Standardfehler der Koeffizienten können als Indikator für die Präzision der Schätzung des Koeffizienten interpretiert werden. Im Kern erlauben sie uns zu beurteilen, ob im Rauschen der Daten ein Signal zu finden ist, das sich stark genug gegenüber der Nullhypothese abhebt, um ernst genommen zu werden.

Dazu können Sie sich als grobe *Faustregel* einprägen, dass die die Koeffizienten mindestens doppelt so groß wie deren Standardfehler sein sollten, damit die Nullhypothese mit einer Fehlerwahrscheinlichkeit von 5% verworfen werden kann.

Dies ist bei beiden obigen Koeffizienten der Fall, und wird auch durch die Sterne neben den Standardfehlern symbolisch ausgedrückt. Wie wir im Kapitel zu den Hypothesentests ausführlich zeigen werden, symbolisiert ein Stern, dass die Nullhypothese, dass der 'wahre' Koeffizient (einer unbeobachteten Grundgesamtheit) gleich Null ist, auf einem Signifikanzniveau von 10% verworfen werden kann, zwei Sterne implizieren ein Signifikanzniveau von 5%, und drei Sterne ein Signifikanzniveau von 1%. All dies wird später noch ausführlich diskutiert werden.

Die obige Schreibweise in Zeilen eignet sich schlecht, wenn viele Regressoren verwendet werden, oder wenn mehrere Regressionsgleichungen miteinander verglichen werden sollen. Deshalb wird in der Literatur fast ausschließlich eine Darstellung in Tabellenform gewählt. Tabelle 2.4 zeigt in Spalte (1) den gleichen Regressionsoutput wie oben in Tabellenform.

2.5 Das Bestimmtheitsmaß

> *"The secret of success is honesty and fair dealing. If you can fake those, you've got it made."*
> (vermutl. Groucho Marx, 1890–1977)

Die Regressionsgerade kann die Daten – je nach der Beschaffenheit der Daten – mehr oder weniger gut beschreiben.

Abbildung 2.10 zeigt zwei Extremfälle, im linken Panel liegen die Punkte sehr nahe an der Regressionsgerade, d.h. der *'Fit'* ist sehr gut, und die Daten werden durch die Regressionsgerade gut beschrieben – der Informationsverlust ist bei Beschreibung der Daten durch die Regressionsgerade eher gering. Im Gegensatz dazu werden die Daten im rechten Panel durch die Regressionsgerade weniger gut beschrieben, d.h. der 'Fit' ist schlecht. Wenn man im zweiten Fall *ausschließlich* die Regressionsgerade kennt, erhält man nur eine schlechte Vorstellung von den zugrunde liegenden Daten – der Informationsverlust bei Beschreibung der Daten durch eine Regressionsgerade ist groß.

Praktisch wäre, wenn wir eine einfache Kennzahl hätten, die uns angibt, wie 'gut' die Anpassung der Regressionsgeraden an die Beobachtungspunkte ist. Eine solche Kennzahl für die Güte des 'Fits' existiert tatsächlich, nämlich das 'Bestimmtheitsmaß' R^2.

Wir werden gleich zeigen, dass das Bestimmtheitsmaß als der Anteil der durch x erklärten Streuung von y an der gesamten Streuung von y interpretiert werden kann.

Da es sich um einen Anteil handelt, kann das Bestimmtheitsmaß R^2 für gewöhnliche Regressionen mit Interzept ausschließlich Werte zwischen Null und Eins annehmen. Umso besser der 'Fit' ist, umso näher liegt das Bestimmtheitsmaß bei Eins. Das linke Panel von Abbildung 2.10 zeigt einen relativ guten 'Fit' mit einem Bestimmtheitsmaß von $R^2 = 0.94$. Wenn das Bestimmtheitsmaß den Wert Eins annimmt ($R^2 = 1$) liegen die Beobachtungspunkte exakt auf der Regressionsgeraden. Umgekehrt liegt das Bestimmtheitsmaß umso näher bei Null, umso schlechter der 'Fit' ist. Das rechte Panel in Abbildung 2.10 zeigt einen sehr schlechten 'Fit' mit einem Bestimmtheitsmaß von $R^2 = 0.03$.

Das Bestimmtheitsmaß interpretiert man am einfachsten als ein deskriptives Maß zur Beurteilung der *'Güte der Anpassung'* der Regressionsgeraden an die Beobachtungspunkte.

Im Wesentlichen beruht es auf einer Streuungszerlegung, wir zerlegen die gesamte Streuung von y in einen 'erklärten' und einen 'unerklärten' Teil; Abbildung 2.11 zeigt die Idee.

Zuerst beachte man, dass eine Regressionsgerade mit Interzept immer durch den Mittelwert von x und y verläuft.

Dies folgt direkt aus den Bedingungen erster Ordnung und kann einfach gezeigt werden, indem wir den Mittelwert $\bar{x}$ in die Gleichung für die gefitteten Werte $\widehat{y}_i = b_1 + b_2 x_i$ einsetzen, also

$$\widehat{y}_{\bar{x}} = b_1 + b_2 \bar{x}$$

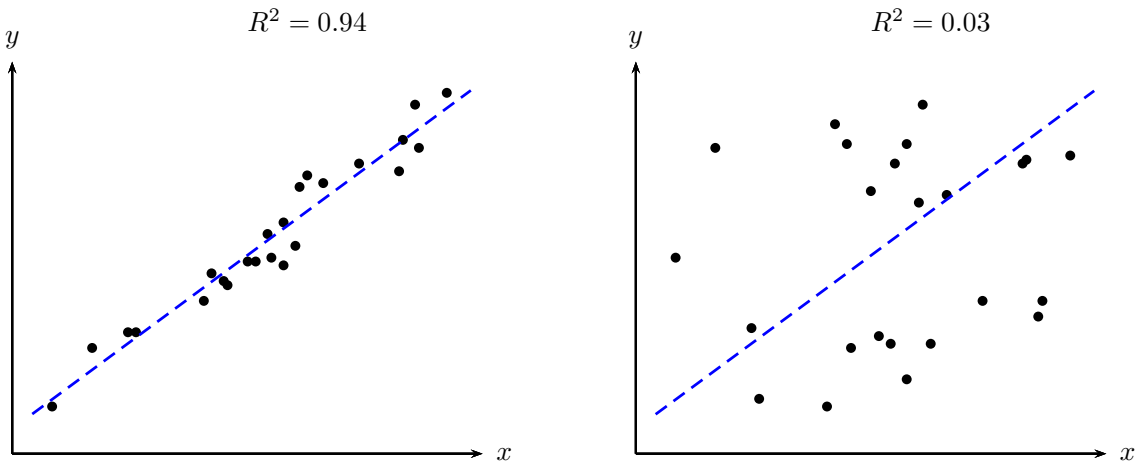


Abbildung 2.10: Der Zusammenhang zwischen zwei Variablen kann durch eine Regressionsgerade mehr oder weniger gut beschrieben werden.

wobei $\widehat{y}_{\bar{x}}$ den Wert von $\widehat{y}$ für $\bar{x}$ bezeichnet.

Wenn die Regressionsgerade durch den Punkt $(\bar{x}, \bar{y})$ läuft muss $\widehat{y}_{\bar{x}} = \bar{y}$ sein. Dies ist tatsächlich so, um dies zu sehen setzen wir die OLS Formel für das Interzept $b_1 = \bar{y} - b_2\bar{x}$ in obige Gleichung ein und erhalten

$$\begin{aligned} \widehat{y}_{\bar{x}} &= b_1 + b_2\bar{x} \\ &= \underbrace{\bar{y} - b_2\bar{x}}_{b_1} + b_2\bar{x} \\ &= \bar{y} \end{aligned}$$

Man beachte, dass dies nur gilt, wenn die Regression ein Interzept enthält, denn wir haben hier $b_1 = \bar{y} - b_2\bar{x}$ verwendet um dies zu zeigen.

Übungsbeispiel: Zeigen Sie, dass der Mittelwert der gefitteten y gleich dem Mittelwert der y ist, d.h. $\overline{\widehat{y}} = \bar{y}$. Gilt dies auch für Regressionen ohne Interzept?

Kommen wir zurück und erinnern wir uns, dass die OLS Methode in erster Linie eine Zerlegungsmethode ist, sie hilft uns eine Variable y_i in eine systematische Komponente $\widehat{y}_i$ und den unsystematischen 'Rest' e_i zu zerlegen.

Nehmen wir zum Beispiel an, es gebe einen positiven Zusammenhang zwischen Körpergröße x und Gewicht y. Dieser Zusammenhang ist natürlich nicht exakt, Sie kennen die Geschichte vom spannenlangen Hansel und der nudeldicken Dirn, aber zumindest im Durchschnitt erwarten wir von größeren Personen ein höheres Gewicht.

Was ist die beste Schätzung für das Gewicht einer Person, wenn wir die Körpergröße dieser Person nicht kennen? Genau, das Durchschnittsgewicht aller Personen $\bar{y}$, oder in anderen Worten, das Gewicht einer Person mit Durchschnittsgröße $\bar{x}$, denn wir haben gerade gezeigt, dass die Regressionsgerade immer durch den Punkt $(\bar{x}, \bar{y})$ läuft. Wenn die Person tatsächlich das Gewicht y_i hat machen wir den Fehler von $y_i - \bar{y}$.

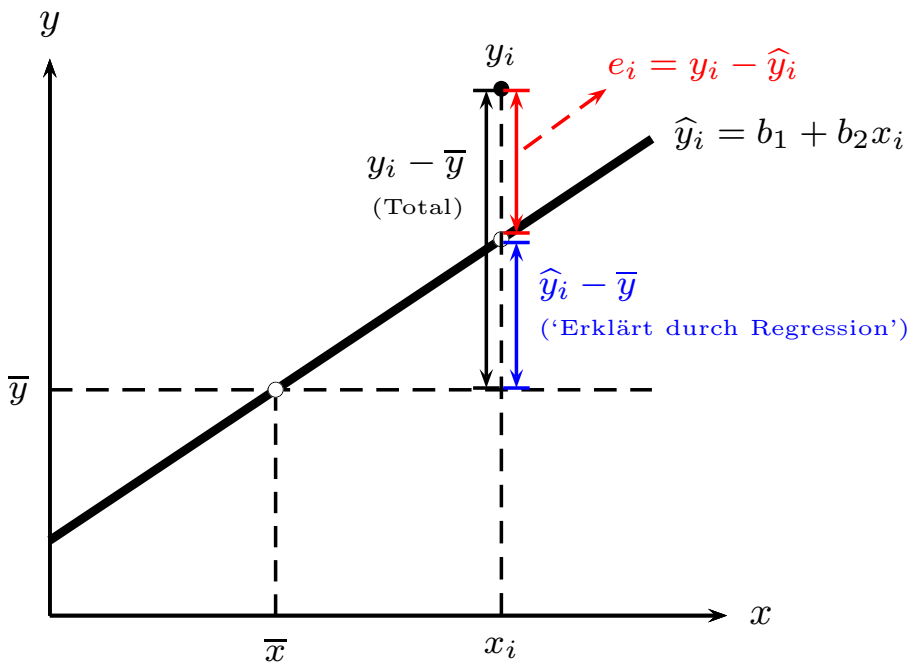


Abbildung 2.11: Zerlegung der gesamten Streuung von y in einen 'erklärten' und einen 'unerklärten' Teil.

Angenommen wir erfahren nun, dass diese Person 190 cm groß ist. In diesem Fall werden wir diese Information nützen um unsere Schätzung zu revidieren, $\widehat{y}_i = b_1 + b_2\, 190$. Wenn wir das tatsächliche Gewicht y_i nicht kennen erlaubt uns diese Information zwar die Schätzung zu verbessern, aber trotzdem ist es nur eine Schätzung, wir müssen immer noch mit einem Fehler $y_i - \widehat{y}_i = e_i$ rechnen.

Diese Überlegung erlaubt uns den Fehler, den wir ohne Kenntnis von x_i machen würden, d.h. $y_i - \bar{y}$, in zwei Teile zu zerlegen, in einen Teil den wir durch Kenntnis von x 'erklären' können $\widehat{y}_i - \bar{y}$, und in den Rest $y_i - \widehat{y}_i = e_i$.

Abbildung 2.11 fasst diese Überlegungen zusammen. Wir haben eine einzelne Beobachtung (x_i, y_i) herausgegriffen und beginnen damit, für diese Beobachtung die gesamte Abweichung von y_i vom Mittelwert $\bar{y}$, also die Distanz $y_i - \bar{y}$, in eine 'durch die Regression erklärte' Distanz $\widehat{y}_i - \bar{y}$ und in eine 'unerklärte' Distanz $e_i = y_i - \widehat{y}_i$ zu zerlegen.

Für eine einzelne Beobachtung wie in Abbildung 2.11 gilt

$$y_i - \bar{y} = (\widehat{y}_i - \bar{y}) + (y_i - \widehat{y}_i)$$

Unter Streuung verstehen wir hier die Summe der quadrierten Abweichungen. Deshalb quadrieren wir den obigen Ausdruck und summieren über alle Beobachtungen

$$\begin{aligned}
(y_i - \bar{y})^2 &= [(\widehat{y}_i - \bar{y}) + (y_i - \widehat{y}_i)]^2 \\
&= (\widehat{y}_i - \bar{y})^2 + (y_i - \widehat{y}_i)^2 + 2(\widehat{y}_i - \bar{y})(y_i - \widehat{y}_i) \\
\sum_i (y_i - \bar{y})^2 &= \sum_i (\widehat{y}_i - \bar{y})^2 + \sum_i (y_i - \widehat{y}_i)^2 + \\
&\quad +2 \sum_i (\widehat{y}_i - \bar{y})(y_i - \widehat{y}_i) \qquad (2.10)
\end{aligned}$$

Wir werden nun zeigen, dass der dritte Term auf der rechten Seite aufgrund der Eigenschaften der OLS Methode immer gleich Null ist, wenn die Regression ein Interzept enthält. Diese Eigenschaft folgt aus den Bedingungen erster Ordnung $\sum_i e_i = 0$ und $\sum_i x_i e_i = 0$ (Gleichungen (2.2) und (2.3), Seite 15).
Dies kann einfach gezeigt werden, der dritte Term von Gleichung (2.10) ist

$$\begin{aligned} 2\sum_i (\widehat{y}_i - \bar{y})(y_i - \widehat{y}_i) &= 2\sum_i (\widehat{y}_i - \bar{y})e_i \\ &= 2\sum_i \widehat{y}_i e_i - 2\bar{y}\sum_i e_i \end{aligned}$$

Da für Regressionen mit Interzept immer gilt $\sum_i e_i = 0$ (Gleichung (2.2), Seite 15) bleibt nur zu zeigen, dass $\sum_i \widehat{y}_i e_i = 0$.
Dazu setzen wir $\widehat{y}_i = b_1 + b_2 x_i$ ein

$$\begin{aligned} \sum_i \widehat{y}_i e_i &= \sum_i (b_1 + b_2 x_i)e_i \\ &= \sum_i (b_1 e_i + b_2 x_i e_i) \\ &= b_1 \sum_i e_i + b_2 \sum_i x_i e_i = 0 \end{aligned}$$

Dieser Ausdruck ist ebenfalls Null, weil die Bedingungen erster Ordnung für die OLS Residuen garantieren, dass $\sum_i e_i = 0$ und $\sum_i x_i e_i = 0$. Damit wurde gezeigt, dass für Regressionen mit Interzept der Kreuzterm von Gleichung (2.10) immer gleich Null ist (d.h. $\sum_i (\widehat{y}_i - \bar{y})(y_i - \widehat{y}_i) = 0$).
Deshalb zerfällt die Gesamtstreuung von y um den Mittelwert in bloss zwei Terme, in die durch x 'erklärte' Streuung und in die 'unerklärte' Streuung

$$\sum_i (y_i - \bar{y})^2 = \sum_i (\widehat{y}_i - \bar{y})^2 + \sum_i (y_i - \widehat{y}_i)^2$$

bzw.

$$\underbrace{\sum (y_i - \bar{y})^2}_{\text{TSS}} = \underbrace{\sum (\widehat{y}_i - \bar{y})^2}_{\text{ESS}} + \underbrace{\sum e_i^2}_{\text{SSR}}$$

wobei TSS für 'Total Sum Squared' steht, also die gesamte Streuung der y_i um den Mittelwert $\bar{y}$. ESS ist die 'Explained Sum Squared', die Streuung der gefitteten Werte $\widehat{y}_i$ um den Mittelwert $\bar{y}$, und SSR steht für 'Sum of Squared Residuals', die Streuung der y_i um die Regressionsgerade, das ist die Quadratsumme der Residuen.

Das Bestimmtheitsmaß ist schließlich definiert als Anteil der durch die Regressionsgerade *erklärten Streuung* ESS an der *gesamten Streuung* TSS

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = \frac{\text{TSS} - \text{SSR}}{\text{TSS}} = 1 - \frac{\text{SSR}}{\text{TSS}} = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2} \tag{2.11}$$

In anderen Worten, das Bestimmtheitsmaß R^2 gibt an, welcher Anteil der gesamten Streuung von y durch die Regressionsgerade (oder genauer, durch die erklärende Variable x) erklärt wird.

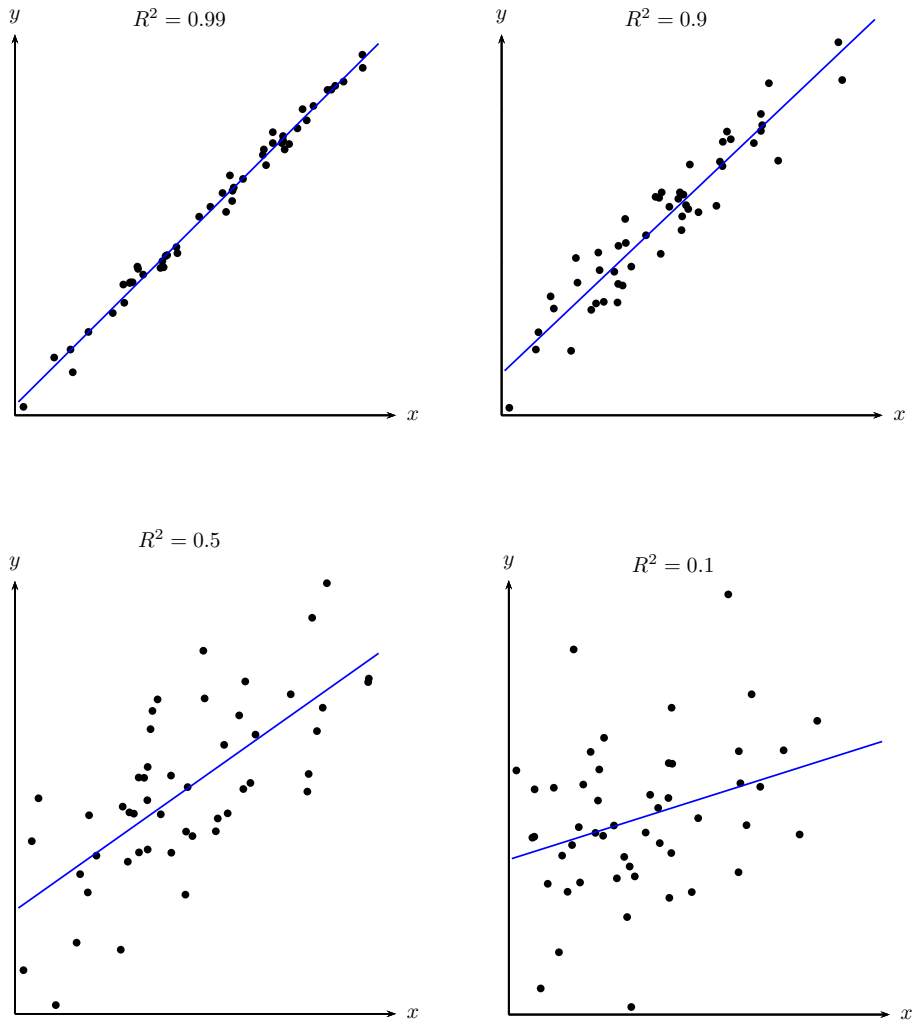


Abbildung 2.12: Das Bestimmtheitsmaß R^2 ist ein Indikator für die Streuung um die Regressionsgerade.

Da es sich um einen Anteil handelt liegt das Bestimmtheitsmaß für Regressionsgleichungen mit Interzept immer zwischen Null und Eins (dies muss für Regressionsgleichungen ohne Interzept *nicht* gelten! Warum?).

Um einen Eindruck vom Fit bei unterschiedlich großem R^2 zu geben zeigt Abbildung 2.12 einige Regressionsgeraden mit unterschiedlichem R^2.

Andererseits können völlig unterschiedliche Daten zu gleichen R^2 führen, der Klassiker dazu sind die Anscombe Daten, siehe Abbildung 2.13.

Da das R^2 fast immer mit dem Regressionsoutput angegeben wird und einfach zu verstehen ist neigen Anfänger häufig dazu, dem R^2 eine zu große Bedeutung beizulegen. Insbesondere ist der Irrglaube weit verbreitet, dass ein hohes R^2 mit einer genaueren Messung der Regressionskoeffizienten einher gehe, und deshalb ein hohes R^2 'gut' für die Interpretation der Ergebnisse sei. Dies ist falsch, wenn z.B. eine Regressionsgleichung fehlspezifiziert ist, kann sie ein sehr hohes R^2 aufweisen, obwohl die Regressionsgleichung mehr oder weniger unbrauchbar ist. Andererseits kann eine Regressionsgleichung mit einem niedrigen R^2 eine sehr genaue Messung der Regressionskoeffizienten erlauben, wenn genügend Beobachtungen zur Verfügung stehen.

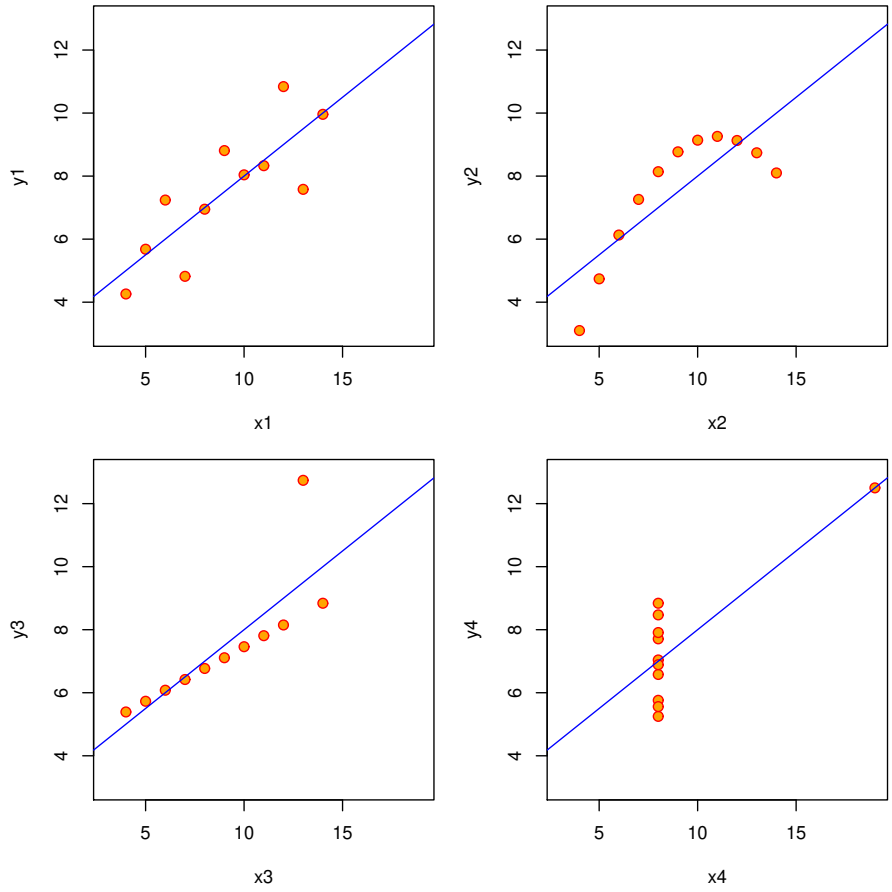


Abbildung 2.13: Anscombe Datensatz, alle vier Regressionen weisen fast das gleiche $R^2 = 0.666$ auf!
Quelle: Anscombe, Francis J. (1973) Graphs in statistical analysis. The American Statistician, 27, 17–21; entnommen dem R package `datasets`, anscombe, Examples.

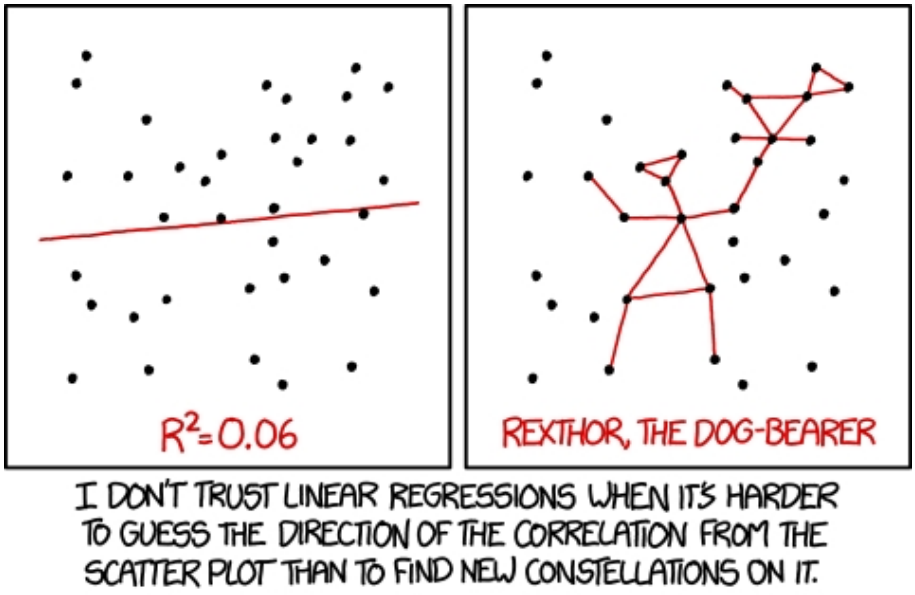


Abbildung 2.14: Quelle xkcd, `http://xkcd.com/1725/`

Übungsbeispiele:

1. Zeigen Sie, dass das Bestimmtheitsmaß R^2 das Quadrat des (Pearsonschen) Korrelationskoeffizienten zwischen den beobachteten Werten y und den gefitteten Werten $\widehat{y}$ ist, d.h. $R^2 = [\text{corr}(y, \widehat{y})]^2 := r^2_{y,\widehat{y}}$.

   Hinweise: Der Pearsonsche Korrelationskoeffizient ist definiert als

   $$r_{y,\widehat{y}} := \frac{\text{cov}(y, \widehat{y})}{\sqrt{\text{var}(y)\,\text{var}(\widehat{y})}}$$

   Berücksichtigen Sie, dass $y = \widehat{y} + e$ und und die Varianzrechenregeln $\text{cov}[x, (y + z)] = \text{cov}(x, y) + \text{cov}(x, z)$. Außerdem erinnern wir uns, dass

   $$R^2 = \frac{\text{ESS}}{\text{TSS}} = \frac{\text{var}(\widehat{y})}{\text{var}(y)}$$

   und dass in Regressionen mit Interzept $\text{cov}(\widehat{y}, e) = 0$ (warum eigentlich?).

2. Zeigen Sie, dass in einer bivariaten Regression das Bestimmtheitsmaß auch gleich dem Quadrat eines Korrelationskoeffizienten zwischen y und x ist (dies gilt nur für bivariate Regressionen).

   $$R^2 = r^2_{y,\widehat{y}} = \frac{[\text{cov}(y, \widehat{y})]^2}{\text{var}(y)\,\text{var}(\widehat{y})} = \frac{[\text{cov}(y, x)]^2}{\text{var}(y)\,\text{var}(x)} := r^2_{y,x}$$

   Lösung: Zeigen Sie zuerst, dass

   $$\begin{aligned} \text{cov}(y, \widehat{y}) &= \text{cov}(y, b_1 + b_2 x) = b_2\,\text{cov}(y, x) \\ \text{var}(\widehat{y}) &= \text{var}(b_1 + b_2 x) = b_2^2\,\text{var}(x) \end{aligned}$$

   Einsetzen gibt

   $$R^2 = r^2_{y,\widehat{y}} := \left(\frac{\text{cov}(y, \widehat{y})}{\sqrt{\text{var}(y)\,\text{var}(\widehat{y})}}\right)^2 = \frac{b_2^2[\text{cov}(y, x)]^2}{\text{var}(y) b_2^2\,\text{var}(x)} = \frac{[\text{cov}(y, x)]^2}{\text{var}(y)\,\text{var}(x)} = r^2_{y,x}$$

2.6 Multiple Regression

> *"With four parameters I can fit an elephant, and with five I can make him wiggle his trunk."*
> (John von Neumann, 1903–1957)

Bisher haben wir uns nur mit der Messung des Zusammenhangs zwischen zwei Variablen x und y befasst. Die meisten Zusammenhänge in der realen Welt sind natürlich deutlich komplexer, fast immer wirken mehrere erklärende Variablen auf eine abhängige y Variable ein. Zum Beispiel wird der Preis von Gebrauchtautos nicht ausschließlich durch das Alter erklärt, sondern auch durch den Kilometerstand, Austattung, frühere Unfälle, Farbe und vieles mehr.

Glücklicherweise lässt sich die OLS Methode sehr einfach für den Fall mit mehreren erklärenden Variablen verallgemeinern.

Der Fall mit zwei erklärenden Variablen kann noch grafisch in einem 3-dimensionalen Raum dargestellt werden; Abbildung 2.15 zeigt eine solche 3-dimensionale Abbildung mit der abhängigen y Variable auf der Vertikalachse und zwei erklärenden Variablen x_2 und x_3 auf den Horizontalachsen. Während wir im bivariaten Modell eine Regressions*gerade* suchten, die die Daten möglichst gut abbildet, suchen wir im Fall mit zwei erklärenden Variablen eine *Regressionsebene*, die die Quadratsumme der Residuen minimiert. Das linke Panel in Abbildung 2.15 zeigt die Beobachtungspunkte im Raum, das rechte Panel zeigt die dazugehörige Regressionsebene mit den auf dieser Ebene liegenden gefitteten Werten $\widehat{y}_i$. Höherdimensionale Fälle, d.h. Fälle mit mehr als zwei erklärenden Variablen, können graphisch nicht mehr dargestellt werden, die mathematische Berechnung ist aber ebenso einfach.

Für zwei erklärende Variablen kann die Regressionsfunktion geschrieben werden als

$$y_i = b_1 + b_2 x_{i2} + b_3 x_{i3} + e_i \quad (\text{mit } i = 1, \dots n)$$

wobei n wieder die Anzahl der Beobachtungen bezeichnet. Man beachte, dass wir nun zwei Subindizes für die erklärenden x benötigen, der erste Subindex $i = 1, \dots, n$ bezeichnet nach wie vor die Beobachtung (bzw. die Zeile der Datenmatrix), der zweite Subindex bezeichnet die erklärende Variable (d.h. die Spalte der Datenmatrix).

Wir können die drei unbekannten Koeffizienten b_1, b_2 und b_3 gleich wie früher berechnen, indem wir die die Quadratsumme der Residuen minimieren:

$$\min_{b_1,b_2,b_3} \sum e_i^2 = \min_{b_1,b_2,b_3} \sum (y_i - b_1 - b_2 x_{i2} - b_3 x_{i3})^2$$

Gesucht sind die Werte b_1, b_2 und b_3, die die folgenden Bedingungen 1. Ordnung erfüllen:

$$\begin{aligned}
\frac{\partial \sum e_i^2}{\partial b_1} &= 2 \sum (y_i - b_1 - b_2 x_{i2} - b_3 x_{i3})(-1) \stackrel{!}{=} 0 \\
\frac{\partial \sum e_i^2}{\partial b_2} &= 2 \sum (y_i - b_1 - b_2 x_{i2} - b_3 x_{i3})(-x_{i2}) \stackrel{!}{=} 0 \\
\frac{\partial \sum e_i^2}{\partial b_3} &= 2 \sum (y_i - b_1 - b_2 x_{i2} - b_3 x_{i3})(-x_{i3}) \stackrel{!}{=} 0
\end{aligned}$$

y	x_1	x_2
2	9	1
5	4	2
4	7	3
8	2	4
9	3	5
9	1	6

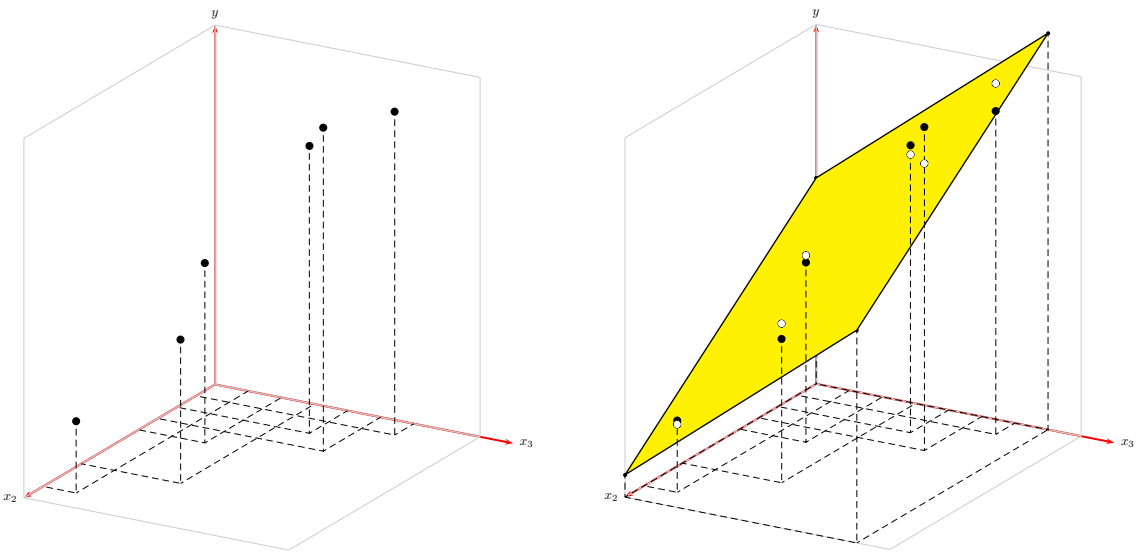


Abbildung 2.15: 3-dimensionale Abbildung der Daten und der Regressionsebene $\widehat{y}_i = 5.73 - 0.51x_{i2} + 0.76x_{i3}$ (gefittete Werte auf der Regressionsebene sind als hohle Kreise dargestellt)

Man beachte, dass diese Gleichungen wieder $\sum e_i = 0$, $\sum e_i x_{i2} = 0$ und $\sum e_i x_{i3} = 0$ implizieren, da $(y_i - b_1 - b_2 x_{i2} - b_3 x_{i3}) = e_i$.

Als Lösungen dieser drei Bedingungen erster Ordnung erhält man nach einiger Rechnerei

$$\begin{aligned} b_2 &= \frac{(\sum \ddot{y}_i \ddot{x}_{i2})(\sum \ddot{x}_{i3}^2) - (\sum \ddot{y}_i \ddot{x}_{i3})(\sum \ddot{x}_{i2} \ddot{x}_{i3})}{(\sum \ddot{x}_{i2}^2) \sum \ddot{x}_{i3}^2) - (\sum \ddot{x}_{i2} \ddot{x}_{i3})^2} \\ b_3 &= \frac{(\sum \ddot{y}_i \ddot{x}_{i3})(\sum \ddot{x}_{i2}^2) - (\sum \ddot{y}_i \ddot{x}_{i2})(\sum \ddot{x}_{i2} \ddot{x}_{i3})}{(\sum \ddot{x}_{i2}^2 \sum \ddot{x}_{i3}^2) - (\sum \ddot{x}_{i2} \ddot{x}_{i3})^2} \\ b_1 &= \bar{y} - b_2 \bar{x}_2 - b_3 \bar{x}_3 \end{aligned}$$

wobei wir hier zur einfacheren Darstellung eine neue Notation einführen, zwei Punkte über einer Variable bedeuten, dass von jeder Beobachtung i einer Variable der Mittelwert dieser Variable subtrahiert wurde, d.h. $\ddot{y}_i := (y_i - \bar{y})$, $\ddot{x}_{i2} := (x_{i2} - \bar{x}_2)$ und $\ddot{x}_{i3} := (x_{i3} - \bar{x}_3)$ (siehe auch Abschnitt 2.12.1 Mittelwerttransformationen). Der Laufindex $i = 1, \ldots, n$ kennzeichnet natürlich wieder die einzelne Beobachtung.

Es sei noch angemerkt, dass die OLS Methode natürlich auch mit mehr als zwei erklärenden Variablen funktioniert, allerdings werden die Ausdrücke in Summennotation ziemlich unübersichtlich. Wir werden später zeigen, dass man das multiple Regressionsmodell mit Hilfe von Matrizen sehr viel übersichtlicher anschreiben und auch einfacher lösen kann.

Glücklicherweise sind diese Formeln für die OLS Schätzer in so gut wie allen statistischen Programmpaketen implementiert (selbst in Excel), hier geht es nur darum zu erkennen, dass die Berechnung der OLS-Schätzer im multivariaten Fall nach dem gleichen Grundprinzip erfolgt wie im bivariaten Fall.

Mit mehr als zwei erklärenden Variablen wird das multiple Regressionsmodell häufig geschrieben als

$$y_i = b_1 + b_2 x_{i2} + \ldots + b_h x_{ih} + \ldots + b_k x_{ik} + e_i$$

wobei k die Anzahl der erklärenden Variablen inklusive der Regressionskonstante angibt, und das Interzept b_1 wie üblich der Koeffizient der Regressionskonstanten $x_{i1} = 1$ ist. Für dieses Modell benötigen wir zwei Laufindizes, i als Laufindex über die einzelnen Beobachtungen mit $i = 1, \ldots, n$, und einen Laufindex h über die erklärenden Variablen mit $h = 1, \ldots, k$.

Damit eine Lösung existiert muss die Anzahl der erklärenden Variablen k kleiner (oder gleich) der Anzahl der Beobachtungen n sein, d.h. $k \leq n$, und die erklärenden Variablen müssen untereinander linear unabhängig sein (keine perfekte Multikollinearität).

Zur Verdeutlichung noch einmal ausführlich in Vektornotation

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = b_1 \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} + b_2 \begin{pmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{n2} \end{pmatrix} + \cdots + b_h \begin{pmatrix} x_{1h} \\ x_{2h} \\ \vdots \\ x_{nh} \end{pmatrix} + \cdots + b_k \begin{pmatrix} x_{1k} \\ x_{2k} \\ \vdots \\ x_{nk} \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

Ein wesentlicher Teil des Charmes linearer Regressionsmodelle liegt in der einfachen Interpretation der Koeffizienten als **marginale Effekte**, denn aufgrund der linearen Funktionsform sind die Regressionskoeffizienten einfach die partiellen Ableitungen und können als solche interpretiert werden. Das einzig Neue, was dazukommt, ist die *ceteris paribus* Interpretation.

Für das Regressionsmodell

$$\widehat{y}_i = b_1 + b_2 x_{i2} + b_3 x_{i3}$$

gibt der Regressionskoeffizient b_2 an, um wieviele Einheiten sich $\widehat{y}$ ändert, wenn x_2 um eine Einheit zunimmt *und* x_3 *konstant bleibt*, d.h. *ceteris paribus!* Analoges gilt für b_3

$$b_2 = \left.\frac{d\hat{y}}{dx_2}\right|_{dx_3=0} = \frac{\partial \widehat{y}}{\partial x_2} \qquad \text{und} \qquad b_3 = \left.\frac{d\hat{y}}{dx_3}\right|_{dx_2=0} = \frac{\partial \widehat{y}}{\partial x_3}$$

Diese ceteris-paribus Interpretation wird durch Verwendung des *partiellen Ableitungszeichens* ∂ zum Ausdruck gebracht.

Achtung: Diese ceteris-paribus Interpretation der Koeffizienten gilt nur in Bezug auf die in der Regression berücksichtigten Variablen, nicht für Variablen außerhalb des Modells!

Wenn im Autobeispiel km und Alter auf den Preis regressiert werden bezieht sich die ceteris-paribus Interpretation nur auf km und Alter, wie ändert sich der gefittete Preis bei einer Zunahme der km Zahl bei konstantem Alter, und vice versa, aber nicht in Bezug auf andere Variablen wie z.B. Ausstattungsmerkmale.

Beispiel In einem früheren Abschnitt haben wir den Zusammenhang zwischen dem Preis von Gebrauchtautos und deren Alter untersucht. Natürlich wird der Preis nicht nur vom Alter abhängen, sondern auch von zahlreichen anderen Faktoren, wie zum Beispiel dem Kilometerstand.[10]

Eine Regression des Verkaufspreises auf Alter *und* Kilometerstand gibt

$$\begin{array}{rcccccc} \widehat{\text{Preis}} & = & 22649.884 & - & 1896.264 \text{ Alter} & - & 0.031 \text{ km} \\ & & (411.87)^{***} & & (235.215)^{***} & & (0.008)^{***} \end{array}$$

$$R^2 = 0.907, \qquad n = 40$$

Diese Regression beschreibt den Zusammenhang zwischen Preis und Alter sowie Kilometerstand für 40 Beobachtungen.

Wie früher können wir den gefitteten Preis für ein Auto mit gegebenen Alter und Kilometerstand als lineare Approximation an den Mittelwert dieser Unterkategorie interpretieren, z.B. ist die lineare Approximation für einen Durchschnittspreis von Autos mit einem Alter von vier Jahren und einem Kilometerstand von 100 000 km gleich

$$(\widehat{y}|x_2 = 4, x_3 = 100000) = 22649.884 - 1896.264 * 4 - 0.031 * 100000 = 11963.79$$

[10] Dies ist ein sehr einfaches Beispiel für ein hedonistisches Preismodell (*'hedonic pricing model'*). Dabei wird im wesentlichen der Preis eines Gutes durch seine Eigenschaften erklärt. Weit verbreitet sind solche Preismodelle z.B. für Immobilienmärkte.

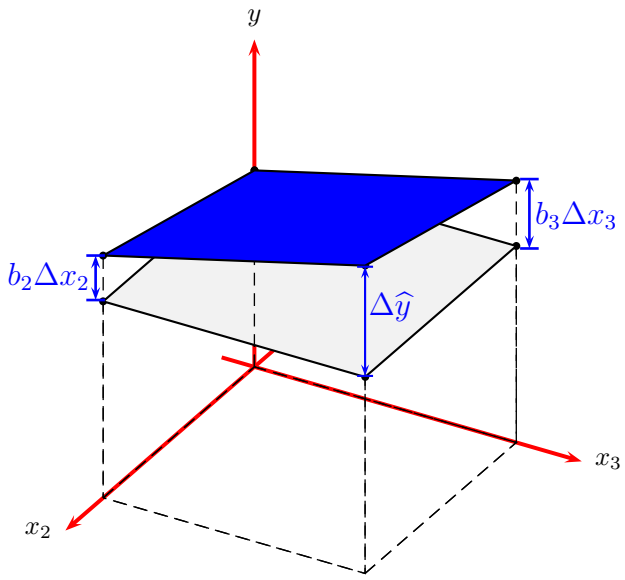


Abbildung 2.16: Ceteris-paribus Interpretation der Koeffizienten;
Für die Regressionsebene $\widehat{y}_i = b_1 + b_2 x_{i2} + b_3 x_{i3}$ folgt durch Bildung erster Differenzen $\Delta\widehat{y} = b_2\Delta x_2 + b_3\Delta x_3$, woraus für die Koeffizienten folgt

$$b_2 = \left.\frac{\Delta\widehat{y}}{\Delta x_2}\right|_{\Delta x_3=0} = \frac{\partial\widehat{y}}{\partial x_2}, \qquad \text{und} \qquad b_3 = \left.\frac{\Delta\widehat{y}}{\Delta x_3}\right|_{\Delta x_2=0} = \frac{\partial\widehat{y}}{\partial x_3}$$

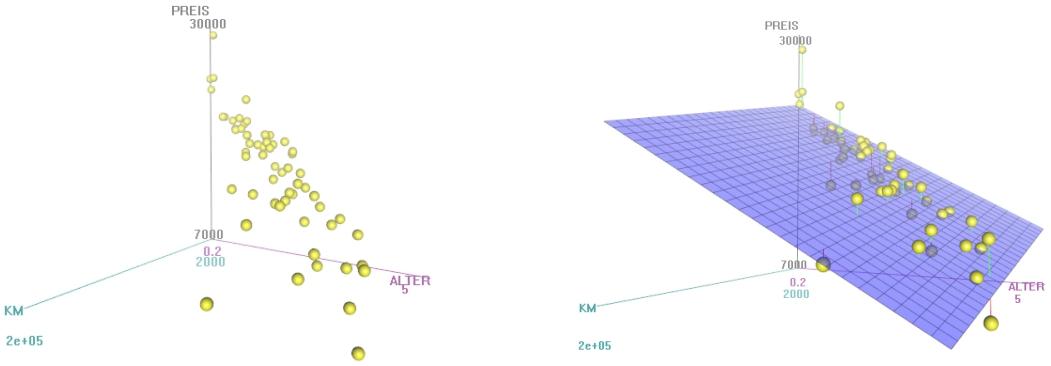


Abbildung 2.17: 3-dimensionale Abbildung des Autobeispiels mit Hilfe des R packages `Rcmdr` (**?**).

wobei $\widehat{y}$ den gefitteten Preis, x_2 das Alter und x_3 den Kilometerstand bezeichnet.

Meist interessieren wir uns aber für die einzelnen Koeffizienten. Das Interzept hat in diesem Fall eine einfache Interpretation, es gibt den durchschnittlichen Wert eines 'gebrauchten Neuwagens' an, d.h. eines Gebrauchtautos mit Alter = 0 und km = 0, allerdings ist das Interzept nur selten von Interesse, weshalb es bei der Interpretation der Ergebnisse nur selten erwähnt wird.

Interessanter sind meistens die Steigungskoeffizienten. Aufgrund dieser Regression würden wir damit rechnen, dass der Preis eines Gebrauchtautos dieser Marke durchschnittlich um 1896 Euro fällt, wenn das Alter um ein Jahr zunimmt *und der Kilometerstand konstant bleibt* (d.h. ceteris paribus)

$$\frac{\partial \widehat{\text{Preis}}}{\partial \text{Alter}} = 1896.264$$

Ebenso müssen wir damit rechnen, dass der Preis mit jedem gefahrenen Kilometer um ca. 0.031 Euro fällt (d.h. um ca. 3 Cent/km bzw. um 31 Euro pro tausend Kilometer), *wenn das Alter unverändert bleibt* (ceteris paribus)

$$\frac{\partial \widehat{\text{Preis}}}{\partial \text{km}} = 0.031$$

Aufgrund der linearen Funktionsform gilt diese Interpretation nicht nur infinitesimal, sondern auch für diskrete Änderungen der erklärenden Variablen. Wenn mit einem 'durchschnittlichen' Auto z.B. über einen Zeitraum von zwei Jahren 30000 Kilometer zurücklegt werden, muss aufgrund dieser Regression mit einem durchschnittlichen Wertverlust von $1896.264 \times 2 + 0.031 \times 30000 = 4722.838$ Euro gerechnet werden. Aber natürlich bezieht sich dies nicht auf die tatsächlichen Durchschnittspreise, sondern auf die auf der Regressionsebene liegenden gefitteten Preise $\widehat{y}$ (d.h. auf die lineare Approximation).

Um die ceteris paribus Interpretation zu betonen sagt man manchmal auch, dass im multiplen Regressionsmodell für den Einfluss der anderen erklärender Variablen *kontrolliert* wird, d.h. der Koeffizient des Alters misst den durchschnittlichen Wertverlust pro Jahr, wenn für den Kilometerstand kontrolliert wird. Dieser Sprachgebrauch geht auf die experimentellen Ursprünge der Regressionsanalyse zurück.

In dieser ceteris-paribus Interpretation der Koeffizienten als marginale Effekte liegt ein großer Vorteil des multiplen Regressionsmodells, es erlaubt die Kontrolle mehrerer Einflussfaktoren, die gleichzeitig auf die abhängige Variable y einwirken. Diese ceteris paribus Interpretation der Koeffizienten ist natürlich auch dann gültig, wenn die Daten nicht auf eine ceteris paribus Art erhoben wurden. Um z.B. die isolierten Einflüsse des Alters auf den Preis *bei konstantem Kilometerstand* zu ermitteln benötigen wir keine Daten von Autos mit unterschiedlichem Alter und *gleichem Kilometerstand*, aufgrund der linearen Funktionsform können die marginalen ceteris paribus Effekte selbst dann berechnet werden, wenn jede Alter – Kilometerstand Kombination nur einmalig beobachtet wird.

Die lineare Regression ermöglicht deshalb auch für nichtexperimentelle Daten eine ceteris paribus Interpretation der Koeffizienten.

Aber hier gilt es zwei wichtige Punkt zu beachten:

1. Diese ceteris-paribus Interpretation bezieht sich ausschließlich auf die *im Modell explizit berücksichtigten* x Variablen! Wenn Sie z.B. eine Lohngleichung schätzen und Ausbildung, Berufserfahrung und Geschlecht berücksichtigen, dann bezieht sich die ceteris paribus Interpretation *nur* auf diese berücksichtigten Variablen Bildung, Erfahrung und Geschlecht, aber nicht auf *nicht berücksichtigte* (und oft unbeobachtbare) Variablen wie z.B. Qualität der Bildung, soziale Fähigkeiten, Durchsetzungsfähigkeit, Intelligenz, usw.

   Darin liegt ein zentraler Unterschied zu *Randomisierten kontrollierten Experimenten* (RCT), die zumindest prinzipiell eine indirekte 'Kontrolle' auch der unbeobachteten Faktoren erlaubt. Mehr dazu im Kapitel zu *Endogenität*.

2. Der marginale Effekt bezieht sich auf den systematischen Teil der Regression $\widehat{y}$, also auf die lineare Approximation an den bedingten Mittelwert, nicht auf das y_i eines Individuums. Es ist verlockend sich den marginalen Effekt als Auswirkung auf y vorzustellen, aber dies ist falsch. $\widehat{y}_i$ ist ein bedingter Mittelwert und verrät uns nicht mehr über eine einzelne Beobachtung wie uns das Durchschnittseinkommen eines Landes über des Einkommen einer Einzelperson dieses Landes verrät!

   Wie wir gleich sehen werden führt eine Fehlspezifikation (z.B. die Nicht-Berücksichtigung einer relevanten Variable) in der Regel dazu, dass ein Koeffizient die Auswirkungen auf y nicht korrekt misst. Dies wird immer dann der Fall sein, wenn eine Veränderung von x auch den Fehlerprozess e beeinflusst. Mehr dazu später im Abschnitt zur 'Nichtberücksichtigung relevanter Variablen' und vor allem im späteren Kapitel zur Endogenität.

Aber selbstverständlich ist die ceteris paribus Interpretation auch dann zulässig, wenn die erklärenden Variablen untereinander korreliert sind, wie dies z.B. in unserem Beispiel mit Kilometerstand und Alter der Autos zu erwarten ist.

Möglich wird diese ceteris paribus Interpretation allerdings ausschließlich durch die Annahme der Funktionsform. Falls die Daten durch eine lineare Funktionsform nur sehr schlecht approximiert werden oder wesentliche erklärende Variablen fehlen wird diese Interpretation zu irreführenden Schlussfolgerungen führen.

Tatsächlich haben wir durch die Wahl der linearen Funktionsform die Daten gewissermaßen auf das Prokrustes-Bett[11] unserer Spezifikation gespannt; dazu werden wir später mehr zu sagen haben.

Man beachte außerdem, dass wir bisher nur die 'durchschnittlichen' Zusammenhänge für die gegebenen 40 Beobachtungen beschrieben haben, es handelte sich bisher also um eine rein deskriptive Analyse.

Das korrigierte Bestimmtheitsmaß $\overline{R}^2$ (*adjusted* R^2): Alles, was früher über das Bestimmtheitsmaß R^2 gesagt wurde, gilt auch für das multiple Regressionsmodell, falls die Regression ein Interzept enthält ist das R^2 der Anteil der durch

[11] Prokrustes – eine Figur aus der griechischen Mythologie – war bekannt dafür Reisenden ein Bett anzubieten, und die unglücklichen Wanderer dann mit Brachialgewalt an die Größe des Bettes 'anzupassen'. War der Wanderer groß hackte er ihm die Füße ab, war der Wanderer klein zog er ihn in die Länge.

Eine stochastische Interpretation: In den allermeisten Fällen werden Regressionen verwendet, um Schlüsse auf eine unbeobachtete Grundgesamtheit zu tätigen, also im Sinne der *induktiven Statistik*. Die Details dazu werden wir in den folgenden drei Kapiteln ausführlich erläutern, hier nur eine kurze Vorschau.
Die Zahlen in Klammern unter den Koeffizienten sind (wenn nichts anderes erwähnt wird) die Standardfehler der Koeffizienten, also ein Maß für die Genauigkeit der Messung.
Als *grobe Faustregel* können Sie sich merken, dass der Absolutbetrag der Koeffizienten mindestens doppelt so groß sein sollte wie der darunterstehende Standardfehler (für $n > 30$). Dies wird häufig durch zwei oder drei Sterne neben den Koeffizienten oder Standardfehlern kenntlich gemacht (meistens bedeutet ein Stern, dass der Koeffizient auf einem Signifikanzniveau von 10% von Null verschieden ist, zwei Sterne 5% und drei Sterne ein Signifikanzniveau von 1%).

$$\begin{array}{rcccccc} \widehat{\text{Preis}} & = & 22649.884 & - & 1896.264\ \text{Alter} & - & 0.031\ \text{km} \\ & & (411.87)^{***} & & (235.215)^{***} & & (0.008)^{***} \end{array}$$

$$R^2 = 0.907, \qquad n = 40$$

Für die Interpretation kann man in folgenden Schritten vorgehen:

1. Welcher Zusammenhang wird dargestellt, und was erwarten wir zu sehen? Stimmen die Vorzeichen mit unseren Erwartungen überein?
2. Sind die Koeffizienten statistisch signifikant von Null verschieden? Wenn nein, interpretieren Sie diese Koeffizienten nicht weiter.
3. Interpretieren Sie die quantitative Bedeutung der statistisch signifikanten Koeffizienten unter Beachtung der ceteris paribus Annahme. Das Interzept wird nur interpretiert, wenn besondere Gründe dafür sprechen.
4. Erwähnen Sie kurz die Güte der Anpassung (R^2) und inwiefern wir vermuten können, dass die der Schätzung zugrunde liegenden Annahmen erfüllt sind (die Details dazu folgen in den nächsten Kapiteln).

Die obige Regressionsgleichung zeigt die Abhängigkeit des Preises einer bestimmten Type von Gebrauchtautos in Abhängigkeit von deren Alter und Kilometerzahl. Wir würden a priori erwarten, dass der durchschnittliche Preis mit zunehmendem Alter und zunehmender Kilometerzahl fällt. Die Gleichung zeigt, dass dies tatsächlich der Fall ist, beide Vorzeichen sind negativ.
Wenn wir die Schlussfolgerungen aus dieser *Stichprobe* (mit $n = 40$) verallgemeinern wollen, müssen wir mögliche Stichprobenfehler berücksichtigen. Da das Verhältnis von Koeffizient zu Standardfehler derart klein ist können wir in diesem Fall fast ausschließen, dass es sich hierbei lediglich um ein Zufallsresultat handelt (Details dazu folgen später). Die Wahrscheinlichkeit dafür, dass in der Grundgesamtheit *kein* Zusammenhang zwischen Preis und Alter bzw. Kilometerzahl besteht ist in diesem Fall jeweils kleiner als ein Prozent.
Daraus schließen wir, dass der durchschnittliche Preis mit jedem Altersjahr bei konstanter Kilometerzahl (d.h. ceteris paribus) um ca. 1900 Euro fällt, und dass der Preis mit jedem zusätzlichen Kilometer bei konstantem Alter um ca. 3 Cent abnimmt. Diese beiden Variablen erklären ca. 90% der Streuung der Preise, die Anpassung der Regressionsgerade ist also sehr gut.

alle x Variablen gemeinsam erklärten Streuung an der Gesamtstreuung von y (d.h. ESS/TSS).

Ein kleines Problem gibt es allerdings im multiplen Regressionsmodell: weil die Streuung (Varianz) nie negativ werden kann, wird durch die Einbeziehung eines weiteren Regressors das R^2 immer größer werden (oder zumindest nie kleiner werden). Dies ist einleuchtend, durch die Einbeziehung eines zusätzlichen Regressors kann der Fit nie schlechter werden. Deshalb eignet sich das übliche Bestimmtheitsmaß nicht für einen Vergleich von Regressionen mit einer unterschiedlichen Anzahl von erklärenden x Variablen.

Mit dem korrigierten Bestimmtheitsmaß $\overline{R}^2$ wird versucht dieses Problem zumindest zu zu mildern, indem ein Korrekturfaktor eingeführt wird.

$$\text{normal: } R^2 = 1 - \frac{\sum_i e_i^2}{\sum_i (y_i - \overline{y})^2}$$

$$\text{korrigiert: } \overline{R}^2 = 1 - \frac{\frac{\sum_i e_i^2}{(n-k)}}{\frac{\sum_i (y_i - \overline{y})^2}{(n-1)}} = 1 - \left(\frac{\sum_i e_i^2}{\sum_i (y_i - \overline{y})^2} \right) \left(\frac{n-1}{n-k} \right)$$

Mit einer zunehmenden Zahl erklärender Variablen k wird der Faktor $(n-1)/(n-k)$ größer und kompensiert damit dafür, dass $\sum_i e_i^2$ mit zunehmendem k kleiner wird. Deshalb eignet sich das korrigierte Bestimmtheitsmaß $\bar{R}^2$ eher für einen Vergleich zweier Regressionen mit einer unterschiedlichen Anzahl erklärender Variablen.

Im Rahmen der stochastischen Regressionsanalyse werden wir später sehen, dass die Quadratsumme der Residuen $(n-k)$ *Freiheitsgrade* hat, während die Gesamtstreuung von y im Nenner $(n-1)$ *Freiheitsgrade* hat, deshalb kann man sich als Merkhilfe vorstellen, dass für das korrigierte Bestimmtheitsmaß $\bar{R}^2$ die entsprechenden Streuungen einfach um die Freiheitsgrade bereinigt werden.

Alternative – und theoretisch besser fundierte – Kennzahlen für die Modellselektion sind das *Akaike Informationskriterium* (AIC) und das *Bayessche Informationskriterium* (BIC), die v.a. in der Zeitreihenökonometrie häufig angewandt werden.

2.6.1 Nichtberücksichtigung relevanter Variablen

Kehren wir nochmals zu unserem Beispiel mit den Gebrauchtautos zurück. Die multiple Regression zur Erklärung des Preises ist Preis $= b_1 + b_2$Alter $+ b_3$km $+ e$; Spalte (1) von Tabelle 2.5 zeigt zu Vergleichszwecken noch einmal das Ergebnis dieser Schätzung. Spalte (2) zeigt das Ergebnis einer Regression *nur* auf das Alter, und Spalte (3) das Ergebnis einer Regression *nur* auf den Kilometerstand. Nachdem diese beiden Regressionen weniger erklärende Variablen haben werden wir diese 'kurze' Modelle nennen.

In den beiden 'kurzen' Modellen (2) und (3) erhalten wir absolut gesehen deutlich größere Steigungskoeffizienten als die im 'langen' (multiplen) Modell (1). Was ist passiert?

Wenn wir *nur* auf das Alter regressieren misst der Steigungskoeffizient nicht nur den Einfluss des Alters, sondern indirekt auch den Einfluss des nicht berücksichtigten

Tabelle 2.5: Preise von Gebrauchtautos.

Abh.Var.: Preis	(1)	(2)	(3)
Const.	22 649.884	23 056.714	20 279.226
Alter	−1 896.264	−2 635.669	
km	−0.031		−0.082
R^2	0.907	0.868	0.743
n	40	40	40

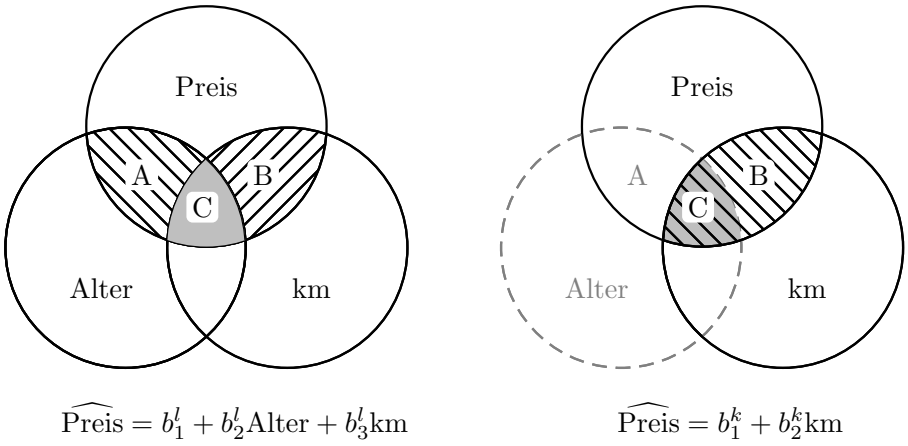


Abbildung 2.18: 'Langes' und 'kurzes' Modell; Im 'langen' Modell (linkes Panel) geht die Überschneidungsfläche C nicht in die Schätzung der Steigungskoeffizienten ein. Falls das Alter fälschlich nicht berücksichtigt wird geht die Fläche C in die Schätzung des Koeffizienten für den Kilometerstand ein (*'Omitted Variables Bias'*, rechtes Panel).

Kilometerstands. Da das Alter und der Kilometerstand von Gebrauchtautos üblicherweise positiv korreliert sind, überschätzen wir den Einfluss des Alters, ein Teil des Preisverlusts ist auf den durchschnittlich höheren Kilometerstand älterer Autos zurückzuführen.

Einen intuitiven Einblick gibt das Venn Diagramm in Abbildung 2.18. Die Streuung der Variablen Preis, Alter und Kilometerstand wird durch Kreise symbolisiert, und die Korrelation zwischen den Variablen durch die Überschneidungen der Kreise.

Im korrekt spezifizierten Modell (linkes Panel) geht die Fläche A in die Schätzung des Koeffizienten für das Alter ein und die Fläche B in die Schätzung des Koeffizienten für den Kilometerstand. Die Überschneidungsfläche C, die aus der Korrelation zwischen Alter und Kilometerstand resultiert, kann nicht klar einer der Variablen zugeordnet werden, und geht deshalb nicht in die Schätzung der Steigungskoeffizienten ein (sehr wohl aber in das R^2).

Anders im Fall des falsch spezifizierten Modell im rechten Panel. Wenn das Alter nicht als erklärende Variable berücksichtigt wird, gehen die Flächen B *und* C in die Schätzung des Koeffizienten für den Kilometerstand ein, die Fläche C zumindest teilweise zu unrecht, da diese auch dem nicht berücksichtigten Alter zuzuschreiben

ist.

Dies gibt dem Kilometerstand fälschlich eine größere Bedeutung als ihm eigentlich zukommt, da er zum Teil auch den Effekt des nicht berücksichtigten Alters mit einfängt! Die Folgen sind gravierend, der Koeffizient des Kilometerstands misst nicht länger den korrekten marginalen Effekt, sondern ist gewissermassen 'verschmutzt' durch die fälschlich *nicht* berücksichtigte Variable Alter (man beachte, dass die ceteris paribus Interpretation nur für die im Modell berücksichtigten Variablen gilt). Deshalb erhalten wir einen weit überhöhten Preisverlust von 8 Cent pro Kilometer anstelle der 3 Cent des 'langen' Modells, die bei einer Berücksichtigung von Kilometerstand *und* Alter resultieren.

Analoges gilt, wenn wir nur auf das Alter regressieren und den Kilometerstand nicht berücksichtigen. In diesem Fall würden wir einen Teil des Preisverlustes, der eigentlich Kilometerstand zuzuschreiben ist, zu unrecht dem Alter zuschreiben.

Dieses Problem ist in die Literatur als *'Omitted Variables Bias'* bekannt und wird uns später im Rahmen der stochastischen Regressionsanalyse noch ausführlich beschäftigen. Hier sei nur vorausgeschickt, dass ein *'Omitted Variables Bias'* nur dann auftreten kann, wenn die nicht berücksichtigte Variable sowohl mit der abhängigen Variable y als auch mit dem berücksichtigten Regressor x korreliert ist.

Das linke Panel des Venn Diagramms in Abbildung 2.18 kann uns noch eine weitere Einsicht vermitteln. Wenn die Regressoren Alter und Kilometerstand sehr hoch korreliert sind führt dies dazu, dass die Überschneidungsfläche C sehr groß wird, und die Flächen A und B entsprechend klein werden. Da aber nur die die Flächen A und B in die Schätzung der Koeffizienten eingehen, wird die Schätzung entsprechend ungenau, dies führt im wesentlichen zum gleichen Problem wie eine (zu) kleine Stichprobe. Dieses Problem einer hohen Korrelation zwischen den erklärenden Variablen wird in der Ökonometrie *Multikollinearität* genannt.

Im Extremfall, wenn die Regressoren Alter und Kilometerstand perfekt korreliert sind (d.h. linear abhängig sind) liegen die Kreise für Alter und Kilometerstand aufeinander, und die Koeffizienten können nicht mehr einzeln geschätzt werden, bzw. sind nicht mehr definiert. Dieser Extremfall wird *perfekte Multikollinearität* genannt. Auch diese Fälle von Multikollinearität werden wir in einem späteren Kapitel noch ausführlich diskutieren.

Zuerst wollen wir aber das Problem fehlender relevanter Variablen noch etwas näher beleuchten und zeigen, was bei der Nichtberücksichtigung relevanter Variablen passiert.

Die Algebra der Nichtberücksichtigung relevanter Variablen

Wir starten mit dem einfachsten multiplen Regressionsmodell, wobei wir alle Variablen mittelwerttransformieren, d.h. $\ddot{x}_i := x_i - \bar{x}$ (siehe Abschnitt 2.12.1). Durch die Mittelwerttransformation fällt das Interzept weg, was die folgende Darstellung vereinfacht (um die Lesbarkeit zu erhöhen verzichten wir zudem auf den Beobachtungsindex i).

Wir vergleichen nun die Koeffizienten eines *langen* Modells (durch einen hochgestellten Index l gekennzeichnet)

$$\ddot{y} = b_2^l \ddot{x}_2 + b_3^l \ddot{x}_3 + e^l$$

mit dem Steigungskoeffizienten eines 'kurzen' Modells, in dem $\ddot{x}_3$ nicht berücksichtigt wird

$$\ddot{y} = b_2^k \ddot{x}_2 + e^k$$

Wir nehmen an, dass das *lange* Modell den datengenerierenden Prozess korrekt abbildet, und dass das *kurze* Modell *fehlspezifiziert* ist.

Der OLS Steigungskoeffizient des fehlspezifizierten 'kurzen' Modells ist

$$b_2^k = \frac{\operatorname{cov}(x_2, y)}{\operatorname{var}(x_2)} = \frac{\sum \ddot{x}_2 \ddot{y}}{\sum \ddot{x}_2^2}$$

Um die Konsequenzen dieser Fehlspezifikation zu erkennen (d.h. die Nichtberücksichtigung von $\ddot{x}_3$) setzen wir in die obige OLS-Formel für den Steigungskoeffizienten des kurzen Modells b_2^k für $\ddot{y}$ das korrekt spezifizierte 'lange' Modell $\ddot{y} = b_2^l \ddot{x}_2 + b_3^l \ddot{x}_3 + e^l$ ein und vereinfachen

$$\begin{aligned} b_2^k &= \frac{\sum \ddot{x}_2 (b_2^l \ddot{x}_2 + b_3^l \ddot{x}_3 + e^l)}{\sum \ddot{x}_2^2} \\ &= \frac{\sum \ddot{x}_2 b_2^l \ddot{x}_2 + \sum \ddot{x}_2 b_3^l \ddot{x}_3 + \sum \ddot{x}_2 e^l}{\sum \ddot{x}_2^2} \\ &= \frac{b_2^l \sum \ddot{x}_2^2 + b_3^l \sum \ddot{x}_2 \ddot{x}_3 + \sum \ddot{x}_2 e^l}{\sum \ddot{x}_2^2} \\ &= b_2^l + b_3^l \frac{\sum \ddot{x}_2 \ddot{x}_3}{\sum \ddot{x}_2^2} + \frac{\sum \ddot{x}_2 e^l}{\sum \ddot{x}_2^2} \end{aligned}$$

Aufgrund der Bedingungen erster Ordnung ist $\sum_i \ddot{x}_{i2} e_i^l = 0$, deshalb gilt

$$b_2^k = b_2^l + b_3^l \frac{\sum \ddot{x}_2 \ddot{x}_3}{\sum \ddot{x}_2^2} = b_2^l + b_3^l \frac{\operatorname{cov}(x_2, x_3)}{\operatorname{var}(x_2)} \tag{2.12}$$

Es gibt also einen einfachen Zusammenhang zwischen den Steigungskoeffizienten des 'kurzen' und 'langen' Modells.

Kommt Ihnen der Ausdruck $\operatorname{cov}(x_2, x_3)/\operatorname{var}(x_2)$ bekannt vor? Genau, dies ist die OLS Formel für den Steigungskoeffizienten einer Regression von x_3 auf x_2

$$x_3 = a_1 + a_2 x_2 + e^*, \quad \Rightarrow \quad a_2 = \frac{\operatorname{cov}(x_2, x_3)}{\operatorname{var}(x_2)}$$

wobei e^* wie üblich die Residuen dieser Regression bezeichnet.

Deshalb können wir den Zusammenhang zwischen den Steigungskoeffizienten des 'kurzen' und 'langen' Modells einfacher schreiben als

$$\boxed{b_2^k = b_2^l + b_3^l \frac{\operatorname{cov}(x_2, x_3)}{\operatorname{var}(x_2)} = b_2^l + b_3^l \, a_2} \tag{2.13}$$

Wenn – und nur wenn – b_3^l und a_2 *gleichzeitig* von Null verschieden sind, führt die Nichtberücksichtigung von x_3 dazu, dass sich die Koeffizienten des 'kurzen' und 'langen' Modells unterscheiden werden.

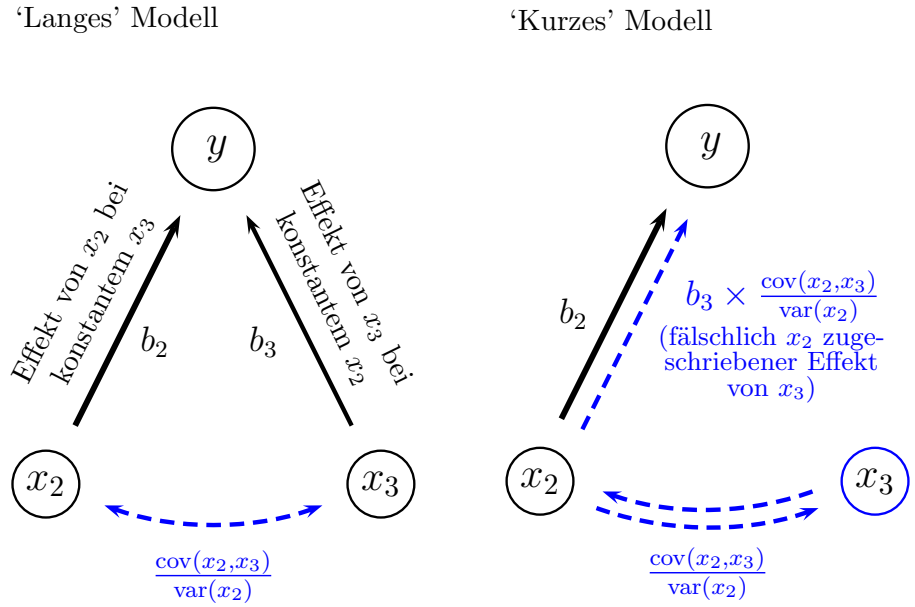


Abbildung 2.19: Nichtberücksichtigung einer relevanten Variable x_3 führt dazu, dass ein Teil der Auswirkungen von x_3 fälschlich x_2 zugeschrieben wird. Wenn das 'wahre' Modell $y = b_1 + b_2^l x_2 + b_3^l x_3 + e^l$ ist und irrtümlich ein kurzes Modell $y = b_1^k + b_2^k x_2 + e^k$ geschätzt wird ist $b_2^k = b_2^l + b_3^l \frac{\text{cov}(x_2,x_3)}{\text{var}(x_2)}$.

Abbildung 2.19 zeigt das Problem noch einmal: wenn x_3 nicht berücksichtigt wird, wird x_2 neben seiner direkten Wirkung b_2^l auch noch fälschlich ein Teil der Wirkung von x_3 zugeschrieben, da x_2 als Proxy für x_3 wirkt. Die Größe dieses 'Proxy-Effekts' hängt von zwei Faktoren ab: erstens vom Effekt von x_3 auf y, also von b_3^l, und zweitens von dem Zusammenhang zwischen x_2 und x_3.

Für den Fall mit mehreren nicht berücksichtigten Variablen sind die Formeln etwas komplexer, aber die Essenz bleibt erhalten.

Beispiel: Was bedeutet das nun für unser Beispiel mit den Gebrauchtautos? In Tabelle (2.5) haben wir die Schätzung für ein 'langes' und für zwei 'kurze' Modelle. Um den Zusammenhang zu demonstrieren beschränken uns auf das 'kurze' Modell mit dem Alter.

Zur Erinnerung, das 'lange' Modell aus Tabelle (2.5) war

$$\widehat{\text{Preis}} = 22649.884 - 1896.264\ \text{Alter} - 0.031\ \text{km}$$
$$R^2 = 0.907, \qquad n = 40$$

und die Hilfsregression $\text{km} = a_1 + a_2\,\text{Alter} + v$ ist

$$\widehat{\text{km}} = -13119.185 + 23843.819\,\text{Alter}, \qquad R^2 = 0.6357,\ n = 40$$

Den Steigungskoeffizienten des 'kurzen' Modells aus Spalte (2) von Tabelle (2.5) erhalten wir alternativ auch aus $b_2^l + b_3^l \times a_2 = -1896.264 - 0.031 \times 23843.819 = -2635.669 = b_2^k$ (kleine Abweichungen sind auf Rundungsfehler zurückzuführen).

Tabelle 2.6: Gleichung (2.12) erlaubt eine Abschätzung der Richtung des Fehlers bei der Schätzung eines 'kurzen' Modells $y = b_1^k + b_2^k x_2 + e^k$ anstelle eines 'langen' Modells $y = b_1 + b_2^l x_2 + b_3^l x_3 + e^l$.
Da $b_2^k = b_2^l + b_3^l \times \mathrm{cov}(x_2, x_3)/\,\mathrm{var}(x_2)$ gilt:

	$\mathrm{cov}(x_2, x_3) > 0$	$\mathrm{cov}(x_2, x_3) < 0$
$b_3^l > 0$	$b_2^k > b_2^l$	$b_2^k < b_2^l$
$b_3^l < 0$	$b_2^k < b_2^l$	$b_2^k > b_2^l$

Wozu war das nun alles gut? Die ganze Tragweite dieses Resultats wird erst später im Rahmen der stochastischen Regressionsanalyse deutlich werden; dort werden wir sehen, dass die Nichtberücksichtigung relevanter Variablen zu *endogenen Regressoren* führt und einen *"Omitted Variable Bias"* verursacht.

Aber bereits jetzt erlaubt uns dieses Resultat die Abschätzung eines möglichen 'Fehlers'. Ob der Steigungskoeffizient des 'langen' Modells größer oder kleiner als der Steigungskoeffizient des 'kurzen' Modells ist hängt nämlich nur vom Vorzeichen des Ausdrucks $b_3^l \times a_2$ ab.

Angenommen, wir hätten keine Daten über den Kilometerstand der Autos gesammelt und nur Preise und Alter der Autos. Wir vermuten, dass der Preis mit zunehmender Kilometerzahl fällt (d.h. $b_3^l < 0$), und das Kilometerzahl und Alter positiv korreliert sind (d.h. $a_2 > 0$, bzw. $\mathrm{cov}(\mathrm{km}, \mathrm{Alter}) > 0$). Da $b_2^k = b_2^l + b_3^l \times a_2$ und $b_3^l \times a_2 < 0$ folgt $b_2^k < b_2^l$, der Einfluss des Alters auf den Preis wird in der 'kurzen' Regression also vermutlich überschätzt (man beachte, dass die Koeffizienten negativ sind, also $b_2^k = -2635 < -1896 = b_2^l$).

2.6.2 Das Frisch-Waugh-Lovell (FWL) Theorem

Bereits in der allerersten Ausgabe der Econometrica (1933) haben Ragnar Frisch und Frederick V. Waugh auf eine interessante Eigenschaft des multiplen Regressionsmodells hingewiesen, die uns auch ein tieferes Verständnis für die Interpretation der Regressionskoeffizienten geben kann.

Dieses Ergebnis wurde später von Michael C. **?** verallgemeinert; er zeigte, dass dies auch für Gruppen von Variablen gilt. Seither ist dieses Resultat als *Frisch-Waugh-Lovell* (FWL) Theorem bekannt.

Im wesentlichen zeigt das FWL Theorem, dass ein interessierender Koeffizient einer multiplen Regression alternativ auch mit Hilfe mehrerer bivariater (kurzer) Regressionen berechnet werden kann.

Als **?** dieses Ergebnis bewiesen waren Computer noch kaum verfügbar, und weil multiple Regressionen weit aufwändiger zu berechnen waren als bivariate Regressionen hatte dieses Ergebnis damals durchaus praktische Bedeutung. Heute ist Rechenzeit billig, trotzdem ist dieses Resultat immer noch wichtig. Es gestattet uns tiefere Einsichten in die 'OLS-Mechanik', trägt zum Verständnis der Regressionskoeffizienten in multiplen Regressionen bei, und hat zahlreiche Anwendungen in fortgeschrittenen Bereichen der Ökonometrie, z.B. in der Panelökonometrie.

Konkret besagt das FWL Theorem folgendes: wenn uns z.B. der Koeffizient b_2 der multiplen Regression

$$y = b_1 + b_2 x_2 + b_3 x_3 + e$$

interessiert, können wir diesen alternativ auch mit Hilfe von drei bivariaten Regressionen berechnen.

Zuerst regressieren wir die beiden interessierenden Variblen y und x_2 auf die zu eliminierende Variable x_3

$$\begin{aligned} y &= c_1 + c_2 x_3 + e_y \\ x_2 &= a_1 + a_2 x_3 + e_{x_2} \end{aligned}$$

wobei e_y die Residuen der ersten bivariaten Gleichung und e_{x_2} die Residuen der zweiten bivariaten Gleichung bezeichnet.

Man beachte, dass die Residuen jeweils den um den linearen Einfluss von x_3 'bereinigten' Effekt enthalten:[12] $e_y = y - (y|x_3)$ und $e_{x_2} = x_2 - (x_2|x_3)$

$$e_y = b_2\, e_{x_2} + e$$

Das FWL Theorem garantiert, dass eine *bivariate* Regression dieser beiden Residuen exakt den gleichen Koeffizienten b_2 und die gleichen Residuen e wie die ursprüngliche lange Regression liefert!

Dabei machen wir uns zunutze, dass OLS eine eine Zerlegungsmethode ist; in den Residuen der beiden 'kurzen' Regressionen auf x_3 wurde der (lineare) Einfluss von x_3 auf y bzw. x_2 eliminiert. Im Englischen wird dies häufig *'partialling out'* genannt. Wie schon erwähnt wurde dieses Resultat von **?** für mehrere Variablen verallgemeinert.

Beweis: Der Beweis dieses Theorems erfolgt üblicherweise unter Zuhilfenahme von Matrixalgebra. Wir werden hier einen deutlich einfacheren Beweis skizzieren, der **?** folgt.

Unser Ausgangspunkt ist wieder die multiple Regression

$$y_i = b_1 + b_2 x_{i2} + b_3 x_{i3} + e_i \tag{2.14}$$

Die folgenden Ausführungen beruhen auf zwei Eigenschaften der OLS Methode:

1. Die erklärenden Variablen x_2 und x_3 sind per Konstruktion mit den Residuen e unkorreliert. Dies folgt unmittelbar aus den Bedingungen erster Ordnung $\sum_{i=1}^{n} x_{ih} e_i = 0$ für alle $h = 2, \ldots, k$.

2. Wenn eine erklärende x Variable weder mit der abhängigen Variable y noch mit den restlichen erklärenden x Variablen korreliert ist, dann ist der Koeffizient dieser Variable gleich Null. Wenn z.B. in Gleichung (2.14) $\mathrm{cov}(y, x_3) = 0$ und $\mathrm{cov}(x_2, x_3) = 0$ ist, dann folgt $b_3 = 0$.

[12] wir erinnern uns, dass der systematische Teil $\widehat{y}$ die bedingten Mittelwerte sind, also $\widehat{y} = c_1 + c_2 x_3 = (y|x_3)$.

Wir beginnen damit, die abhängige Variable y und die erklärende Variable x_2 mittels zweier OLS Hilfsregressionen in die durch x_3 erklärte systematische Komponente und die Residuen zu zerlegen

$$y_i = c_1 + c_2 x_{i3} + e_{i;y} \tag{2.15}$$
$$x_{i2} = a_1 + a_2 x_{i3} + e_{i;x_2} \tag{2.16}$$

Man beachte, dass aufgrund der Bedingungen erster Ordnung $\text{cov}(x_3, e_y) = 0$ und $\text{cov}(x_3, e_{x_2}) = 0$.

Wir setzen diese beiden Gleichungen in die lange Gleichung (2.14) ein und erhalten

$$c_1 + c_2 x_{i3} + e_{i;y} = b_1 + b_2(a_1 + a_2 x_{i3} + e_{i;x_2}) + b_3 x_{i3} + e_i$$

daraus folgt nach umstellen

$$\begin{aligned} e_{i;y} &= (b_1 - c_1) + b_2(a_1 + a_2 x_{i3} + e_{i;x_2}) - c_2 x_{i3} + b_3 x_{i3} + e_i \\ &= (b_1 - c_1 + b_2 a_1) + b_2 e_{i;x_2} + (b_2 a_2 - c_2 + b_3) x_{i3} + e_i \end{aligned}$$

Wenn eine erklärende Variable *weder* mit der abhängigen Variable (y) *noch* mit einer anderen erklärenden Variable (x_2) korreliert ist, muss der Koeffizient dieser Variable Null sein.

Aus den Bedingungen erster Ordnung der Gleichungen (2.15) und (2.16) wissen wir aber, dass $\text{cov}(x_3, e_y) = 0$ (Gleichung 2.15) und dass $\text{cov}(x_3, e_{x_2}) = 0$ (Gleichung 2.16), deshalb muss der Koeffizient von x_3 gleich Null sein, d.h. $b_2 a_2 - c_2 + b_3 = 0$. Deshalb ist

$$e_{i;y} = (b_1 - c_1 + b_2 a_1) + b_2\, e_{i;x_2} + e_i$$

Zudem wissen wir bereits, dass bei einer Regression von mittelwerttransformierten Variablen das Interzept gleich Null ist. In unserem Fall sind sowohl die abhängige Variable $e_{i;y}$ als auch die erklärende Variable $e_{i;x_2}$ Residuen aus Regressionen mit einem Interzept, deshalb muss deren Mittelwert gleich Null sein (Bedingung erster Ordnung!), die Residuen sind also bereits mittelwerttransformiert. Aus diesem Grund ist das Interzept ebenfalls Null ($b_1 - c_1 + b_2 a_1 = 0$) und wir erhalten als Resultat

$$e_{i;y} = b_2\, e_{i;x_2} + e_i$$

Man beachte, dass b_2 aus dieser Gleichung exakt dem b_2 aus 'langen' Regression (2.14) entspricht, das heißt, wir erhalten bei einer Regression der Residuen der beiden Hilfsregressionen (2.15) und (2.16) exakt den gleichen Koeffizienten b_2 und auch die gleichen Residuen e_i wie aus der 'langen' Regression (2.14). ■

Wir können deshalb sagen, dass der Koeffizient b_2 der 'langen' Regression (2.14) die Auswirkungen von x_2 auf y, beschreibt, nachdem der lineare Einfluss von x_3 eliminiert wurde, oder in andern Worten, *nachdem für x_3 kontrolliert wurde.*

Wir haben bereits erwähnt, dass dieses Theorem allgemeiner gilt, es kann auch der lineare Einfluss mehrerer Variablen eliminiert werden, indem man in den Hilfsregressionen auf diese Gruppe von Variablen regressiert.

Beispiel: Wir können dieses Ergebnis wieder anhand des Beispiels mit den Gebrauchtautos demonstrieren. Wir verwenden zwei Hilfsregressionen, um den linearen Einfluss der Kilometer auf den Preis und das Alter zu eliminieren.

Dazu berechnen wir die Residuen der beiden Gleichungen

$$\begin{aligned} \text{Preis} &= a_1 + a_2\,\text{km} + e_P \quad \rightarrow \quad e_P \\ \text{Alter} &= c_1 + c_2\,\text{km} + e_A \quad \rightarrow \quad e_P \end{aligned}$$

und regressieren dann (ohne Interzept!)

$$e_P = b_2\, e_A + e$$

In R kann dies z.B. mit dem Code in Script 2.1 bewerkstelligt werden (bzw. in STATA siehe Script 2.2).

Script 2.1: Beispiel zu Frisch-Waugh-Lovell Theorem, `R-Code`

```
rm(list=ls(all=TRUE))
d <- read.csv2("http://www.hsto.info/econometrics/data/auto40.csv")

res_Preis <- resid(lm(Preis ~ km, data = d))
res_Alter <- resid(lm(Alter ~ km, data = d))
eq_res <- lm(res_Preis ~ res_Alter -1)  # ohne Interzept!

eq_res
# Coefficients:
# res_Alter
#     -1896

eq_long <- lm(Preis ~ Alter + km, data = d)
eq_long
# Coefficients:
# (Intercept)        Alter           km
#   22650            -1896       -0.031

all.equal(resid(eq_long), resid(eq_res))
# TRUE
```

Script 2.2: Beispiel zu Frisch-Waugh-Lovell Theorem, `STATA-Code`

```
clear all
insheet using http://www.hsto.info/econometrics/data/auto40.csv, ///
        delimiter(";")
destring alter, dpcomma replace  // Komma -> Punkt
regress preis km
predict res_preis, res
regress alter km
predict res_alter, res
regress res_preis res_alter
* Zum Vergleich die lange Regression
regress preis alter km
```

Wir haben eine Konsequenz des FWL Theorem bereits früher genützt, ohne explizit darauf hinzuweisen, nämlich bei der Mittelwerttransformation $\ddot{x} := x_i - \bar{x}$. Wir haben behauptet, dass wir aus mittelwertransformierten Daten die gleichen Koeffizienten berechnen können wie aus den ursprünglichen Daten. Erinnern wir uns, dass eine Regression auf die Regressionskonstante den Mittelwert $\bar{y}$ liefert; die Residuen dieser Regression auf die Regressionskonstante sind deshalb einfach die mittelwerttransformierten Daten. Das FWL Theorem sagt uns, dass wir aus einer Regression dieser Residuen aufeinander den gleichen Steigungskoeffizienten erhalten wie aus den Ursprungsdaten.

Achtung: das FWL-Theorem gilt selbstverständlich auch für die Koeffizienten der stochastischen Regressionsanalyse, aber es gilt nicht für die *Standardfehler* der Koeffizienten! Der Grund dafür ist, dass in der Residuen-Regression nicht berücksichtigt wird, dass durch die beiden vorhergehenden Hilfsregressionen Freiheitsgrade verloren gehen.

Partielle Streudiagramme für multiple Regressionen

Unter anderem können wir das FWL Theorem auch dazu nützen, um die Zusammenhänge zwischen abhängiger und erklärenden Variablen *multipler* Regression grafisch darzustellen.

Erinnern wir uns, in einem zweidimensionalen Streudiagramm können wir nur das Resultat einer bivariaten Regression darstellen. Wenn aber weitere Variablen auf y und x einwirken führt dies dazu, dass diese nicht berücksichtigten Variablen den Zusammenhang zwischen y und x verzerren, man spricht von einem einem *'Omitted Variables Bias'* (siehe Abschnitt 2.6.1).

Deshalb können grafische Darstellungen bivariater Zusammenhänge in Streudiagrammen sehr irreführend sein, ein scheinbarer Zusammenhang könnte auch auf nicht berücksichtigte Variablen zurückzuführen sein (Scheinkorrelation).

Das FWL Theorem bietet eine einfache Möglichkeit partielle Zusammenhänge korrekt darzustellen, indem wir zuerst mittels Hilfsregressionen den linearen Einfluss aller anderen (verfügbaren) Variablen eliminieren, und anschließend die Residuen dieser Hilfsregressionen in einem Streudiagramm darstellen.[13] Solche Streudiagramme werden manchmal 'Partielle (Regressions-) Streudiagramme' (*'partial regression plots'*, manchmal auch *'added variable plots'*, *'adjusted variable plots'* oder *'individual coefficient plots'*) genannt.

Beispiel: Kehren wir zurück zu unserem frühere Beispiel mit den Gebrauchtautos. Weil die erklärenden Variablen Alter und Kilometerstand korreliert sind und beide den Preis beeinflussen wird in einem bivariaten Streudiagramm Preis vs. km ein zu optimistisches Bild vom Zusammenhang gezeichnet; das nicht berücksichtigte Alter beeinflusst das Bild indirekt (vgl. Abbildung 2.18).

Abbildung 2.20 zeigt den Zusammenhang zwischen Preis und Kilometerstand von Gebrauchtautos links ohne Berücksichtigung des Alters, und rechts nachdem für

[13] Allerdings ist dabei zu beachten, dass dadurch die Skalierung geändert wird. Durch Addition der Mittelwerte zu den Residuen kann erreicht werden, dass auch die partiellen Regressionen durch den Mittelwert der Variablen läuft.

das Alter kontrolliert wurde (den Den R und Stata Programmcode finden Sie im Appendix auf Seite 147).

Wir sehen, dass die bivariate Regression (links) den Einfluss der Kilometer überschätzt, und dass die partielle Regression (rechts) stark von einer einzelnen Beobachtung beeinflusst wird (für eine bessere Vergleichbarkeit wurde für beide Grafiken die gleiche Skala gewählt).

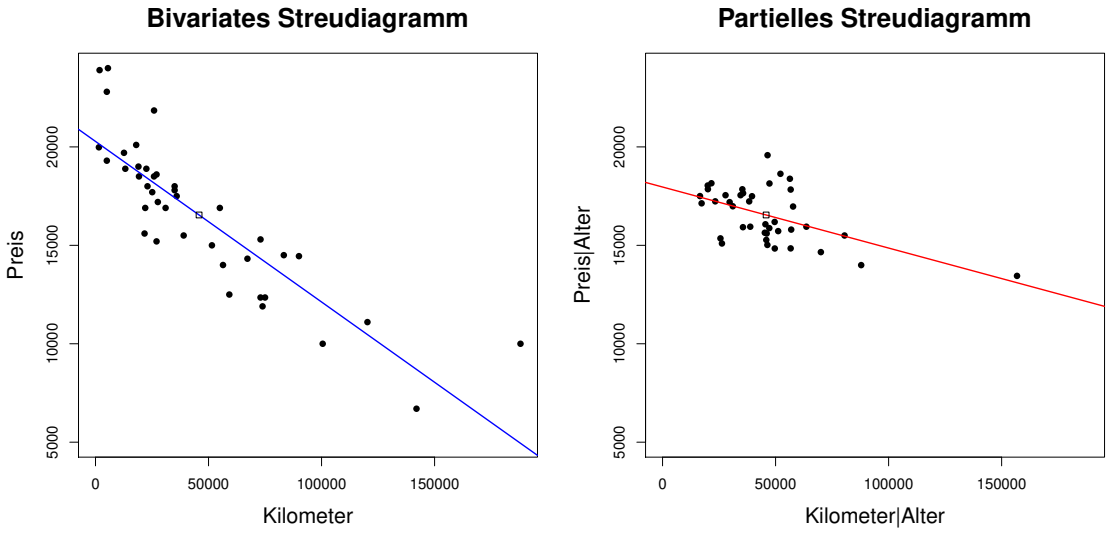


Abbildung 2.20: Bivariate und partielle Regression: bei der partiellen Regression wird für das Alter der Autos kontrolliert, d.h. es werden Residuen von Regressionen auf das Alter geplottet (damit die Regressionsgerade trotzdem durch die Mittelwerte läuft werden die jeweiligen Mittelwerte zu den Residuen addiert).
Für eine bessere Vergleichbarkeit wurde für beide Grafiken die gleiche Achsenskalierung gewählt (den R und Stata Programmcode finden Sie im Appendix auf Seite 147).

Halten wir also zusammenfassend noch einmal fest, nicht berücksichtigte relevante Variablen können über ihren Einfluss auf die berücksichtigten Variablen ein verzerrtes Bild zeichnen, und eine einfache Interpretation der Steigungskoeffizienten als marginale Effekte in diesem Fall zu irreführenden Schlussfolgerungen führen!

Tatsächlich haben wir die Daten gewissermaßen auf das Prokrustes-Bett[14] unserer linearen Spezifikation gespannt!

[14]Prokrustes – eine Figur aus der griechischen Mythologie – war bekannt dafür Reisenden ein Bett anzubieten, und sie dann an die Größe des Bettes 'anzupassen'. War der Wanderer groß hackte er ihm die Füße ab, war der Wanderer klein zog er ihn in die Länge.

2.7 Dummy Variablen

> *"Let us remember the unfortunate econometrician who, in one of the major functions of his system, had to use a proxy for risk and a dummy for sex."*
> (**?**, 892)

Dummy Variablen gehören zum praktischsten, was die einführende Ökonometrie zu bieten hat. Sehr häufig interessieren wir uns nämlich für Vergleiche zwischen Gruppen, z.B. zwischen Ländern, Branchen, oder für die Konsequenzen der Zugehörigkeit zu bestimmten Gruppen (z.B. Geschlecht). Bisher haben wir ausschließlich Variablen untersucht, die innerhalb eines Bereichs jeden Wert annehmen konnten, d.h. *intervall-* bzw. *verhältnisskalierte*[15] Variablen. Um z.B. die Zuordnung einer Person zu einer Gruppe modellieren zu können genügen Variablen, die nur zwei Werte annehmen können, z.B. Eins (1) für *'wahr'* und Null (0) für *'falsch'*. Deshalb werden solche Variablen häufig 0-1 Variablen, binäre Variablen oder auch qualitative Variablen genannt. In der Ökonometrie hat sich dafür die Bezeichnung *Dummy Variablen* eingebürgert.

Mit Hilfe solcher Dummy Variablen können im Rahmen eines Regressionsmodells die Auswirkungen qualitativer Unterschiede untersucht werden, wie zum Beispiel Lohnunterschiede zwischen Männern und Frauen. Dummy Variablen sind ein äußerst nützliches und flexibles Instrument, mit der eine Vielzahl von Fragen untersucht werden kann, wie zum Beispiel Lohnunterschiede zwischen Männern und Frauen, ob Länder in den Tropen langsamer wachsen als Länder in den gemäßigten Klimazonen, ob und wie sich die marginale Konsumneigung nach einer Steuerreform ändert, oder inwieweit sich das Ausgabeverhalten von Verheirateten gegenüber Ledigen unterscheidet.

Dummy Variablen können nur zwei Werte annehmen, Null und Eins, und werden für die Kodierung von Gruppen verwendet. Wenn ein (binäres) Merkmal vorliegt wird der Dummy Variable die Zahl Eins zugeordnet, und wenn dieses Merkmal *nicht* vorliegt die Zahl Null. Einer Dummy Variable OECD wird z.B. die Zahl Eins zugeordnet, wenn ein Land OECD Mitglied ist, und Null, wenn es kein OECD Mitglied ist. Oder, einer Dummy Variable w (für weiblich) wird die Zahl Eins zugeordnet, wenn es sich bei der Person um eine Frau handelt, und Null sonst. Natürlich könnte man ebenso gut eine Dummy Variable m für männlich anlegen

$$w_i = \begin{cases} 1 & \text{wenn Person } i \text{ eine Frau ist,} \\ 0 & \text{sonst (d.h. Mann)} \end{cases} \qquad m_i = \begin{cases} 1 & \text{wenn Mann, und} \\ 0 & \text{sonst} \end{cases}$$

[15] Bei intervallskalierten Daten ist die Reihenfolge festgelegt und die Differenzen zwischen zwei Werten können inhaltlich interpretiert werden. Bei verhältnisskalierten Variablen existiert zusätzlich ein absoluter Nullpunkt. In diesem Abschnitt werden wir uns mit Fällen beschäftigen, in denen zumindest eine erklärende Variablen nominal- oder ordinalskaliert ist. Bei einer *Nominalskala* können die Ausprägungen in keine *natürliche Reihenfolge* gebracht werden. Beispiele für nominalskalierte Merkmale sind Geschlecht, Religion, Hautfarbe, etc. Bei einer *Ordinalskala* besteht zwar eine natürliche Rangordnung, aber die Abstände zwischen den Merkmalsausprägungen sind nicht sinvoll quantifizierbar. Beispiele sind Schulnoten, Güteklassen bei Lebensmitteln, usw.

Praxistipp: In der Logik ist es üblich wahren Aussagen die Zahl Eins und falschen Aussagen die Zahl Null zuzuordnen. Bei der Wahl des Namens von Dummy Variablen empfiehlt es sich deshalb den Namen derart zu wählen, dass aus dem Namen geschlossen werden kann, welcher Ausprägung der Wert Eins zugewiesen wurde. Würde man zum Beispiel einer Dummy Variablen den Namen 'Geschlecht' geben, so kann aus diesem Variablennamen nicht geschlossen werden, welchem Geschlecht der Wert Eins zugeordnet wurde. Wenn wir die Dummy Variable hingegen 'weiblich' nennen ist klar, dass dieser Variable der Wert 1 für Frauen und 0 für Männer zugeordnet ist. Dies kann die Interpretation von Dummy Variablen erheblich erleichtern, wie wir gleich sehen werden.

Wir beginnen mit einem einfachen Beispiel, Tabelle 2.7 zeigt Stundenlöhne (StdL) von 12 Personen ($n = 12$), sowie deren Geschlecht, Familienstand und Bildungsjahre.

Tabelle 2.7: Stundenlöhne (StdL) von Männern (m) und Frauen (w), Familienstatus ($v_i = 1$ für verheiratet und Null sonst; $u_i = 1$ für unverheiratet und Null sonst) sowie Bildung (in Jahren).
Beachten Sie, dass $w = 1 - m$ (bzw. $m = 1 - w$ oder $m + w = 1$) und $u = 1 - v$.
(`https://www.uibk.ac.at/econometrics/data/stdl_bsp1.csv`)

i	StdL	m	w	v	u	Bildg
1	16	1	0	0	1	17
2	12	0	1	0	1	16
3	16	1	0	1	0	18
4	14	1	0	1	0	13
5	12	1	0	0	1	8
6	12	0	1	1	0	15
7	18	1	0	1	0	19
8	14	0	1	0	1	17
9	14	0	1	1	0	16
10	14	1	0	1	0	9
11	10	0	1	1	0	11
12	13	0	1	0	1	15

Wir können aus den Daten in Tabelle 2.7 einfach den durchschnittlichen Stundenlohn $\overline{\text{StdL}}$ sowie die *bedingten* durchschnittlichen Stundenlöhne für Männer, Frauen, Verheiratete und Unverheiratete berechnen:

Mittelwert von StdL:

$$\overline{\text{StdL}} = (16 + 12 + 16 + \cdots + 13)/12 = 13.75$$

Bedingte Mittelwerte von StdL:

$$\begin{aligned}(\overline{\text{StdL}}|m=1) &= (16+16+14+12+18+14)/6 = 15\\ (\overline{\text{StdL}}|w=1) &= (12+12+14+14+10+13)/6 = 12.5\end{aligned}$$

$$\begin{aligned}(\overline{\text{StdL}}|v=1) &= (16+14+12+18+14+14+10)/7 = 14\\ (\overline{\text{StdL}}|v=0) &= (16+12+12+14+13)/5 = 13.4\end{aligned}$$

In einem früheren Beispiel haben wir gezeigt, dass eine Regression *nur* auf die Regressionskonstante (d.h. einen Einsen-Vektor) den Mittelwert der abhängigen Variable liefert (siehe Seite 20). Wir werden nun gleich sehen, dass wir auch die bedingten Mittelwerte einfach mit Hilfe einer OLS Regression berechnen können, nämlich durch eine Regression auf eine Dummy Variable.

Für die Daten aus Tabelle 2.7 liefert eine Regression auf die Dummy Variable m

$$\widehat{\text{StdL}}_i = b_1 + b_2 m_i = 12.5 + 2.5\, m_i$$

Wenn $m_i = 0$, also für Frauen, erhalten wir $\widehat{\text{StdL}}_i = b_1 + b_2 \times 0 = b_1$; deshalb vermuten wir, dass das Interzept b_1 den durchschnittlichen Stundenlohn von Frauen liefert.

Für Männer ist $m_i = 1$, also $\widehat{\text{StdL}}_i = b_1 + b_2 \times 1 = b_1 + b_2$, deshalb vermuten wir, dass $b_1 + b_2$ den durchschnittlichen Stundenlohn von Männern angibt, und der Steigungskoeffizient b_2 die Differenz zwischen durchschnittlichen Stundenlöhnen von Männern und Frauen misst.

Dies ist tatsächlich richtig, der bedingte Mittelwert des Stundenlohns für Frauen beträgt 12.5 Euro, und Männer verdienen in diesem Beispiel im Durchschnitt um 2.5 Euro mehr als Frauen, also 15 Euro.

$$\overline{\text{StdL}}|(m=0) = b_1, \qquad \overline{\text{StdL}}|(m=1) = b_1 + b_2$$

Da das Interzept jeweils den Mittelwert der 'Null-Kategorie' angibt (d.h. den Mittelwert der Kategorie, welcher in der Dummy Variable der Wert Null zugewiesen wurde), wird diese 'Null-Kategorie' häufig *Referenzkategorie* genannt.

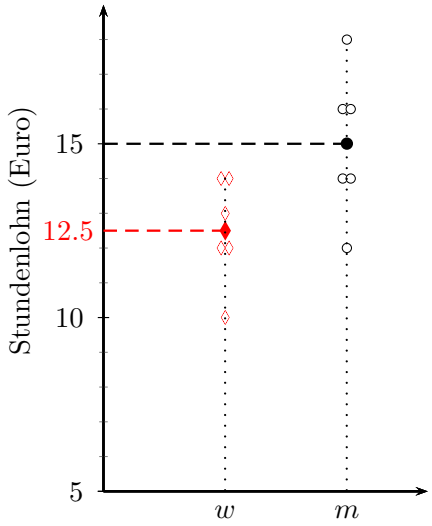
Der Steigungskoeffizient misst die Differenz zwischen dem Mittelwert dieser Referenzkategorie und dem Mittelwert der 'Eins-Kategorie' (d.h. der Kategorie, welcher in der Dummy Variable der Wert Eins zugewiesen wurde), in diesem Beispiel also um wie viel Euro der durchschnittliche Stundenlohn von Männern (mit $m_i = 1$) höher ist als der durchschnittliche Stundenlohn der Referenzkategorie (d.h. Frauen mit $m_i = 0$).

$$\overline{\text{StdL}}|(m=1) - \overline{\text{StdL}}|(m=0) = 15 - 12.5 = 2.5 = b_2$$

Alternativ hätten wir natürlich auch eine Regression auf die Dummy Variable w (für weiblich) rechnen können; diese liefert

$$\widehat{\text{StdL}}_i = 15 - 2.5\, w_i$$

Da $w_i = 1$ für Frauen und $w_i = 0$ für Männer bilden in diesem Fall Männer die Referenzkategorie, deren mittlerer Stundenlohn im Interzept gemessen wird ($b_1 =$



$$\widehat{\text{StdL}}_i = 12.5 + 2.5\, m_i$$

bzw.

$$\widehat{\text{StdL}}_i = 15 - 2.5\, w_i$$

Abbildung 2.21: Stundenlöhne von Männern und Frauen, siehe Tabelle 2.7.

15), und der durchschnittliche Stundenlohn für Frauen ist um 2.5 Euro *niedriger* als der von Männern ($b_2 = -2.5$). Abbildung 2.21 zeigt dies für die Daten aus Tabelle 2.7.

Man könnte auf die Idee kommen, eine Regression auf eine Regressionskonstante *und* die beiden Dummy Variablen w und m zu rechnen, also $y_i = b_1 + b_2 w_i + b_3 m_i + e_i$. Dies funktioniert aber nicht, da in diesem Fall eine lineare Beziehung zwischen den Regressoren besteht (die Summe der beiden Dummies ergibt die Regressionskonstante, d.h. $w_i + m_i = 1$). Wann immer eine exakte lineare Abhängigkeit zwischen Regressoren besteht ist die OLS Funktion nicht definiert, es existieren unendlich viele Lösungen.[16]

Dies ist im einfachsten Fall leicht zu erkennen; wenn alle Ausprägungen des Regressors die gleichen Ausprägungen haben (z.B. $x_i = 5$) wäre x ein Vielfaches der Regressionskonstante, und die Varianz einer Konstanten ist natürlich Null. Da $b_2 = \text{cov}(x, y)/\text{var}(x)$ und $\text{var}(x) = 0$ existiert in diesem Fall keine Lösung für b_2. Wir werden später zeigen, dass dies für alle linearen Abhängigkeiten zwischen Regressoren gilt.

Aber wir können eine Regression auf beide Dummy Variablen m und w *ohne* Regressionskonstante rechnen. In diesem Fall liefern die geschätzten Koeffizienten einfach die Mittelwerte beider Kategorien

$$\begin{aligned} \widehat{\text{StdL}}_i &= b_2 w_i + b_3 m_i \\ &= 12.5\, w_i + 15\, m_i \end{aligned}$$

Übung mit Lösungshinweisen: Sei y_i eine von insgesamt n Beobachtungen einer intervallskalierten Variable, und d eine Dummy Variable; n_1 ist die Anzahl der Elemente dieser Dummy Variable mit der Ausprägung Eins und n_0 die Anzahl der Elemente mit dem Wert Null ($n_1 + n_0 = n$).

[16] Wir werden diesen Fall später unter der Bezeichnung *perfekte Multikollinearität* ausführlich diskutieren.

Der Mittelwert von y ist $\bar{y} := \frac{1}{n}\sum_{i=1}^{n} y_i$, und der Mittelwert von d ist $\bar{d} := \frac{1}{n}\sum_{i=1}^{n} d_i = \frac{n_1}{n}$ (warum?).
Den Mittelwert aller y_i für die gilt $d_i = 0$ nennen wir $\bar{y}_0$, und den Durchschnitt aller y_i mit $d_i = 1$ mit $\bar{y}_1$.
Wir werden nun in mehreren Schritten zeigen, dass allgemein gilt

$$\begin{aligned} y_i &= b_1 + b_2 d_i + e_i \\ &= \bar{y}_0 + (\bar{y}_1 - \bar{y}_0) d_i + e_i \end{aligned}$$

d.h., das Interzept b_1 dieser Regression ist der Mittelwert der Gruppe mit $d_i = 0$ (Referenzgruppe), und der Steigungskoeffizient b_1 ist die Differenz der Gruppe mit $d_i = 1$ zur Referenzgruppe.

1. Sei $\bar{y}_1$ der bedingte Mittelwert der y_i für die gilt $d_i = 1$, und $\bar{y}_0$ der bedingte Mittelwert aller y_i für die $d_i = 0$ (d.h. $\bar{y}_0 := \bar{y}|(d_i = 0)$ und $\bar{y}_1 := \bar{y}|(d_i = 1)$). Zeigen Sie allgemein, dass

   $$\bar{y} = \frac{n_0}{n}\bar{y}_0 + \frac{n_1}{n}\bar{y}_1$$

   Lösungshinweis: die Daten werden zuerst sortiert, sodass zuerst alle Beobachtungen mit $d_i = 0$ kommen, und anschließend alle Beobachtungen mit $d_i = 1$

   $$\begin{aligned} \bar{y} &= \frac{1}{n}\left(\frac{n_0}{n_0}\sum_{i=1}^{n_0} y_i + \frac{n_1}{n_1}\sum_{j=n_0+1}^{n} y_j\right) \\ &= \frac{n_0}{n}\left(\frac{1}{n_0}\sum_{i=1}^{n_0} y_i\right) + \frac{n_1}{n}\left(\frac{1}{n_1}\sum_{j=n_0+1}^{n} y_j\right) \\ &= \frac{n_0}{n}\bar{y}_0 + \frac{n_1}{n}\bar{y}_1 \end{aligned}$$

   Die Summe der mit den Anteilen gewichteten *bedingten* Mittelwerte ist der Gesamtmittelwert!

2. Zeigen Sie, dass die empirische Varianz $\mathrm{var}(y) := \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$ auch geschrieben werden kann als

   $$\mathrm{var}(y) = \frac{1}{n}\sum_i y_i^2 - \bar{y}^2 := \overline{y^2} - \bar{y}^2$$

   Lösungsskizze:

   $$\begin{aligned} \mathrm{var}(y) &:= \frac{1}{n}\sum_i (y_i - \bar{y})^2 = \frac{1}{n}\sum_i \left(y_i^2 - 2\bar{y}y_i + \bar{y}^2\right) \\ &= \frac{1}{n}\left(\sum_i y_i^2 - 2\bar{y}\sum_i y_i + \sum_i \bar{y}^2\right) \\ &= \frac{1}{n}\sum_i y_i^2 - 2\frac{1}{n}n\bar{y}^2 + \frac{1}{n}n\bar{y}^2 \\ &= \frac{1}{n}\sum_i y_i^2 - \bar{y}^2 := \overline{y^2} - \bar{y}^2 \end{aligned}$$

   da aus $\bar{y} := \frac{1}{n}\sum_i y_i$ folgt $\sum_i y_i = n\bar{y}$

3. Zeigen Sie, dass die empirische Kovarianz $\text{cov}(y,x) := \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})$ auch geschrieben werden kann als

$$\text{cov}(y,x) = \frac{1}{n}\sum_i y_i x_i - \bar{y}\bar{x} := \overline{xy} - \bar{y}\bar{x}$$

4. Zeigen Sie, dass für eine Dummy Variable d gilt

$$\text{var}(d) = \frac{n_1}{n}\left(1-\frac{n_1}{n}\right)$$

Beachten Sie, dass $\sum_i d_i^2 = n_1$ (weil $1^2 = 1$)

Lösungsskizze: beachte, dass für eine Dummy Variable $\sum_i d_i = n_1$ und $\bar{d} = \frac{n_1}{n}$. Deshalb

$$\begin{aligned}\text{var}(d) &= \frac{1}{n}\sum_i \left(d_i-\bar{d}\right)^2 = \overline{d^2} - \bar{d}^2 \\ &= \frac{n_1}{n} - \left(\frac{n_1}{n}\right)^2 \\ &= \frac{n_1}{n}\left(1-\frac{n_1}{n}\right)\end{aligned}$$

5. Zeigen Sie, dass für eine Dummy Variable d gilt

$$\begin{aligned}\text{cov}(y,d) &= \frac{1}{n}\sum_i (y_i-\bar{y})(d_i-\bar{d}) = \frac{n_1}{n}(\bar{y}_1-\bar{y}) = \\ &= \frac{n_1}{n}\left[\frac{n_0}{n}(\bar{y}_1-\bar{y}_0)\right] = \frac{n_1}{n}\left(\frac{n-n_1}{n}\right)(\bar{y}_1-\bar{y}_0) \\ &= \text{var}(d)(\bar{y}_1-\bar{y}_0)\end{aligned}$$

Hinweis: $\text{cov}(y,d) = \frac{1}{n}\sum_i y_i d_i - \bar{y}\bar{d}$, $\quad \bar{y} = \frac{n_0}{n}\bar{y}_0 + \frac{n_1}{n}\bar{y}_1$; und

$$\frac{1}{n}\sum_i y_i d_i = \frac{n_1}{n}\bar{y}_1, \qquad \frac{n_0}{n} = \frac{n-n_1}{n} = 1-\frac{n_1}{n}, \qquad \bar{d} = \frac{n_1}{n} \qquad \text{(warum?)}$$

6. Zeigen Sie, dass in einer Regression auf eine Dummy Variable $y_i = b_1 + b_2 d_i + e_i$ der Steigungskoeffizient

$$b_2 = \frac{\text{cov}(y,d)}{\text{var}(d)} = \bar{y}_1 - \bar{y}_0$$

Hinweis: $1-\frac{n_1}{n} = \frac{n_0}{n}$ (warum?)

7. Zeigen Sie, dass das Interzept b_1 berechnet werden kann als

$$b_1 = \bar{y} - b_2\bar{d} = \bar{y}_0$$

Lösungsskizze:

$$\begin{aligned}b_1 &= \frac{n_0}{n}\bar{y}_0 + \frac{n_1}{n}\bar{y}_1 - (\bar{y}_1-\bar{y}_0)\frac{n_1}{n} \\ &= \left(\frac{n_0+n_1}{n}\right)\bar{y}_0 \\ &= \bar{y}_0\end{aligned}$$

Deshalb liefert das Interzept b_1 einer Regression auf eine Dummy Variable $y_i = b_1 + b_2 d_i + e_i$ den Mittelwert der Referenzkategorie $\bar{y}_0$, und der Steigungskoeffizient misst die Differenz der Mittelwerte beider Kategorien ($b_2 = \bar{y}_1 - \bar{y}_0$).

□

Partielle Effekte: Den einfachsten Fall haben wir im vorhergehenden Beispiel bereits diskutiert, eine einfache Regression auf eine Regressionskonstante und eine Dummy Variable d

$$\widehat{y} = b_1 + b_2 d$$

die uns im Interzept den Mittelwert der Referenzkategorie (für die $d_i = 0$ ist) als Steigungskoeffizienten die Differenz der Mittelwerte der beiden Kategorien liefert.

Diese Differenz entspricht dem *marginalen Effekt* bei metrisch skalierten Regressoren, aber da sich Dummy Variablen per Definition nicht infinitesimal ändern können ist es kaum angebracht, von einem *marginalen Effekt* zu sprechen; immerhin kann es sich dabei um Unterschiede wie z.B. zwischen Männern und Frauen handeln, eine partielle Ableitung macht hier wenig Sinn.

Deshalb ist es klüger die Unterschiede in y für die beiden Kategorien zu vergleichen, und in Analogie zu marginalen Effekt spricht man bei Dummy Variablen häufig von einem *partiellen Effekt*: wie groß ist ceteris paribus der mittlere Unterschied von y zwischen den beiden Kategorien, z.B. Männern und Frauen?

Wie wir schon gesehen haben misst der Koeffizient der Dummy Variablen die Differenz zur 'Referenzkategorie' $d_i = 0$

$$\begin{aligned} \widehat{y}|(d=1) &= b_1 + b_2 \\ \widehat{y}|(d=0) &= b_1 \end{aligned}$$

und die Differenz ist der *"partielle Effekt"*

$$[\widehat{y}|(d=1)] - [\widehat{y}|(d=0)] = b_1 + b_2 - b_1 = b_2$$

Daran ändert sich nichts Wesentliches, wenn weitere erklärende x Variablen als Regressoren berücksichtigt werden

2.7.1 Unterschiede im Interzept

Wir erweitern unser Dummy Modell, indem wir *zusätzlich* eine metrisch skalierte Variable berücksichtigen. Dazu kehren wir zu unserem Beispiel mit den Stundenlöhnen zurück (siehe Tabelle 2.7) und berücksichtigen zusätzlich die Bildungszeit in Jahren ('Bildg').

Eine Regression auf die Dummy Variable m (für männlich) und 'Bildg' gibt

$$\widehat{\text{StdL}} = 5.78 + 2.95\, m + 0.45\, \text{Bildg}$$

(mit $R^2 = 0.87$ und $n = 12$).

Was ist passiert? Plötzlich ist das Interzept deutlich kleiner und die Differenz zwischen den Stundenlöhnen von Männern und Frauen noch größer (zur Erinnerung, eine Regression nur auf die Dummy Variable lieferte $\widehat{\text{StdL}} = 12.5 + 2.5\,m$).

Das nun viel kleinere Interzept ist schnell erklärt, es gibt den hypothetischen mittleren Stundenlohn für Frauen mit Null Bildungsjahren an; in diesem Datensatz existiert keine solche Person.

Aber warum scheinen Männer nun im Durchschnitt um 2.95 Euro mehr zu verdienen als Frauen? Die Antwort folgt aus der ceteris paribus Bedingung, *bei gleicher Bildung*!

Erinnern wir uns an das Kapitel über die Nichtberücksichtigung relevanter Variablen zurück. Dort haben wir argumentiert, dass zwischen dem Steigungskoeffizienten eines 'kurzen' und 'langen' Modells folgende Beziehung besteht:

$$b_2^k = b_2 + b_3 a_2$$

wobei b_2^k der Koeffizient der Dummy Variable des kurzen Modells $\widehat{\text{StdL}} = 12.5 + 2.5\,m$ ist, und $b_2 = 2.95$ $b_3 = 0.45$ die obigen Koeffizienten des langen Modells sind.

Im kurzen Modell 'fehlt' die Variable Bildung. Diese nichtberücksichtigte Variable Bildung regressieren wir in einer Hilfsregression auf die berücksichtigte Dummy Variable m und erhalten den Steigungskoeffizienteb a_2

$$\widehat{\text{Bildg}} = a_1 + a_2\,m = 15 - 1\,m$$

Diese Hilfsregression sagt uns, dass Frauen (die Referenzkategorie) im Durchschnitt 15 Bildungsjahre aufweisen, und Männer um ein Jahr weniger (also 14 Bildungsjahre).

Wenn wir dies in die Formel für die nicht-berücksichtigte Variable einsetzen erhalten wir

$$2.95 + 0.45 \times (-1) = 2.5$$

Aus der 'langen' Regression mit den Bildungsjahren lernen wir also, dass ein einfacher Vergleich von Durchschnittszahlen in die Irre führen kann.

In diesem Beispiel haben Frauen im Durchschnitt ein höheres Bildungsniveau, und im 'kurzen' Modell wird dies nicht berücksichtigt. Da im 'kurzen' Modell 'Bildg' kein Regressor ist fällt diese im 'kurzen' Modell nicht unter die ceteris paribus Annahme, im 'langen' Modell hingegen schon.

Das Bildungsniveau ist positiv mit dem Stundenlohn korreliert ($b_3 > 0$), und Männer im Durchschnitt ein niedrigeres Bildungsniveau haben ($a_2 < 0$, unterschätzt das 'kurze' Modell den ceteris paribus Unterschied, vgl. Tabelle 2.6 (Seite 51).

Bei gleicher Bildung (ceteris paribus) wäre die Lohndifferenz mit 2.95 Euro deutlich größer als der Unterschied der einfachen Mittelwerte von 2.5 Euro! Dies wird in Abbildung 2.22 gezeigt.

Für Österreich erhält man auf Grundlage der EU-Silc (2018) Daten folgende Schätzung

$$\begin{array}{rcccccc} \text{StdL} & = & 10.966 & + & 0.615\ \text{potBildg} & - & 3.121\ \text{Weibl} \\ & & (0.48)^{***} & & (0.035)^{***} & & (0.294)^{***} \end{array}$$

$$R^2 = 0.087, \qquad n = 4151$$

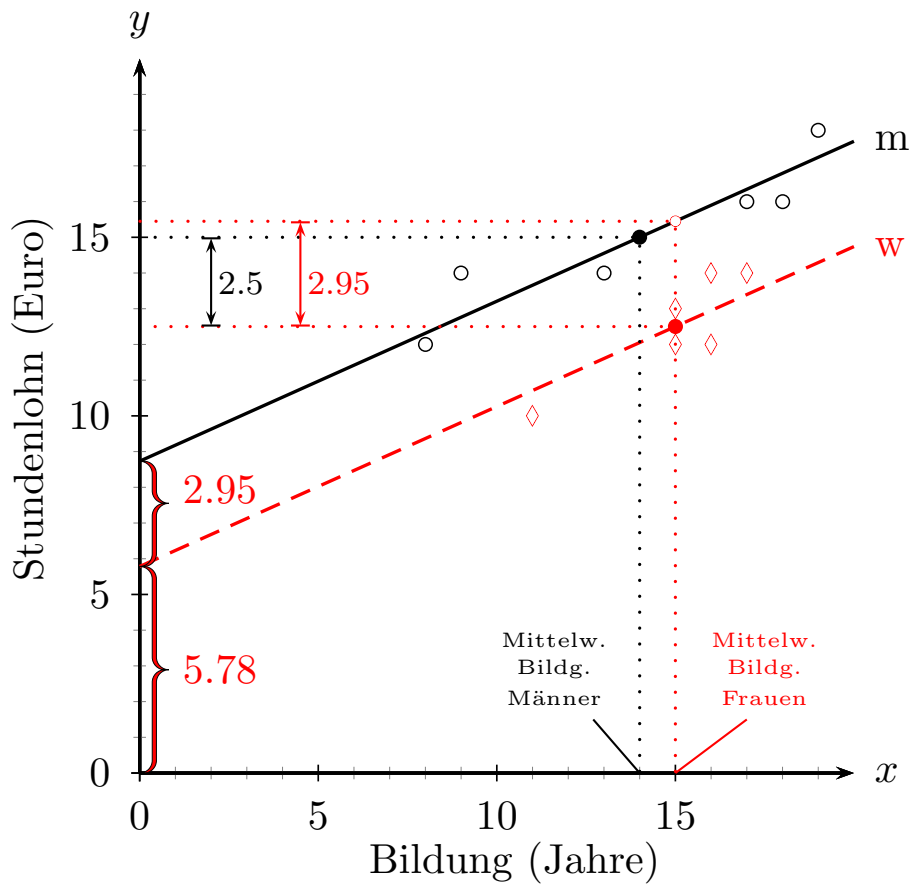


Abbildung 2.22: Unterschiede im Interzept; $\widehat{\text{StdL}} = 5.78 + 2.95\,m + 0.45\,\text{Bildg}$
Der einfache mittlere Lohnunterschied zwischen Männern und Frauen beträgt 2.5 Euro, aber dieser berücksichtigt nicht, dass hier die durchschnittliche Bildungsdauer von Frauen 15 Jahre beträgt, also um ein Jahr mehr als die durchschnittliche Bildungsdauer von Männern (14 Jahre). Der *ceteris paribus Unterschied* (d.h. bei gleicher Bildungsdauer) beträgt 2.95 Euro! Man beachte auch die Bedeutung der angenommenen linearen Funktionsform.

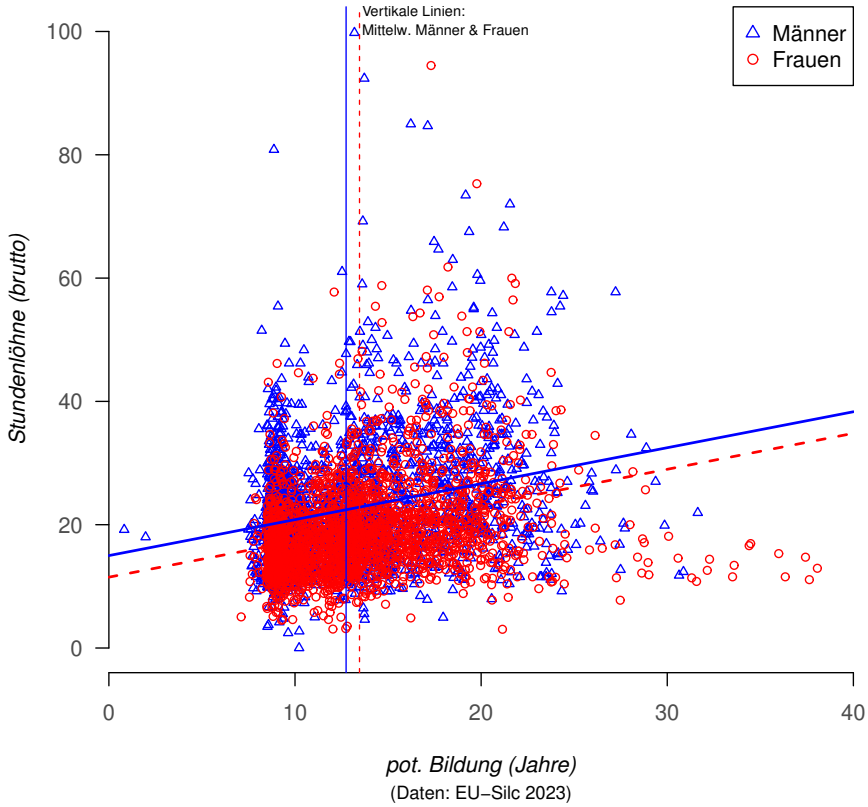


Abbildung 2.23: Stundenlöhne und (potentielle) Bildungsjahre in Österreich ($n = 4532$).
Quelle: EU-Silc 2023, Statistik Austria

die in Abbildung 2.23 dargestellt ist. Man beachte, dass die Verteilung der Stundenlöhne sehr rechtsschief ist, und deshalb bedingte Mittelwerte keine sehr gut geeignete Kennzahl sind; wir werden im Abschnitt zu logarithmischen Funktionen darauf zurückkommen.

Durch diese Spezifikation mit einer einfachen Dummy haben wir zwar zugelassen, dass sich das Interzept zwischen Männern und Frauen unterscheiden kann, aber wir haben a priori unterstellt, dass Bildung für Männern und Frauen die gleichen Auswirkungen hat, mit jedem zusätzlichen Bildungsjahr steigt der mittlere Stundenlohn für Männer und Frauen um 0.45 Euro. Dies ist natürlich eine sehr restriktive Annahme, die wir aber leicht lockern können.

2.7.2 Unterschiede in der Steigung

Wenn man das Produkt einer Dummy Variable mit einer anderen metrisch skalierten Variable als zusätzlichen Regressor einführt erlaubt dies unterschiedliche Steigungen der Regressionsgeraden für beide Kategorien.

Dies ist möglich, indem man das *Produkt* Dummyvariablen als zusätzlichen Regressor einführt. Ein solches Produkt zweier Regressoren wird *Interaktionseffekt* genannt und wird uns später noch ausführlicher beschäftigen.

Hier genügt es festzustellen, dass das Produkt zweier Dummyvariablen immer 1 ist, wenn *beide* Dummyvariablen den Wert 1 haben, und 0 sonst.

Im Beispiel mit den Stundenlöhnen

$$\widehat{\text{StdL}} = b_1 + b_2\,\text{Bildg} + b_3(m \times \text{Bildg})$$

In diesem Fall können sich die *Steigungen* der Regressiongeraden beider Kategorien unterscheiden, für die Kategorie $m = 0$ ist die Steigung b_2, und für die Kategorie $m = 1$ ist die Steigung $b_2 + b_3$.

$$\begin{aligned} \widehat{y} &= b_1 + b_2 x + b_3(m \times x) \\ \widehat{y}|(m=1) &= b_1 + (b_2 + b_3)x \\ \widehat{y}|(m=0) &= b_1 + b_2 x \end{aligned}$$

Die Steigungen sind

$$\frac{\partial \widehat{y}|(m=1)}{\partial x} = b_2 + b_3; \qquad \frac{\partial \widehat{y}|(m=0)}{\partial x} = b_2$$

Der Koeffizient des Interaktionsterms b_3 misst den *Unterschied der Steigungen* zwischen beiden Kategorien, denn

$$\frac{\partial \widehat{y}|(m=1)}{\partial x} - \frac{\partial \widehat{y}|(m=0)}{\partial x} = b_3$$

Für unser Beispiel mit den Stundenlöhnen erhalten wir

$$\widehat{\text{StdL}} = 8.27 + 0.29\,\text{Bildg} + 0.19(m \times \text{Bildg}), \qquad (R^2 = 0.83,\ n = 12)$$

Demnach würde der mittlere Stundenlohn von Frauen ($m = 0$) mit einem zusätzlichen Bildungsjahr um 0.29 Euro zunehmen, der mittlere Stundenlohn von Männern würde mit jedem zusätzlichen Bildungsjahr um $0.29 + 0.19 = 0.48$ Euro steigen. Diese Spezifikation erlaubt also die Modellierung unterschiedlicher Auswirkungen der metrisch skalierten Variable auf die beiden der Dummy Variable zugrunde liegenden Kategorien.

Allerdings impliziert diese Spezifikation für beide Kategorien das gleiche Interzept (siehe Abbildung 2.24), was in den meisten Fällen eine theoretisch nur schwer begründbare Restriktion darstellt. Es ist fast immer klüger unterschiedliche Ordinatenabschnitte *und* unterschiedliche Steigungen zuzulassen.

2.7.3 Unterschiede im Interzept und Steigung

Wir können die Spezifikation leicht verallgemeinern und das Unterschiede im Interzept *und* der Steigung zulassen. Dazu müssen wir nur sowohl eine Dummy als auch eine Interaktionsvariable zwischen Dummy Variable und metrisch skalierten x Variable verwenden

$$\begin{aligned} \widehat{y} &= b_1 + b_2 x + b_3 m + b_4(m \times x) \\ \widehat{y}|(m=1) &= (b_1 + b_3) + (b_2 + b_4)x \\ \widehat{y}|(m=0) &= b_1 + b_2 x \end{aligned}$$

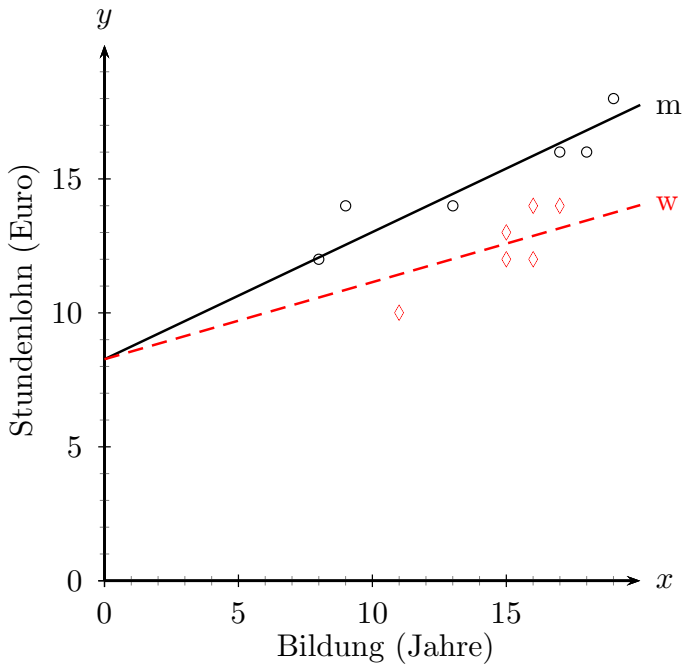


Abbildung 2.24: Unterschiede in der Steigung;
$\widehat{\text{StdL}} = 8.27 + 0.29\,\text{Bildg} + 0.19(m \times \text{Bildg})$

Der Unterschied zwischen den beiden Kategorien ist wieder

$$\widehat{y}|(m=1) - \widehat{y}|(m=0) = b_3 + b_4 x$$

Man beachte, dass man die gleichen Koeffizienten erhält, wenn man für beide Gruppen eine eigene Regression rechnen würde

$$\begin{aligned} \text{für } m=0: \quad \widehat{y}^0 &= b_1 + b_2 x \\ \text{für } m=1: \quad \widehat{y}^1 &= c_1 + c_2 x \end{aligned}$$

mit $c_1 = b_1 + b_3$ und $c2 = b_2 + b_4$.[17]
Für unser Beispiel mit den Stundenlöhnen erhalten wir

$$\widehat{\text{StdL}} = 2.95 + 6.30m + 0.64\,\text{Bildg} - 0.23(m \times \text{Bildg})$$

(mit $R^2 = 0.89,\ n = 12$); siehe Abbildung 2.25.
Wenn wir getrennte Regressionen für Männer und Frauen rechnen erhalten wir:
Für Frauen ($m = 0$):

$$\widehat{\text{StdL}} = 2.95 + 0.64\,\text{Bildg}$$

Für Männer ($m = 1$):

$$\widehat{\text{StdL}} = 9.25 + 0.41\,\text{Bildg}$$

Bitte beachten Sie, dass es sich bei diesem Beispiel um rein hypothetische Zahlen handelt.

[17] Allerdings werden sich die Standardfehler bei diesen Ansätzen unterscheiden, da das Dummy Variablen Modell implizit für beide Gruppen die gleiche Varianz σ^2 (*Homoskedastizität*) unterstellt. Deshalb sollte vor Anwendung des Dummy Variablen Modells getestet werden, ob die Varianzen tatsächlich in allen Gruppen gleich sind. Wie das geht erfahren Sie im Kapitel über Heteroskedastizität.

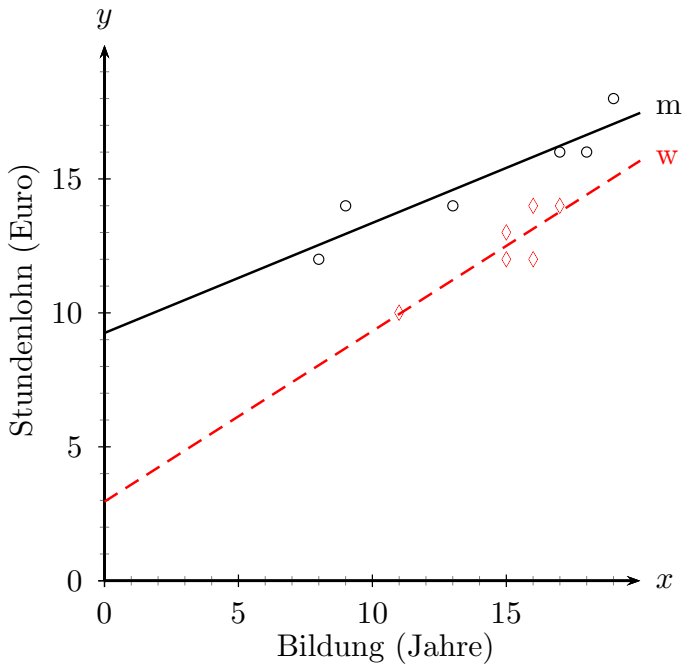


Abbildung 2.25: Unterschiede im Interzept und Steigung;
$\widehat{\text{StdL}} = 2.95 + 6.30m + 0.64\,\text{Bildg} - 0.23(m \times \text{Bildg})$

2.7.4 Eine kategoriale Variable mit mehr als zwei Ausprägungen

Viele kategoriale Variablen haben mehr als zwei Ausprägungen, z.B. die Nationalität, der höchste Bildungsabschluss, Schulnoten usw. Oft werden die unterschiedlichen Ausprägungen in Zahlen kodiert, z.B. werden in den EU-SILC Daten der Statistik Austria den höchsten Bildungsabschlüssen die Zahlen 1 bis 6 zugeordnet, siehe Tabelle 2.8, und nicht verfügbare Werte werden mit negativen Zahlen (hier -3 & -1) kodiert.

Tabelle 2.8: EU-SILC 2018 (Statistik Austria 2020)

```
P137000  Höchster Bildungsabschluss
-3  Weiß nicht
-1  keine Angabe
 1  Pflichtschule
 2  Lehre mit Berufsschule
 3  Fach- oder Handelsschule
 4  Matura
 5  Abschluss an einer Universität, (Fach-)Hochschule
 6  Anderer Abschluss nach der Matura
```

Mit diesen Zahlen kann man natürlich nicht sinnvoll rechnen, sie haben keine inhaltliche Bedeutung. Die Zahlen wurden lediglich zugeordnet, um die Daten platzsparend speichern zu können.

Trotzdem kann man mit derart kodierten Daten sehr einfach arbeiten, man muss nur eine *Referenzkategorie* wählen, und für jede weitere Kategorie eine eigene Dummy Variable.

In diesem konkreten Beispiel mit den Bildungsabschlüssen würde man z.B. zuerst die negativen Werten mit '`NA`' (für *'not available'*) ersetzen, als Referenzkategorie z.B. '`Pflichtschule`' wählen, und für jede weitere Kategorie eine Dummy-Variable anlegen die den Wert 'Eins' annimmt, wenn eine Person in diese Kategorie fällt, und 'Null' sonst. Für diesen Fall würden wir also fünf Dummy Variablen benötigen (dies erzeugen die meisten Programme automatisch, `R` verwendet dafür `factor` Objekte).

Das Interzept misst dann wieder den Wert der Referenzkategorie, wenn alle Regressoren (inklusive Dummies) gleich Null sind, und die Koeffizienten der Dummies messen wieder den ceteris paribus Unterschied zur Referenzkategorie.

Wichtig ist dabei, dass jede Beobachtung in genau eine Kategorie fällt, anderenfalls (wenn z.B. bei einer Variable 'Nationalität' eine Person mehrere Staatsbürgerschaften besitzt) wird der im übernächsten Abschnitt diskutierte Fall *mehrerer kategorialer Variablen* relevant.

Kehren wir zur Demonstration noch einmal zurück zu unserem Beispiel mit den Gebrauchtautos, wobei wir wieder das auf ganze Jahre gerundete Alter '`AlterJ`' verwenden; diese Variable hat die Ausprägungen $\{0, 1, 2, 3, 4, 5\}$.

In Spalte (1) von Tabelle 2.9 werden die Preise auf dieses Alter mit den 6 Ausprägungen regressiert. Wie schon früher gezeigt misst das Interzept (= 22 709.3) den durchschnittlichen Preis von Gebrauchtautos mit einem gerundeten Alter von Null Jahren, und der Koeffizient des Alters (= −2 517.27) misst die durchschnittliche Abnahme des Preises mit jedem weiteren Jahr. Dies sind die gleichen Größen, die wir bereits in Tabelle 2.3 (Seite 23) erhalten haben (abgesehen von unterschiedlichen Rundungen).

Man beachte, dass diese Spezifikation nur eine konstante Abnahme des Preises zulässt, d.h. diese Spezifikation erzwingt eine Approximation, der zufolge die Abnahme des Preises im 1. Jahr genau gleich groß sein muss wie im 5. Jahr. Dies mag in diesem Fall noch eine annehmbare Approximation darstellen, aber spätestens im Falle von nominal- oder ordinalskalierten Variablen macht eine solche Approximation überhaupt keinen Sinn mehr.

In solchen Fällen legen wir (bzw. das Programm) für jede Ausprägung der kategorialen Variable mit Ausnahme der Referenzkategorie eine eigene Dummy Variable an und verwenden diese als Regressoren.

Im Beispiel mit den Gebrauchtautos wählen wir z.B. das Alter von Null Jahren (d.h. AlterJ = 0) als Referenzkategorie, und legen für alle anderen Altersstufen eine eigene Dummy Variable an.

Spalte (2) von Tabelle 2.9 zeigt das Ergebnis der Regression. Wie erwartet gibt das Interzept (= 23 566.67) den Durchschnittspreis von Autos mit AlterJ = 0 an, und die Koeffizienten der Dummy Variablen messen den durchschnittlichen Unterschied im Preis zu dieser Referenzkategorie. Autos mit einem Alter von vier Jahren sind im Durchschnitt also um 11 163, 81 Euro billiger als Autos mit dem Alter von Null Jahren. Man beachte, dass diese Spezifikation keine konstante Abnahme des Preises 'erzwingt', sondern von Jahr zu Jahr unterschiedliche Abnahmen des Preises zulässt.

Tabelle 2.9: Drei verschiedene Spezifikationen für die Preise von Gebrauchtautos, Alter gerundet auf ganze Jahre. Vergleiche die Ergebnisse mit Tabelle 2.3 (Seite 23).

	Abhängige Variable: Preis		
	(1)	(2)	(3)
Interzept	22 709.30	23 566.67	
AlterJ (Jahre)	−2 517.27		
AlterJ= 0			23 566.67
AlterJ= 1		−4 158.10	19 408.57
AlterJ= 2		−5 870.83	17 695.83
AlterJ= 3		−7 785.42	15 781.25
AlterJ= 4		−11 163.81	12 402.86
AlterJ= 5		−13 666.67	9 900.00
n	40	40	40
R^2	0.82	0.84	[0.99]

Vergleichen Sie dies wieder mit Tabelle 2.3 (Seite 23); die Koeffizienten der Dummy Variablen messen den Unterschied im mittleren Preis zur Referenzkategorie AlterJ = 0.

Spalte (3) von Tabelle 2.9 zeigt schließlich das Ergebnis einer Regression, in der kein Interzept berücksichtigt wird, dafür aber alle sechs Dummy Variablen für das Alter. Wie erwartet messen die Koeffizienten der Dummy Variablen in diesem Fall einfach den durchschnittlichen Preis der Autos mit dem betreffenden Alter, wie wieder ein Vergleich mit Tabelle 2.3 (Seite 23) zeigt. Man beachte, dass das R^2 der Gleichung in Spalte (3) nicht wie üblich interpretiert werden darf, weil diese Regression kein Interzept enthält!

Hinweis: Den Code zur Erzeugung von Tabelle 2.9 finden Sie im Appendix, Seite 148.

2.7.5 Beispiel: Heterogenität und das Simpson-Paradox

Unsere Untersuchungsobjekte unterscheiden sich in der Regel in unzähligen Merkmalen, und wir sind bei unseren Untersuchungen aus praktischen Überlegungen fast immer gezwungen, einen Großteil dieser Heterogenität außer Acht zu lassen.

Solche *nicht berücksichtigte Heterogenität* kann dramatische Folgen haben, wie das folgende hypothetische Beispiel demonstriert.

Angenommen, eine Universität mit zwei überlaufenen Studienrichtungen – Psychologie und Mathematik – Zulassungsprüfungen ein.

Nach Durchführung der Prüfung wird folgendes Ergebnis bekannt gegeben:

	Frauen	Männer
Bewerber*innen	500	500
davon zugelassen	200	300
in Prozent	40%	60%

Angesichts der deutlich höheren Zulassungsquote bei Männern kommt die Universitätsleitung in Erklärungsnotstand.

Da kommt eine brillante Statistikerin der Universität auf die Idee, sich die Ergebnisse für die beiden Studienrichtungen getrennt anzusehen, siehe Tabelle (2.10).

Psychologie	Frauen	Männer
Bewerber*innen	100	400
zugelassen	80	280
in Prozent	80%	70%

Mathematik	Frauen	Männer
Bewerber*innen	400	100
zugelassen	120	20
in Prozent	30%	20%

Tabelle 2.10: Zulassungsquoten für Frauen und Männer in zwei Studienrichtungen. Obwohl insgesamt 60% der Männer und nur 40% der Frauen zugelassen wurden, ist die Zulassungsquote der Frauen in beiden Studienrichtungen höher als die der Männer.

Das Ergebnis verblüfft, offensichtlich haben Frauen in beiden Studienrichtungen eine deutlich höhere Zulassungsquote als Männer, obwohl die aggregierten Zahlen das Gegenteil vermuten ließen. Wie kann das geschehen?

Das Ergebnis scheint auf den ersten Moment widersprüchlich, aber die Zahlen sprechen eine eindeutige Sprache.

Grafisch kann man sich dieses Paradox auch am Beispiel einer Vektoraddition verdeutlichen, siehe Abbildung 2.26.

Dieses Beispiel ist nicht bei den Haaren herbeigezogen, dieser Fall trat z.B. bei Zulassungen zu Graduate Schools der University of California, Berkeley, im Herbst 1973 auf. Die Zahlen zeigten, dass insgesamt mehr Männer als Frauen zugelassen wurden, und die Differenz war so groß, dass sie nicht durch Zufall erklärt werden konnte. Die Aufschlüsselung nach Fakultäten zeigte allerdings, dass Frauen nicht diskriminiert wurden, sondern im Gegenteil leicht, aber statistisch signifikant, bevorzugt wurden (siehe `https://de.wikipedia.org/wiki/Simpson-Paradoxon`).

Benannt wurde dieses Paradoxon nach Edward Simpson, der die Möglichkeit eines solchen paradoxen Ergebnisses 1951 publizierte. Allerdings wurden die Folgen von *'omitted variables'* bereits weit früher von Karl Pearson (1899) und Udny Yule (1903) beschrieben.

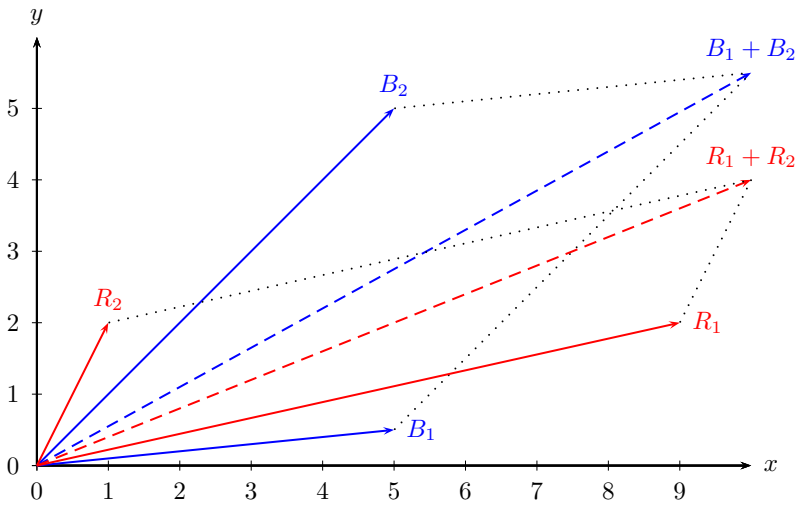


Abbildung 2.26: Simpson-Paradox: obwohl beide rote Vektoren (R) steiler sind als die blauen Vektoren (B) ist die Vektorsumme der blauen Vektoren steiler als die Vektorsumme der roten Vektoren.

Dieses scheinbare Paradox können wir mit den bisher vorgestellten Techniken in einem Regressionszusammenhang einfach reproduzieren.

Im Kern *kann* ein solches Ergebnis immer auftreten, wenn wesentliche Aspekte – wie hier die Studienrichtung – in der aggregierten Untersuchung nicht berücksichtigt wurden, insbesondere wenn sich die Größe und die Anteile der Gruppen der Unterkategorien stark unterscheiden (z.B. beträgt in obigem Beispiel der Anteil der Frauen in der Psychologie nur 20%, in der Mathematik aber 80%).

Wie dieses Beispiel demonstriert kann der Blick auf die aggregierten Daten also zu völlig irreführenden Schlussfolgerungen führen kann, und dass die Berücksichtigung von Gruppenunterschieden manchmal essentiell sein kann. Das Problem ist, dass wir nie wirklich sicher sein können alle entscheidenden Gruppenunterschiede berücksichtigt zu haben. Es wäre also durchaus möglich (wenngleich nicht sehr wahrscheinlich), dass sich die Resultate bei Berücksichtigung weiterer Merkmale wieder umkehren.

Dahinter verbirgt sich ein bekanntes Phänomen, nämlich die Nichtberücksichtigung relevanter Variablen (in diesem Fall kategoriale Variablen für die Fakultätenzuordnung), die zu einem *'omitted variables bias'* führt.

Solche Probleme werden in der Statistik unter dem Schlagwort *'confounding'* diskutiert, in der Ökonometrie sind diese Probleme bekannt als *'unobserved heterogeneity'* oder *'omitted variables bias'*. Da wir nie wissen können, ob wir wirklich alle relevanten Einflussfaktoren berücksichtigt haben, sollten wir bei der Interpretation empirischer Ergebnisse stets vorsichtig bleiben, insbesondere bei Kausalaussagen!

Natürlich wird ein solch extremes Ergebnis nicht immer eintreten, aber allein die Tatsache, dass es eintreten kann, sollte uns vorsichtig stimmen.

Um dieses Ergebnis in einem Regressionszusammenhang zu analysieren benötigen wir lediglich drei Dummy Variablen: zugelassen ja/nein, weiblich ja/nein, Psychologie ja/nein. Die Referenzkategorie sind also Männer, bzw. für die zweite Gleichung Männer, die Mathematik studieren.

Da jede der drei Dummies zwei mögliche Ausprägungen hat benötigen wir insgesamt 6 Kombinationen. Diese werden im Script 2.3 mit der *replication* Funktion von R erzeugt (`rep()`, z.B. erzeugt `rep(1, 80)` einen Vektor der Länge 80 mit lauter Einsen) und packen diese in einen data.frame (z.B. 80 Frauen wurden für das Studium Psychologie zugelassen). Im zweiten data.frame werden die 20 Frauen, die *nicht* zugelassen wurden, mit der rbind() Funktion an den ersten data.frame angefügt, usw.

Die Regression in Tabelle 2.11 zeigt das Resultat.

Tabelle 2.11: Simpson Paradox: Dummy Variablen

	Dependent variable:	
	zugelassen	
	(1)	(2)
Constant	0.600^{***}	0.200^{***}
	(0.022)	(0.035)
fem	-0.200^{***}	0.100^{***}
	(0.031)	(0.035)
psych		0.500^{***}
		(0.035)
Observations	1,000	1,000
R^2	0.040	0.200
Note:	*p<0.1; **p<0.05; ***p<0.01	

Wir erinnern uns, dass eine Regression auf Dummy Variablen die Anteile liefert, und da es sich um ein gesättigtes Dummy Variablen Modell handelt, erhalten wir die exakten Anteile.

Wir erinnern uns, dass der *omitted variables bias* in diesem einfachen Modell folgendermaßen berechnet werden kann

$$b_2^{(1)} = b_2^{(2)} + b_3^{(2)} \frac{\text{cov}(\text{fem}, \text{psych})}{\text{var}(\text{psych})} = 0.1 + 0.5 * (-0.6) = -0.2$$

Script 2.3: R Beispiel zu Simpson-Paradox

```
## Simpson Paradox, Example

# Psychologie Frauen Männer
# Bewerber*innen 100 400
# davon zugelassen 80 280

# Mathematik Frauen Männer
# Bewerber*innen 400 100
# davon zugelassen 120 20

# female, psych
s <- data.frame(zugelassen = rep(1, 80),
                fem = rep(1, 80),
                psych = rep(1, 80))
s <- rbind(s, data.frame(zugelassen = rep(0, 20),
                         fem = rep(1, 20),
                         psych = rep(1, 20)))

# male, psych
s <- rbind(s, data.frame(zugelassen = rep(1, 280),
                         fem = rep(0, 280),
                         psych = rep(1, 280)))
s <- rbind(s, data.frame(zugelassen = rep(0, 120),
                         fem = rep(0, 120),
                         psych = rep(1, 120)))

# female, math
s <- rbind(s, data.frame(zugelassen = rep(1, 120),
                         fem = rep(1, 120),
                         psych = rep(0, 120)))
s <- rbind(s, data.frame(zugelassen = rep(0, 280),
                         fem = rep(1, 280),
                         psych = rep(0, 280)))

# male, math
s <- rbind(s, data.frame(zugelassen = rep(1, 20),
                         fem = rep(0, 20),
                         psych = rep(0, 20)))
s <- rbind(s, data.frame(zugelassen = rep(0, 80),
                         fem = rep(0, 80),
                         psych = rep(0, 80)))


eq_short <- lm(zugelassen ~ fem, data = s)
eq_long <- lm(zugelassen ~ fem + psych, data = s)

stargazer::stargazer(eq_short, eq_long, type = "text",
                     intercept.bottom = FALSE)

# omitted variable
coef(eq_long)[2] + coef(eq_long)[3]*cov(s$fem, s$psych)/var(s$psych)
## -0.2
```

	$i=1$ $\mathbf{y_1}$	$i=2$ $\mathbf{y_2}$	$i=3$ $\mathbf{y_3}$	$i=1$ $\mathbf{x_{21}}$	$i=2$ $\mathbf{x_{22}}$	$i=3$ $\mathbf{x_{23}}$	$i=1$ $\mathbf{x_{31}}$	$i=2$ $\mathbf{x_{32}}$	$i=3$ $\mathbf{x_{33}}$
$t=1$	y_{11}	y_{21}	y_{31}	x_{211}	x_{221}	x_{231}	x_{311}	x_{321}	x_{331}
$t=2$	y_{12}	y_{22}	y_{32}	x_{212}	x_{222}	x_{232}	x_{312}	x_{322}	x_{332}
$t=3$	y_{13}	y_{23}	y_{33}	x_{213}	x_{223}	x_{233}	x_{313}	x_{323}	x_{333}
$t=4$	y_{14}	y_{24}	y_{34}	x_{214}	x_{224}	x_{234}	x_{314}	x_{324}	x_{334}
mean	$\bar{y}_{1\bullet}$	$\bar{y}_{2\bullet}$	$\bar{y}_{3\bullet}$	$\bar{x}_{21\bullet}$	$\bar{x}_{22\bullet}$	$\bar{x}_{23\bullet}$	$\bar{x}_{31\bullet}$	$\bar{x}_{32\bullet}$	$\bar{x}_{33\bullet}$

zum Beispiel:

	GDP_DEU	GDP_AUT	GDP_ITA	C_DEU	C_AUT	C_ITA	I_DEU	I_AUT	I_ITA
2020	y_{11}	y_{21}	y_{31}	x_{211}	x_{221}	x_{231}	x_{311}	x_{321}	x_{331}
2021	y_{12}	y_{22}	y_{32}	x_{212}	x_{222}	x_{232}	x_{312}	x_{322}	x_{332}
2022	y_{13}	y_{23}	y_{33}	x_{213}	x_{223}	x_{233}	x_{313}	x_{323}	x_{333}
2023	y_{14}	y_{24}	y_{34}	x_{214}	x_{224}	x_{234}	x_{314}	x_{324}	x_{334}

Tabelle 2.12: Paneldaten im *'wide'* Format.

2.7.6 Beispiel: Das LSDV und *'Fixed Effects'* Modell

Das bisherige Wissen gestattet uns bereits Einsichten in eines der wichtigsten Modelle der angewandten Ökonometrie, in das *'Fixed Effects'* Modell für Panel Daten.

Häufig beobachten wir mehrere Individuen (Länder, Firmen, Personen, ...) über mehrere Zeitperioden, z.B. das BIP aller OECD Länder von 2005 – 2016, die Bruttolöhne aller Beschäftigten einer Firma über die letzten vier Jahre, mittlere Tages-Temperatur an verschiedenen Messstationen über die letzten 200 Jahre.

Tabelle 2.12 zeigt ein Beispiel für 3 Individuen, 4 Zeitperioden, und 2 Regressoren (x_2, x_3).

Wenn die Daten zwei Dimensionen haben (z.B. Länder und Zeitperioden) benötigen wir zwei Indizes (*'Identifier'*) um eine Beobachtung zu identifizieren; y_{it} bezeichnet z.B. den Wert von y für Individuum i in Periode t, wobei $i = 1, \ldots, n$ über die Individuen und $t = 1, \ldots, T$ über die Zeit läuft.

Diese zweidimensionale Datenstruktur (Individuen und Zeit) ermöglicht verschiedene Auswertungen. Man könnte z.B. für jedes einzelne Individuum eine Regression über die Zeit rechnen, aber in den meisten Fällen wäre dies wenig hilfreich, z.B. wenn wir Daten für mehrere tausend Individuen haben. Genauso könnten wir für jede Periode eine Querschittsregression rechnen, aber auch diese Information ist selten von Interesse.

Eine dritte Möglichkeit wäre von allen Variablen die Zeitmittelwerte zu bilden, und über diese Mittelwerte eine Querschnittsregression zu rechnen. Wenn wir den Durchschnitte über die Zeit mit $\bar{y}_{i\bullet} = 1/T \sum_{t=1}^{T} y_{it}$ notieren (vgl. Abbildung 2.12) erhalten wir das so genannte *'between'* Modell

$$\bar{y}_{i\bullet} = b_1 + b_2 \bar{x}_{2i\bullet} + b_3 \bar{x}_{3i\bullet} + e_i$$

allgemein: ($n = 3,\ T = 4$)

i	t	y	x_2	x_3
1	1	y_{11}	x_{211}	x_{311}
1	2	y_{12}	x_{212}	x_{312}
1	3	y_{13}	x_{213}	x_{313}
1	4	y_{14}	x_{214}	x_{314}
2	1	y_{21}	x_{221}	x_{321}
2	2	y_{22}	x_{222}	x_{322}
2	3	y_{23}	x_{223}	x_{323}
2	4	y_{24}	x_{224}	x_{324}
3	1	y_{31}	x_{231}	x_{331}
3	2	y_{32}	x_{232}	x_{332}
3	3	y_{33}	x_{233}	x_{333}
3	4	y_{34}	x_{234}	x_{334}

Beispiel:

i	t	GDP	Cons	Inv
DEU	2020	y_{11}	x_{211}	x_{311}
DEU	2021	y_{12}	x_{212}	x_{312}
DEU	2022	y_{13}	x_{213}	x_{313}
DEU	2023	y_{14}	x_{214}	x_{314}
AUT	2020	y_{21}	x_{221}	x_{321}
AUT	2021	y_{22}	x_{222}	x_{322}
AUT	2022	y_{23}	x_{223}	x_{323}
AUT	2023	y_{24}	x_{224}	x_{324}
ITA	2020	y_{31}	x_{231}	x_{331}
ITA	2021	y_{32}	x_{232}	x_{332}
ITA	2022	y_{33}	x_{233}	x_{333}
ITA	2023	y_{34}	x_{234}	x_{334}

Tabelle 2.13: Paneldaten im *'long'* Format (*'gestackt'*). Dieses Format kann in R z.B. mit dem `reshape2` package aus dem *wide* Format erzeugt werden (auch in Stata gibt es einen reshape-Befehl).

Die Bezeichnung *'between'* Modell kommt daher, weil nur die Heterogenität *zwischen* den Individuen modelliert wird, die Streuung über die Zeit 'innerhalb' der Individuen bleibt unberücksichtigt.

Eine alternative Lösung mit maximaler Informationsverdichtung wäre, einfach eine Regression über alle Beobachtungen zu rechnen. Dazu müssen die Daten zuerst umgeordnet werden, d.h., sie müssen zuerst in das *'long'* Format transformiert werden, indem die Daten entsprechend angeordnet werden: man 'stapelt' einfach die Beobachtungen für die einzelnen Individuen übereinander (engl. *'stack'*), siehe Tabelle 2.13.[18]

$$y_{it} = b_1^p + b_2^p x_{2it} + b_3^p x_{3it} + e_{it}^p$$

Dieses Modell impliziert, dass die Koeffizienten b_1^p bzw. b_2^p für alle Länder und Zeitperioden den gleichen Wert haben und wird auch *Pool-Modell* genannt (deshalb der hochgestellte Index p).

Dieses 'gepoolte' Modell kann ganz normal mit OLS geschätzt werden.

Für 3 Individuen und 4 Zeitperioden und zwei erklärende Variablen würde das 'gepoolte' Modell mit den '*stacked data*' in Vektorschreibweise folgendermaßen aussehen ($i = 1, \ldots, 3,\ t = 1, \ldots, 4$)

[18] Diese Anordnung der Daten muss natürlich nicht manuell erfolgen, alle Programme verfügen über spezielle Befehle für diese Umorganisation der Daten. In Stata gibt es dafür die Befehle `xtset` und `reshape`; in R z.B. mit dem Package `reshape2`, welches die sehr flexiblen Befehle `melt` und `dcast` zur Verfügung stellen. Mit dem Package `plm` können schließlich alle Arten von Panelmodellen geschätzt werden.

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{24} \\ y_{31} \\ y_{32} \\ y_{33} \\ y_{34} \end{pmatrix} = b_1 \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + b_2 \begin{pmatrix} x_{2\,11} \\ x_{2\,12} \\ x_{2\,13} \\ x_{2\,14} \\ x_{2\,21} \\ x_{2\,22} \\ x_{2\,23} \\ x_{2\,24} \\ x_{2\,31} \\ x_{2\,32} \\ x_{2\,33} \\ x_{2\,34} \end{pmatrix} + b_3 \begin{pmatrix} x_{3\,11} \\ x_{3\,12} \\ x_{3\,13} \\ x_{3\,14} \\ x_{3\,21} \\ x_{3\,22} \\ x_{3\,23} \\ x_{3\,24} \\ x_{3\,31} \\ x_{3\,32} \\ x_{3\,33} \\ x_{3\,34} \end{pmatrix} + \begin{pmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{14} \\ e_{21} \\ e_{22} \\ e_{23} \\ e_{24} \\ e_{31} \\ e_{32} \\ e_{33} \\ e_{34} \end{pmatrix}$$

Die Annahme, dass die Koeffizienten für alle Individuen und Perioden gleich sind, ist natürlich ziemlich restriktiv.

Ein etwas allgemeineres und flexibleres Modell, welches in der Praxis am häufigsten angewandt wird, erlaubt individuenspezifische Interzepte, aber unterstellt für alle Länder die gleichen Steigungskoeffizienten. Dies kann einfach mit Hilfe entsprechender individuenspezifischer Dummy Variablen bewerkstelligt werden. Wir würden z.B. das folgende Modell mit OLS schätzen

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{24} \\ y_{31} \\ y_{32} \\ y_{33} \\ y_{34} \end{pmatrix} = a_1 \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + a_2 \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + a_3 \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + b_2 \begin{pmatrix} x_{2\,11} \\ x_{2\,12} \\ x_{2\,13} \\ x_{2\,14} \\ x_{2\,21} \\ x_{2\,22} \\ x_{2\,23} \\ x_{2\,24} \\ x_{2\,31} \\ x_{2\,32} \\ x_{2\,33} \\ x_{2\,34} \end{pmatrix} + b_3 \begin{pmatrix} x_{3\,11} \\ x_{3\,12} \\ x_{3\,13} \\ x_{3\,14} \\ x_{3\,21} \\ x_{3\,22} \\ x_{3\,23} \\ x_{3\,24} \\ x_{3\,31} \\ x_{3\,32} \\ x_{3\,33} \\ x_{3\,34} \end{pmatrix} + \begin{pmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{14} \\ e_{21} \\ e_{22} \\ e_{23} \\ e_{24} \\ e_{31} \\ e_{32} \\ e_{33} \\ e_{34} \end{pmatrix}$$

wobei wir jetzt für eine bessere Lesbarkeit die Koeffizienten der Dummies und das Interzept mit a bezeichnen (in den meisten Fällen interessieren wir uns nicht für diese Koeffizienten der Dummies, weshalb sie in Publikationen häufig unterdrückt werden).

Hier dient das erste Individuum als Referenzkategorie.

Dieses Modell wird auch *'Least Squares Dummy Variable'* (LSDV) Modell genannt und kann für insgesamt n Individuen und T Zeitperioden folgendermaßen geschrieben werden

$$y_{it} = a_1 + \sum_{i=2}^{n} a_i d_i + b_2 x_{2it} + \cdots + b_k x_{kit} + e_{it}$$

mit $i = 1, \ldots, n$ und $t = 1, \ldots, T$. Man beachte, dass die Dummies d_i keinen Zeitindex benötigen (sie sind *zeitinvariant*)!

Meist wird (insbesondere das *'fixed effects'* Modell) aber kürzer geschrieben

$$y_{it} = a_i + b_2 x_{2it} + \cdots + b_k x_{kit} + e_{it}$$

wobei die a_i symbolisch für die Individueneffekte stehen (z.B. Ländereffekte), also die Individuen-Dummies symbolisieren (d.h. $a_i := a_1 + \sum_{i=2}^{n} a_i d_i$).

Stellen Sie sich vor, Sie möchten dieses Modell für ein Panel mit mehreren tausend Individuen schätzen, dann würden Sie mehrere tausend Dummy Variablen benötigen. Dies würde sogar die Rechenleistung moderner Computer herausfordern.

Glücklicherweise gibt es eine ebenso einfache wie elegante Alternative. Erinnern wir uns an das Frisch-Waugh-Lovell (FWL) Theorem: wir können den linearen Einfluss von Variablen 'eliminieren', indem wir in einem ersten Schritt Hilfsregressionen rechnen, und anschließend die Residuen dieser Hilfsregressionen verwenden.

Wir könnten also in einem ersten Schritt die y und x auf die Dummy Variablen regressieren, und dann die Residuen dieser Hilfsregressionen verwenden, um die interessierenden Steigungskoeffizienten b_h zu berechnen.

Damit scheint im ersten Moment nicht viel gewonnen, für die Hilfsregressionen benötigen wir weiterhin alle Dummy Variablen. Aber überlegen wir, was wir aus einer Regression des gesättigten Dummy Variablen Modells erhalten, genau, *die gruppenspezifischen Mittelwerte* (z.B. Mittelwerte von Frauen und Männern, länderspezifische Mittelwerte, etc.)! Und die Residuen der Hilfsregressionen sind einfach die Abweichungen von diesen gruppenspezifischen Mittelwerten.

Es genügt also, für jedes Individuum und für alle Variablen die Gruppen-Mittelwerte über die Zeit zu bilden, und die individuenspezifischen Abweichungen von diesen Gruppen-Mittelwerten zu berechnen. Eine solche individuenspezifische Mittelwerttransformation kann von Computern sehr effizient und schnell durchgeführt werden.

Also, anstatt eine Regression mit potentiell mehreren tausend Dummy Variablen zu berechnen können wir auch individuenspezifische Mittelwerttransformationen durchführen und eine einfache OLS Regression auf die derart transformierten Daten berechnen

$$(y_{it} - \bar{y}_{i\bullet}) = b_2(x_{2it} - \bar{x}_{2i\bullet}) + b_3(x_{3it} - \bar{x}_{3i\bullet}) + e_{it}$$

wobei $\bar{y}_{i\bullet}$, $\bar{x}_{2i\bullet}$ und $\bar{x}_{3i\bullet}$ individuenspezifischen Zeit-Mittelwerte sind. Das heißt, wir müssen nur die individuenspezifische Mittelwerttransformation durchführen, und können mit den so transformierten Daten eine ganz normale OLS Regression anwenden.

Das nach dieser Methode geschätzte Modell wird *'fixed effects model'* genannt (die Individueneffekte, z.B. Ländereffekte, ändern sich nicht über die Zeit, sind also '*fixed*').

Aufgrund des FWL Theorems führt diese Methode numerisch zu den numerisch exakt gleichen Schätzungen für die Steigungskoeffizienten wie das LSDV Modell, ist aber viel einfacher zu berechnen. Aber Vorsicht, dies gilt nur für die Steigungskoeffizienten und Residuen, die Standardfehler werden sich bei diesen Methoden unterscheiden, weil das *'fixed effects model'* nicht den Verlust von Freiheitsgraden bei der

individuenspezifischen Mittelwerttransformation berücksichtigt. Alle Computerprogramme, die *'fixed effects'* Modelle unterstützen, berücksichtigen dies automatisch und geben die korrekten Standardfehler aus.

Man verliert bei dieser *fixed effects* Methode zwar die einzelnen Individueneffekte a_i (d.h. die Koeffizienten der Individuendummies), aber diese sind ohnehin selten von Interesse, und könnten obendrein ex post wieder berechnet werden.

Zusammenfassend, aufgrund des FWL Theorems können wir die Koeffizienten des *'fixed effects'* Modells gleich interpretieren wie die Koeffizienten eines Least Squares Dummy Variablen Modells (LSDV)!

Da dieses Modell nur die Streuung über die Zeit *'innerhalb'* der Individuen berücksichtigt, wird das *'fixed effects'* Modell auch *'within'* Modell genannt.

Der besondere Reiz des *'fixed effects'* Modells liegt darin, dass die Individueneffekte (bzw. die Dummies für die Individuen) für alles kontrollieren, was sich nicht über die Zeit ändert, d.h. für alle *zeitinvarianten* Effekte (wie z.B. Geschlecht, koloniale Vergangenheit, ...).

Die Individuen-Dummies 'schlucken' gewissermaßen alles was zeitinvariant ist, egal ob wir es beobachten können oder nicht, oder ob wir uns dafür interessieren oder nicht. Das hat zur Folge, dass wir mit Hilfe des *'fixed effects'* Modells keine partiellen Effekte von zeitinvarianten Variablen berechnen können!

Rein technisch wird dies schon daraus ersichtlich, dass individuenspezifische Mittelwerttransformationen für zeitinvariante Variablen immer den Wert Null liefern. Die meisten Computerprogramme unterdrücken solche Variablen automatisch, andere Programme brechen mit einer Fehlermeldung ab.

Nehmen wir zum Beispiel an, wir hätten Paneldaten mit Stundenlöhnen, abgeschlossenem Bildungsniveau, Berufserfahrung und Geschlecht von vielen Personen über viele Jahre.

Wenn wir ein *'fixed effects'* Modell schätzen kontrollieren die (impliziten) Personen-Dummies zwar für *alle* individuenspezifischen Effekte, also z.B. auch für die unbeobachtbare 'emotionale Intelligenz' (wenn sich diese nicht im Zeitablauf ändert!), aber da zugleich auch das Geschlecht und die abgeschlossene Bildung zeitinvariant sind, können wir deren Einfluss ebenfalls nicht messen, sie 'stecken' gewissermaßen alle gemeinsam in den Individuen-Dummies. Wenn obendrein die Berufserfahrung für alle Personen jedes Jahr um ein Jahr zunimmt verlieren wir die *'between'* Information, es bleibt lediglich ein *'within'* Trend erhalten, der ebenso die Auswirkungen der Inflation und ähnliches messen könnte.

Wenn wir uns aber für Variablen *mit* Zeitvariation interessieren ist das *'fixed effects'* Modell äußerst mächtig, da es automatisch für *alle* zeitinvarianten Effekte kontrolliert, egal ob diese beobachtet werden oder nicht.

2.7.7 Mehrere kategoriale Variablen

Im vorhergehenden Abschnitt untersuchten wir sich gegenseitig ausschließende Kategorien, jeder Beobachtung konnte genau eine Ausprägung der kategorialen Variable zugeordnet werden; ein Auto kann z.B. nicht gleichzeitig zwei und vier Jahre alt

sein. In diesem Fall waren die Kategorien disjunkt, und deshalb sind die Interaktionseffekte Null.

Nun sehen wir uns den Fall mit mehreren kategorialen Variablen an, die sich gegenseitig nicht ausschließen; z.B. kann eine Person weiblich sein und als weiteres Merkmal verheiratet oder unverheiratet sein. In diesem Fall sind die Kategorien *nicht* disjunkt und die Interaktionseffekte sind deshalb in der Regel ungleich Null.

Kehren wir nochmals zurück zum Beispiel mit den Stundenlöhnen, siehe 2.7 (Seite 58). Wie erwartet liefert eine Regression auf die Dummy Variable $v_i = 1$ für verheiratet und 'Null sonst' als Interzept den mittleren Stundenlohn der Referenzkategorie, d.h. Unverheirateter, und der Koeffizient der Dummy v zeigt, dass Verheiratete durchschnittlich um 0.6 Euro mehr verdienen,

$$\widehat{\text{StdL}}_i = 13.4 + 0.6 v_i$$

Man könnte vielleicht vermuten, dass eine Regression auf beide Dummy Variablen ($m_i = 1$ für männlich und Null sonst, und $v_i = 1$ für verheiratet und Null sonst) die beiden Abweichungen von der Referenzkategorie ($m_i = 0$ und $v_i = 0$, also eine unverheiratete Frau) misst, aber dem ist nicht so, wie das Ergebnis zeigt

$$\widehat{\text{StdL}}_i = 12.41 + 2.47 m_i + 0.18 v_i$$

Was ist passiert? Wir haben ganz einfach einen Denkfehler gemacht, denn die beiden Dummy Variablen definieren *vier* Kategorien, nicht zwei! Die folgende Tabelle zeigt die vier Kategorien und die jeweiligen Mittelwerte von StdL für dieses Beispiel

		männlich	
		ja (1)	nein (0)
ver-	ja (1)	15.5	12
heiratet	nein (0)	14	13

Wenn wir nur auf die zwei Kategorien m und v regressieren schätzen wir ein zu kurzes Modell, und es tritt wieder das Problem der Nichtberücksichtigung relevanter Variablen (*omitted variables'*) auf.

Nur wenn wir ein Modell wählen, das *alle möglichen* Kategorien berücksichtigt, erhalten wir als Koeffizienten die Mittelwerte der jeweiligen Kategorien.

Ein solches Modell wird *gesättigt* (*'saturated'*) genannt, und nur für solche gesättigten Dummy Variablen Modelle gilt, dass das Interzept den Mittelwert der Referenzkategorie misst, und die Koeffizienten der Dummy Variablen die entsprechenden durchschnittlichen Abweichungen der jeweiligen Kategorie von der Referenzkategorie.

Die einfachste Möglichkeit ein solches Modell zu schätzen besteht darin, für alle Katergorien mit Ausnahme der gewählten Referenzkategorie eine Dummy Variable zu generieren. Wenn wir z.B. unverheiratete Männer ($m = 1$ und $v = 0$) als Referenzkategorie wählen erzeugen wir eine Dummy Variable mv für männlich verheiratet, mit $mv = m \times v$, für weiblich unverheiratet $wu = w \times (1 - v)$, und für weiblich verheiratet $wv = (1 - m) \times v$. Die Regression liefert

$$\widehat{\text{StdL}} = 14.0 + 1.5\, mv - 1.0\, wu - 2.0\, wv$$

Damit erhalten wir das erwartete Ergebnis, der Durchschnitts-Stundenlohn unverheirateter Männer beträgt 14 Euro, verheiratete Männer verdienen im Durchschnitt 1.5 Euro mehr, unverheiratete Frauen verdienen durchschnittlich um einen Euro weniger als unverheiratete Männer, und verheiratete Frauen um zwei Euro weniger.

Diese Parametrisierung liefert direkt einen sehr einfach zu interpretierenden Output. In der Literatur findet man hingegen häufig eine alternative Parametrisierung, die exakt das selbe Ergebnis in einer andern Darstellung liefert, nämlich eine Regression auf beide Dummy Variablen *und* auf den Interaktionseffekt (d.h. das Produkt) der beiden Dummy Variablen. Wenn wir als Referenzkategorie 'weiblich' ($m = 0$) und 'unverheiratet' ($v = 0$) wählen erhalten wir

$$\widehat{\text{StdL}} = 13.0 + 1.0\,m - 1.0\,v + 2.5\,(m \times v)$$

Wir können uns die Dummy Variablen hier gewissermaßen als *'Ein-Aus-Schalter'* vorstellen, falls die Dummy Variable den Wert Eins hat wird der Koeffizient 'eingeschaltet', sonst 'ausgeschaltet', und der Interaktionseffekt ist nur 1 ('eingeschaltet'), wenn *beide* Dummy Variablen den Wert 1 haben.
Wir können die Fälle einfach durchgehen:

1. Weiblich unverheiratet: (Referenzkategorie)
$$\widehat{\text{StdL}}|(m = 0, v = 0) = 13.0 + 1.0 \times 0 - 1.0 \times 0 + 2.5(0 \times 0) = 13$$
2. Weiblich verheiratet:
$$\widehat{\text{StdL}}|(m = 0, v = 1) = 13.0 + 1.0 \times 0 - 1.0 \times 1 + 2.5(0 \times 1) = 12$$
3. Männlich unverheiratet:
$$\widehat{\text{StdL}}|(m = 1, v = 0) = 13.0 + 1.0 \times 1 - 1.0 \times 0 + 2.5(1 \times 0) = 14$$
4. Männlich verheiratet:
$$\widehat{\text{StdL}}|(m = 1, v = 1) = 13.0 + 1.0 \times 1 - 1.0 \times 1 + 2.5(1 \times 1) = 15.5$$

Wie man sich leicht überzeugen kann liefert diese Parametrisierung exakt das gleiche Ergebnis in einer etwas anderen Darstellung, durch die Berücksichtigung des Interaktionseffekts ist das Modell wieder gesättigt, deshalb spielt es formal keine Rolle, welche dieser Parametrisierungen man wählt, es ist eher eine Frage der Zweckmäßigkeit.

Wir haben bisher nur zwei binäre Merkmale untersucht. Natürlich ist es häufig verlockend sich komplexere Kombinationen von Merkmalsausprägungen anzusehen. Für eine typische Lohngleichung ist z.B. häufig nicht nur das Geschlecht und der Familienstand relevant, sondern auch der Bildungsgrad, die Berufserfahrung, in welcher Branche die oder der Beschäftigte arbeitet, und vieles mehr.
Wer allerdings alle diese Merkmalskombinationen mit Dummy Variablen abbilden will landet schnell im *Fluch der Dimensionalität*. Wer z.B. neben (binären) Geschlecht und Familienstand noch sechs Bildungsniveaus, 10 Branchen und vier Regionen unterscheiden möchte landet bei 960 zu unterscheidenden Kategorien. Selbst wenn die Stichprobe groß genug wäre um all die Parameter zu schätzen dürfte es nicht ganz einfach sein, ein solches Ergebnis übersichtlich zu präsentieren.

2.7.8 Beispiel: *'Difference-in-Differences'* Modelle

Stellen Sie sich vor, in einer Stadt wurde eine neue Umfahrungsstrasse gebaut, und Sie werden beauftragt zu schätzen, welche Auswirkungen dies auf die Immobilienpreise *in der betroffenen Region* hatte.

Dieser Auftrag stellt Sie vor eine typische "Was-wäre-wenn" Frage, denn wenn die Straße gebaut wurde fehlt das Kontrafaktum (engl. *counterfactual*; wie wären die Preise, wenn die Straße *nicht* gebaut worden wäre).

Angenommen Sie hätten Daten über die Grundstückpreise *vor* dem Bau der Umfahrungsstrasse. In diesem Fall könnten Sie einfach den Mittelwert der Grundstückpreise *vor* dem Bau der Umfahrungsstrasse mit den Grundstückpreisen *nach* dem Bau der Umfahrungsstrasse vergleichen.

Allerdings ist ein solcher Vergleich schwierig, denn wenn sich während des Baus der Umfahrungsstrasse die Immobilienpreise generell verändert haben, würde man diese Preisänderung fälschlich der Umfahrungsstraße zuschreiben.

In diesem Fall könnte man die Preise *vor* und *nach* dem Bau der Umfahrungsstrasse mit den Grundstückpreisen einer sehr ähnlichen, aber von der Intervention *nicht betroffenen Region* der Stadt vergleichen. Genau dies ist das Grundprinzip des *"Difference-in-Differences"* Ansatzes.

Da diese Art von Analysen früher hauptsächlich in der Medizin angewandt wurden, haben sich in der Literatur die medizinische Terminologie eingebürgert. Man nennt eine Gruppe, der einer Behandlung (*Intervention*) zuteil wurde (bzw. die von einer Veränderung betroffen wurde), *'Treatment Group'*, und die Kontrollgruppe wenig überraschend *'Control Group'*. Um nicht in einer babylonischer Sprachverwirrung zu enden bezeichnen wir die Periode vor und nach dem (*'Treatment'*) mit *'Before'* und *'After'*.

Woher die Bezeichnung *'Difference-in-Differences'* kommt wird unmittelbar klar, wenn wir zum Beispiel zurückkehren. Wir bezeichnen den Mittelwert der Grundstückpreise der *'Treatment Group'* (d.h. der Gruppe, die vom Bau betroffen war) *vor* dem Bau der Umfahrungsstrasse mit T_B, den Mittelwert der *'Treatment Group'* *nach* dem Bau der Umfahrungsstrasse mit T_A, und die Mittelwerte der Preise der Kontrollgruppe mit C_B bzw. C_A, also

	Treatment Group	Control Group
Before	T_B	C_B
After	T_A	C_A

Um die vom Bau der Umfahrungsstrasse '*verursachte*' Preisänderung abzuschätzen können wir einfach die 'Differenz der Differenz' der Mittelwerte bilden, also

$$\text{"Difference-in-Differences"} = (T_A - T_B) - (C_A - C_B)$$

Damit haben wir unser Problem aber erst *fast* gelöst, denn wir werden kaum genügend *vergleichbare* Immobilienpreise in den Gruppen finden. Immobilien unterscheiden sich in Bezug auf Größe, Lage, Ausstattung usw., so dass ein Vergleich schwierig ist.

Glücklicherweise lässt sich dieser "*Difference-in-Differences*" Ansatz sehr einfach in ein Regressionsmodell überführen, und eine Regression erlaubt bekanntlich die Berücksichtigung mehrerer erklärender x Variablen (wie z.B. Größe, Lage, Ausstattung).

Konkret können wir folgende Regressionsgleichung schätzen

$$y_i = b_1 + b_2\text{treat} + b_3\text{after} + b_4(\text{treat} \times \text{after}) + b_5 x_i + e_i$$

mit den Dummies

$$\text{treat} = \begin{cases} 1 & \text{wenn in 'Treatment Group',} \\ 0 & \text{wenn in 'Control Group'.} \end{cases} \qquad \text{after} = \begin{cases} 0 & \text{vor 'Treatment',} \\ 1 & \text{nach 'Treatment'.} \end{cases}$$

und einer (oder mehreren) erklärenden Variablen x.

In der folgenden Tabelle kann man einfach erkennen, dass der Koeffizient des *Interaktionsterms* zwischen der Treatment- und After-Dummy genau der Difference-in Difference Schätzer ist.

	Treatment Group	Control Group	Difference
Before	$b_1 + b_2 + b_5 x$	$b_1 + b_5 x$	b_2
After	$b_1 + b_2 + b_3 + b_4 + b_5 x$	$b_1 + b_3 + b_5 x$	$b_2 + b_4$
Difference	$b_3 + b_4$	b_3	b_4

In Abbildung 2.27 wird dies grafisch gezeigt.

Übung: Was denken Sie, zeigt Abbildung 2.27 eher die Auswirkungen einer Müllverbrennungsanlage oder die einer Freizeitanlage auf die Immobilienpreise in der Umgebung? Welche Grafik würden Sie im anderen Fall erwarten?

Identifizierende Annahmen: Können wir wirklich sicher sein, dass wir den Effekt des Treatments korrekt gemessen haben? Und was bedeutet 'korrekt'?

Uns interessiert natürlich wie sich die Preise entwickelt hätten, wenn z.B. die Strasse nicht gebaut worden wäre (Kontrafaktum), verglichen mit der tatsächlichen Entwicklung der Preise nach dem Bau der Strasse (Faktum). Nachdem wir Kontrafaktum und Faktum nie gleichzeitig beobachten können, ist dieser Vergleich aber nicht durchführbar.

Beim 'Difference-in-Differences' Modell dient natürlich die Control Gruppe als 'Näherung' für das unbeobachtbare Kontrafaktum. Und die Frage ist, wie gut diese 'Näherung' tatsächlich ist.

Eine zentrale Annahme ist, dass *ohne dem Treatment* in der Control- und Treatment Gruppe die gleiche Entwicklung eingetreten wäre, die so genannte *Parallel Trends Assumption*.

Diese Annahme impliziert, dass das Regressionsmodell richtig spezifiziert wurde, z.B. dass alle relevanten Variablen berücksichtigt wurden (d.h. kein *omitted variables bias*).

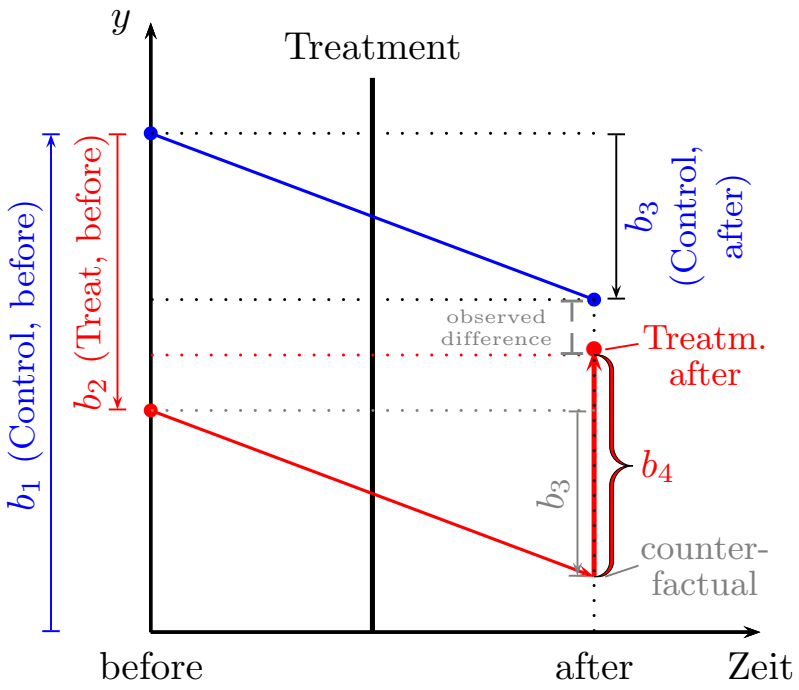


Abbildung 2.27: Difference-in-Differences, der Treatment-Effekt wird durch den Koeffizienten des Interaktionsterms b_4 gemessen;
$\widehat{y}_i = b_1 + b_2\text{treat} + b_3\text{after} + b_4(\text{treat} \times \text{after})$
(hier mit $b_1, b_4 > 0$ und $b_2, b_3 < 0$) Das 'counterfactual' sagt uns, welchen Wert $\widehat{y}$ angenommen hätte, wenn die Treatment Gruppe die gleiche Entwicklung genommen hätte wie die Control Gruppe. Beachten Sie, dass dieses 'counterfactual' nicht direkt beobachtbar ist, sondern auf der identifizierenden Annahme beruht, dass sich in Abwesenheit des Treatments beide Gruppen gleich entwickelt hätten (*Parallel Trends Assumption*).

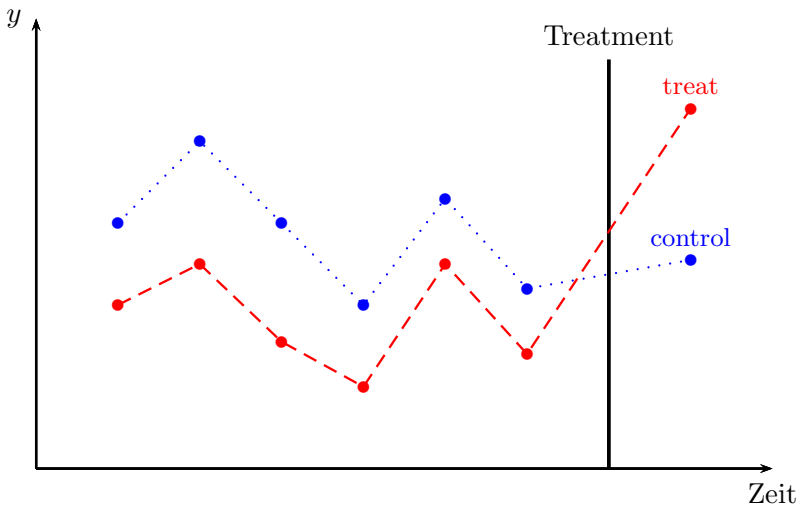


Abbildung 2.28: Eine Kausalinterpretation eines Difference-in-Differences Modells wird 'vertrauenswürdiger', wenn sich *vor dem Treatment* die beiden Gruppen parallel entwickelt haben (das ist aber *kein* Beweis!).

Aber natürlich können wir nie sicher sein, dass in dem hypothetischen Fall *ohne Treatment* in der Treatment-Gruppe das gleiche passiert wäre wie in der Control-Gruppe.

Vor allem gibt es keine Möglichkeit die *Parallel Trends Assumption* zu beweisen, wir können bestenfalls gute Argumente anführen, warum wir vermuten, dass im hypothetischen Fall ohne Treatment in beiden Gruppen das gleiche passiert wäre.

Auch wenn die *Parallel Trends Assumption* nicht bewiesen werden kann, man kann immerhin gute Argumente für deren Gültigkeit suchen. Eines dieser Argumente ist, dass sich – bei wiederholten Beobachtungen – die Trends in beiden Gruppen *vor dem Treatment* annähernd parallel entwickelt haben sollten, siehe Abbildung 2.28.

Eine weitere wichtige Annahme für dass "Difference-in-Differences" Modell ist, dass das Treatment nicht antizipiert wird. Wenn z.B. der Bau einer Strasse bereits angekündigt wurde, führen die *Erwartungen* vermutlich dazu, dass Reaktionen vorweggenommen werden, und deshalb der Effekt des Treatments nicht mehr gemessen werden kann.

Darüber hinaus sollte die Zusammensetzung der Gruppen unverändert bleiben, da z.B. selektive Zu- oder Abwanderung die Ergebnisse verzerren würde. Auch sollten keine systematischen externe Schocks auftreten, die sich auf Control- und Treatment Gruppe unterschiedlich auswirken.

Schließlich darf das Treatment keine indirekten Auswirkungen auf die Control-Gruppe haben, d.h. es dürfen keine 'Spillover' Effekte auftreten. Diese Annahme ist in der Literatur als SUTVA (*Stable Unit Treatment Value Assumption*) bekannt.

Nur wenn diese Annahmen – insbesondere die 'Parallel Trends Assumption' und die Abwesenheit von Antizipationseffekten – erfüllt sind, kann das Ergebnis einer "Difference-in-Differences" Analyse *kausal* interpretiert werden.

Tabelle 2.14: Durchschnittliche Beschäftigtenzahl in Fastfood Restaurants vor und nach Einführung eines Mindestlohns am 1. April 1992 in New Jersey (NJ). Das benachbarte Pennsylvania (PA) dient als Kontrollgruppe. Siehe **?**.

	State		
	PA	NJ	Diff.
Feb	23.33	20.44	−2.89
Nov	21.17	21.03	−0.14
Diff.	2.17	−0.59	2.75

Beispiel Das klassische Beispiel für eine Anwendung des Difference-in-Differences Modells in den Wirtschaftswissenschaften ist eine Studie zu den Auswirkungen einer Erhöhung des Mindestlohnes von **?**. Nicht zuletzt für diese Studie wurde D. Card 2021 mit dem Alfred Nobel Gedächtnispreis für Wirtschaftswissenschaften ausgezeichnet.

Am 1. April 1992 erhöhte New Jersey (NJ) den Mindestlohn von US$4.25 auf US$5.05. **?** erhoben in einer Telefonumfrage bei 320 Fastfood Restaurants in New Jersey und als Kontrollgruppe bei 77 Fastfood Restaurants im benachbarten Pennsylvania die Beschäftigtenzahl. Jede Firma wurde zweimal befragt, einmal vor (Feb) und einmal nach (Nov) Einführung des Mindestlohnes. Fastfood Restaurants wurden gewählt, weil dort der Anteil niedrig bezahlter Beschäftigter besonders hoch ist.

Um den Beschäftigteneffekt der Erhöhung des Mindestlohnes zu ermitteln führten sie eine "Difference-in-Differences" Analyse durch. Die einfachen Mittelwerte und deren Differenzen finden Sie in Tabelle 2.14.

Zur Überraschung vieler Ökonomen beschäftigten die Fastfood Restaurants in der Treatment Gruppe (New Jersey) nach Erhöhung des Mindestlohnes *relativ* mehr Personen als in der Kontrollgruppe Pennsylvania.

Wie vorhin gezeigt kann man dieses Ergebnis auch einfach mit Hilfe einer Regression auf die Dummies NJ (= Treatment Gruppe) und Nov (= After) sowie deren Interaktion erhalten (Empl bezeichnet die Beschäftigtenzahlen).

$$\begin{array}{rcccccccc} \text{Empl} & = & 23.33 & - & 2.89\ \text{NJ} & - & 2.166\ \text{Nov} & + & 2.754\ \text{NJ*Nov} \\ & & (1.072)^{***} & & (1.194)^{**} & & (1.516) & & (1.688) \end{array}$$

$$R^2 = 0.007, \quad n = 794$$
(Standardfehler in Klammern)

Der interessierende Beschäftigungseffekt ist der Koeffizient der Interaktionsvariable NJ*Nov. Wie man sieht ist dieser Koeffizient nicht signifikant von Null verschieden, und das Bestimmtheitsmaß ist etwas klein.

Allerdings schätzten **?** nicht dieses Modell, sondern verwendeten anstelle der NJ Dummy firmen-fixe Effekte. Die Verwendung firmen-fixer Effekte ist numerisch äquivalent zur Berücksichtigung von Firmen-Dummies (jede Firma wurde zweimal befragt, und alle bis auf eine Referenz-Firma erhalten eine Dummy). Damit erhielten sie folgendes Ergebnis

$$\begin{array}{rcccc} \text{Emp} & = & -2.283\ \text{Nov} & + & 2.75\ \text{Nov*NJ} \\ & & (1.036)^{**} & & (1.154)^{**} \end{array}$$

$$R^2 = 0.015, \qquad n = 794$$

(firmen-fixe Effekte, Standardfehler in Klammern)

(NJ ist eine Dummy für New Jersey, und Nov für November) Da das Modell mit *fixed effects* geschätzt wurde werden die Koeffizienten der Firmen-Dummies und das Interzept nicht ausgegeben; der R-Code für die Schätzung dieser Gleichung ist

```
rm(list = ls())
d <- read.csv2("https://www.uibk.ac.at/econometrics/data/cardkrueger94.csv",
           dec = ".")
library(plm)
pd <- pdata.frame(d, index = c("Firm", "Period"))
eq1 <- plm(Emp ~ Nov*NJ, model = "within", data = pd)
```

Mit dieser Spezifikation ist der Koeffizient des Treatment Effekts positiv und auf dem 5% Niveau signifikant von Null verschieden, was dahingehend interpretiert wurde, dass die Erhöhung des Mindestlohnes *positive* Beschäftigungseffekte hatte.

Dieses Ergebnis wird bis heute sehr kontrovers diskutiert, siehe z.B. NZZ vom 23. April 2014. Eine ausführlichere Diskussion des 'Difference-in-Differences' Ansatzes sowie dieses Beispiels finden Sie auch bei **?**, 228, für eine aktuellere und frei verfügbare Übersicht über die Effekte von Mindestlöhnen siehe **?**.

Probleme von Difference-in-Differences Modellen und deren Grenzen wurden früh diskutiert (siehe z.B. **?**), aber diese Modelle erfreuen sich nach wie vor großer Beliebtheit. In den vergangenen Jahren wurde der Anwendungsbereich deutlich erweitert, z.B. individuenspezifische Treatments zu unterschiedlichen Zeitpunkten und ähnliches. Für eine aktuelle Übersicht siehe z.B. **?**.

Übung: In den Daten finden Sie auch den Lohn (wage). Überprüfen Sie mit Hilfe einer 'Difference-in-Differences' Analyse, wie sich die Erhöhung des Mindestlohnes von US$4.25 auf US$5.05 auf die durchschnittliche Lohnhöhe auswirkte.

2.7.9 Alternative Kodierungen*

Die in der Ökonometrie gebräuchlichste Form der Modellierung einer kategorialen Variable mit m verschiedenen Ausprägungen ist, $m-1$ Dummy Variablen anzulegen und diese in einer Regressionsgleichung aufzunehmen. Bei dieser *'Dummy Kodierung'* misst das Interzept den Mittelwert der ('weggelassenen') Referenzkategorie, und der Koeffizient einer Dummy Variable j (mit $j = 1, \ldots, m-1$) misst den ceteris paribus Unterschied zwischen den Mittelwerten der Kategorie j und der Referenzkategorie.

Neben dieser einfachen Dummy Kodierung gibt es noch weitere Möglichkeiten zur Modellierung von Dummy Variablen. Eine ähnlich einfache Methode ist die *'Effektkodierung'*. Dabei misst das Interzept den Mittelwert über alle m Kategorien ('grand

mean'), und der Koeffizient einer Dummy Variable den Unterschied zu diesem 'grand mean'. Jede Kategorie j wird also nicht mehr mit der Referenzkategorie verglichen, sondern mit dem Mittelwert über die gesamte Stichprobe.

Wenn die Kategorien unterschiedlich groß sind unterscheidet man weiters zwischen einer ungewichteten und gewichtetet Effektkodierung, je nachdem ob die relativen Häufigkeiten berücksichtigt werden oder nicht.

Dummies für die ungewichtete Effektkodierung erhält man mit

$$D_j^{\text{E-ungew}} = \begin{cases} 1 & \text{für Kategorie } j; \\ -1 & \text{für Referenzkategorie;} \\ 0 & \text{sonst.} \end{cases}$$

Bei der gewichteten Effektkodierung werden die Dummies ähnlich gebildet, nur für die Referenzkategorie werden

$$D_j^{\text{E-gew}} = \begin{cases} 1 & \text{für Kategorie } j; \\ -\frac{n_j}{n_R} & \text{für Referenzkategorie;} \\ 0 & \text{sonst.} \end{cases}$$

wobei n_j die Anzahl der Fälle in Kategorie j und n_R die Anzahl der Fälle in der Referenzkategorie bezeichnet.

Beispiel Werte von y mit Zuordnung zu vier Kategorien:

	Kat.1	Kat.2	Kat.3	Kat.4
	3	10	2	2
	1	6	3	4
	2		3	−3
	2		4	
Mittelwert	2	8	3	1

Gewichteter Mittelwert (*'grand mean'*): 3; Ungewichteter Mittelwert: 3.5

Datentabelle mit Dummies: Referenzkategorie 1; D2 – D4 ... Dummykodierung, DEU2 – DEU4 ... Effektkodierung ungewichtet, DEG2 – DEG4 ... Effektkodierung gewichtet.

y	Kategorie	D2	D3	D4	DEU2	DEU3	DEU4	DEG2	DEG3	DEG4
3	1	0	0	0	−1	−1	−1	−0.5	−1	−0.75
1	1	0	0	0	−1	−1	−1	−0.5	−1	−0.75
2	1	0	0	0	−1	−1	−1	−0.5	−1	−0.75
2	1	0	0	0	−1	−1	−1	−0.5	−1	−0.75
10	2	1	0	0	1	0	0	1	0	0
6	2	1	0	0	1	0	0	1	0	0
2	3	0	1	0	0	1	0	0	1	0
3	3	0	1	0	0	1	0	0	1	0
3	3	0	1	0	0	1	0	0	1	0
4	3	0	1	0	0	1	0	0	1	0
2	4	0	0	1	0	0	1	0	0	1
4	4	0	0	1	0	0	1	0	0	1
−3	4	0	0	1	0	0	1	0	0	1

Dummy Kodierung:

$$\begin{array}{rcccccccc} y & = & 2.00 & + & 6.00\ \text{D2} & + & 1.00\ \text{D3} & - & 1.00\ \text{D4} \\ & & (1.027)^{*} & & (1.78)^{***} & & (1.453) & & (1.569) \end{array}$$

$$R^2 = 0.635, \quad s = 2.055, \quad F\text{-Stat} = 5.211, \quad n = 13$$
(Standardfehler in Klammern)

Effektkodierung, ungewichtet:

$$\begin{array}{rcccccccc} y & = & 3.50 & + & 4.50\ \text{DEU2} & - & 0.50\ \text{DEU3} & - & 2.50\ \text{DEU4} \\ & & (0.593)^{***} & & (1.186)^{***} & & (0.938) & & (1.027)^{**} \end{array}$$

$$R^2 = 0.635, \quad s = 2.055, \quad F\text{-Stat} = 5.211, \quad n = 13$$
(Standardfehler in Klammern)

Effektkodierung, gewichtet:

$$\begin{array}{rcccccccc} y & = & 3.00 & + & 5.00\ \text{DEG2} & + & 0.00\ \text{DEG3} & - & 2.00\ \text{DEG4} \\ & & (0.57)^{***} & & (1.337)^{***} & & (0.855) & & (1.04)^{*} \end{array}$$

$$R^2 = 0.635, \quad s = 2.055, \quad F\text{-Stat} = 5.211, \quad n = 13$$
(Standardfehler in Klammern)

Welche Kodierung sinnvoll ist hängt im wesentlichen davon ab, welcher Vergleich im jeweiligen Zusammenhang sinnvoller ist, rein statistisch sind diese Kodierungen gleichwertig. Wie man auch am Beispiel sieht, unterscheiden sich die R^2 nicht zwischen den verschiedenen Kodierungen.

2.7.10 Stückweise lineare Funktionen*

Stückweise lineare Funktionen (*piecewise linear functions*) sind der einfachste Fall von *Spline Funktionen.*[19]

Die Idee kann am einfachsten anhand eines Beispiels erläutert werden. Angenommen, das Steuersystem eines Landes kennt zwei Schwellenwerte x^{*1} und x^{*2} beim Einkommen, ab denen unterschiedliche marginale Steuersätze angewandt werden. Möchte man die Steuereinnahmen y in Abhängigkeit vom Einkommen x schätzen, so könnte man für jeden der Einkommensbereiche eine eigene Regression schätzen:

$$\widehat{y}|x = \begin{cases} a_1 + a_2 x, & \text{wenn} \quad x < x^{*1}; \\ b_1 + b_2 x, & \text{wenn} \quad x \geq x^{*1} \quad \text{und} \quad x < x^{*2}; \\ c_1 + c_2 x, & \text{wenn} \quad x \geq x^{*2} \end{cases} \tag{2.17}$$

Die Schwellenwerte (*tresholds*) x^{*1} und x^{*2} werden auch Knoten (*knots*) genannt.

[19] Aus Wikipedia: "Ein Spline n-ten Grades ist eine Funktion, die stückweise aus Polynomen mit maximalem Grad n zusammengesetzt ist. Dabei werden an den Stellen, an denen zwei Polynomstücke zusammenstoßen (man spricht auch von Knoten) bestimmte Bedingungen gestellt, etwa dass der Spline (n-1) mal stetig differenzierbar ist."

Anstelle dreier einzelner Gleichungen kann alternativ auch eine Gleichung mit Dummy Variablen und Interaktionstermen geschätzt werden.
Dazu definieren wir zwei Dummy Variablen

$$\begin{aligned} D_1 &= 1 \quad \text{wenn } x \geq x^{*1} \quad \text{und 0 sonst;} \\ D_2 &= 1 \quad \text{wenn } x \geq x^{*2} \quad \text{und 0 sonst;} \end{aligned}$$

Die folgende schätzbare Gleichung mit den zwei Dummyvariablen und Interaktionstermen stellt eine alternative Spezifikation zu den den drei obigen Einzelregressionen dar, aus der exakt die gleichen Koeffizienten berechnet werden können

$$y = a_1 + a_2 x + b_1 D_1 + b_2 D_1 x + c_1 D_2 + c_2 D_2 x + e \tag{2.18}$$

Allerdings stellt dabei nichts sicher, dass sich die einzelnen Regressionsgeraden genau bei den Schwellenwerten schneiden. Die strichlierten Linien in Abbildung 2.29 zeigen ein Beispiel dafür.
Manchmal erwartet man aber aus theoretischen Gründen, dass sich die Regressionsgeraden genau bei den Schwellenwerten schneiden müssen.
Dies kann man einfach erzwingen, denn diese Bedingung kann man als Restriktion auf die Koeffizienten modellieren.
Wenn sich beim ersten Schwellenwert x^{*1} die Regressionsgeraden schneiden sollen müssen die y bei diesem Wert gleich sein. Aus Gleichung (2.18) folgt deshalb für den ersten Schwellenwert

$$a_1 + a_2 x^{*1} = a_1 + a_2 x^{*1} + b_1 + b_2 x^{*1}$$

Daraus folgt die Parameterrestriktion $b_1 = -b_2 x^{*1}$.
Wenn man diese Parameterrestriktion in Gleichung (2.18) einsetzt folgt

$$\begin{aligned} y &= a_1 + a_2 x - b_2 x^{*1} D_1 + b_2 D_1 x + c_1 D_2 + c_2 D_2 x + e \\ &= a_1 + a_2 x + b_2 D_1 (x - x^{*1}) + c_1 D_2 + c_2 D_2 x + e \end{aligned}$$

Da sich die Regressionsgeraden auch beim zweiten Schwellenwert x^{*2} schneiden müssen, muss zudem gelten

$$a_1 + a_2 x^{*2} + b_1 + b_2 x^{*2} = a_1 + a_2 x^{*2} + b_1 + b_2 x^{*2} + c_1 + c_2 x^{*2}$$

Daraus folgt eine weitere Parameterrestriktion $c_1 = -c_2 x^{*2}$.
Wenn man diese und obige Parameterrestriktion in Gleichung (2.18) einsetzt folgt die schätzbare **stückweise lineare Regressionsfunktion**

$$y = a_1 + a_2 x + b_2 D_1 (x - x^{*1}) + c_2 D_2 (x - x^{*2}) + e$$

Die durchgezogene Linie in Abbildung 2.29 zeigt diese Funktion.
Die Gleichungen der drei Geradensegmente sind

$$\widehat{y} = \begin{cases} a_1 + a_2 x, & \text{für } x \leq x^{*1} \\ (a_1 - b_2 x^{*1}) + (a_2 + b_2) x, & \text{für } x^{*1} < x \leq x^{*2} \\ (a_1 - b_2 x^{*1} - c_2 x^{*2}) + (a_2 + b_2 + c_2) x, & \text{für } x > x^{*2} \end{cases}$$

Daraus ist erkennbar, dass die Steigung des ersten Segmentes a_2 ist, die Steigung des zweiten Segmentes ist $a_2 + b_2$ und die Steigung des dritten Segmentes ist $a_2 + b_2 + c_2$.

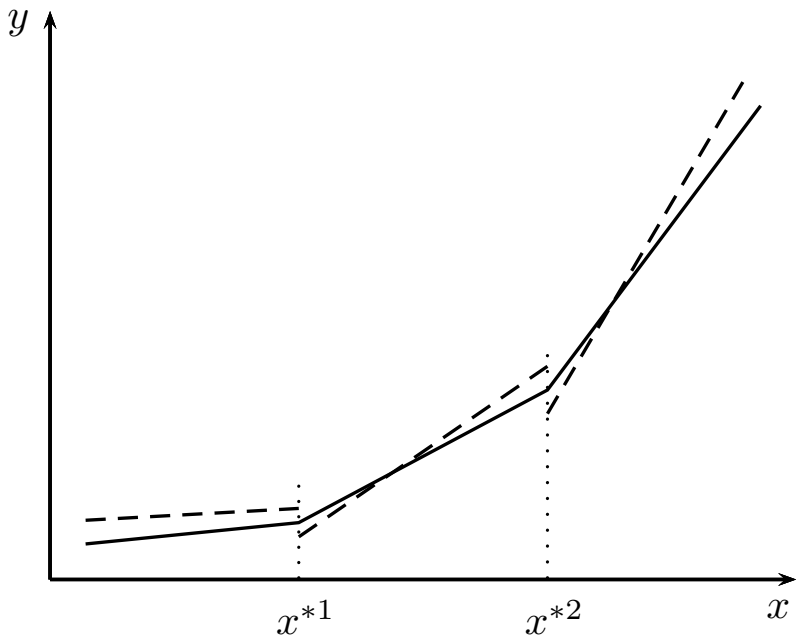


Abbildung 2.29: Einzelregressionen (strichliert) und stückweise lineare Regression (durchgezogene Linie).

2.8 Logarithmische Transformationen

> *"If people do not believe that mathematics is simple, it is only because they do not realize how complicated life is."*
>
> (John von Neumann, 1947)

Wir haben uns bisher fast ausschließlich mit linearen Modellen der Art $\widehat{y} = b_1 + b_2x + b_3x_3$ befasst, in denen die Variablen additiv verknüpft sind.[20] Häufig sind wir aber auch an Modellen interessiert, in denen die Variablen multiplikativ verknüpft sind, wie z.B. bei der bekannten Cobb-Douglas Funktion $Q = AK^{\alpha}L^{\beta}$, dem vermutlich bekanntesten Beispiel einer Exponentialfunktion. Auch für die Modellierung von Wachstumsprozessen spielen solche multiplikative Modelle eine große Rolle. Wir werden gleich sehen, dass solche Modelle mit Hilfe logarithmischer Funktionen einfach linearisiert und geschätzt werden können. Aber nachdem möglicherweise schon der Begriff 'Logarithmus' bei manchen eine leichte Gänsehaut verursacht, und nicht selten mit anderen unliebsamen Erinnerungen an die Schulzeit verdrängt wird, beginnen wir mit einer kurzen Wiederholung.

2.8.1 Wiederholung Exponential- und Logarithmusfunktionen

Als Potenz bezeichnet man einen Term der Art a^x, wobei a Basis und x Exponent (Hochzahl) genannt wird. Für Potenzen gelten die bekannten Rechenregeln, wie z.B. $a^x a^y = a^{x+y}$. Eine *Exponentialfunktion* hat die Form $y = ca^{bx}$ (bzw. in alternativer Schreibweise $x \mapsto ca^{bx}$) mit $x \in \mathbb{R}$, wobei die Basis a sowie c und b fixe Zahlen sind.

Solche Exponentialfunktionen eignen sich u.a. zur Modellierung von Wachstumsprozessen, bei denen sich eine Größe y in gleich langen Zeitintervallen um einen *konstanten Faktor* ändert. Angenommen, der Wert von y sei in der Ausgangsperiode y_0, und dieser Wert nehme in jeder Zeitperiode um 5% zu. Wenn wir die Zeitperiode durch den Subindex ausdrücken folgt $y_0 = y_0(1 + 0.05)^0, y_1 = y_0(1 + 0.05)^1, y_2 = y_1(1+0.05) = y_0(1+0.05)^2, \ldots, y_T = y_0(1+0.05)^T$ mit $t = 0, 1, \ldots, T$. Diese Art des Wachstums kann also durch eine einfache Exponentialfunktion mit Basis $(1 + 0.05)$ und Exponent t beschrieben werden.

Für dieses Beispiel haben wir diskrete Perioden gleicher Länge angenommen. Wenn wir die Periodenlänge gegen Null gehen lassen erhalten wir eine stetige Wachstumsfunktion $y_t = y_0\mathrm{e}^{rt}$, wobei $\mathrm{e} = 2.718281828459\ldots$ die Eulersche Zahl[21] und r die *stetige* Wachstumsrate bezeichnet (siehe Appendix). Aufgrund dieser und einiger weiterer Eigenschaften wird e häufig als *natürliche Basis* bezeichnet und auch als $\exp(\cdot)$ geschrieben (d.h. $\exp(x) := \mathrm{e}^x$). In der Ökonometrie wird fast ausschließlich die natürliche Basis e verwendet, deshalb werden wir uns in den weiteren Ausführungen auf diese Basis beschränken.

[20] Da es in diesem Abschnitt ausschließlich um Funktionsverläufe geht werden wir hier auf den Beobachtungsindex i verzichten.

[21] Beachten Sie den Unterschied zwischen der Eulerschen Zahl e und den Residuen e.

Der *natürliche Logarithmus* ist die Lösung der Exponentialfunktion $y = \mathrm{e}^x$ nach x, d.h. die Logarithmusfunktion ist die Umkehrfunktion zur Exponentialfunktion.

Da die meisten Computerprogramme den log Operator für den natürlichen Logarithmus verwenden folgen wir hier dieser Schreibweise, d.h., log bezeichnet im Folgenden den natürlichen Logarithmus. Deshalb ist $x = \log(y)$ und $\log(\mathrm{e}^x) = x$.

Etwas salopp ausgedrückt sagt uns $\log(y)$, wie oft wir die Basis e mit sich selbst multiplizieren müssen um y zu erhalten. Man beachte, dass die Exponentialfunktion $x \mapsto \mathrm{e}^x$ die Menge der reellen Zahlen $\mathbb{R}$ in die positiven reellen Zahlen $\mathbb{R}^+$ abbildet, da $\mathrm{e}^{-x} = 1/\mathrm{e}^x$. Deshalb ist die Logarithmusfunktion nur für die positiven reellen Zahlen definiert, sie bildet $\mathbb{R}^+ \mapsto \mathbb{R}$ ab; oder einfacher, der Logarithmus negativer Zahlen ist nicht definiert!

Die Bedeutung der logarithmischen Transformation für ökonometrische Anwendungen folgt im wesentlichen aus drei Eigenschaften:

1. *Multiplikative Zusammenhänge können durch Logarithmierung additiv dargestellt werden*, bzw. Exponentialfunktionen werden durch Logarithmierung zu linearen Funktionen

$$\log(xy) = \log(x) + \log(y) \qquad \text{für } x, y > 0$$

   Warum? Um dies zu zeigen definieren wir zwei Zahlen a und b derart, dass $\log(x) = a$ und $\log(y) = b$. Daraus folgt $x = \mathrm{e}^a$ und $y = \mathrm{e}^b$ mit $xy = \mathrm{e}^a\mathrm{e}^b = \mathrm{e}^{a+b}$ aufgrund der Rechenregeln für Potenzen. Deshalb ist $\log(xy) = \log(\mathrm{e}^{a+b}) = a + b := \log(x) + \log(y)$.

   Die wichtigsten Rechenregeln sind

$$\begin{aligned}
\log(xy) &= \log(x) + \log(y) \qquad \text{für } x, y > 0 \\
\log\left(\frac{x}{y}\right) &= \log(x) - \log(y) \\
\log(x^a) &= a\log(x) \\
\log(1/x) &= -\log(x)
\end{aligned}$$

   Deshalb kann z.B. die Cobb-Douglas Funktion $Q = AK^\alpha L^\beta$ linearisiert werden zu $\log(Q) = \log(A) + \alpha\log(K) + \beta\log(L)$.

2. Die Differenz zwischen zwei logarithmierten Werten entspricht näherungsweise einer *relativen* Änderung der ursprünglichen Werte, d.h.

$$\log(x_2) - \log(x_1) \approx \frac{x_2 - x_1}{x_1} := \frac{\Delta x}{x} \qquad \text{für } x_2, x_1 > 0$$

   Wir werden gleich sehen, dass diese Eigenschaft die Interpretation von Regressionskoeffizienten logarithmierter Variablen sehr vereinfacht.

   Für ein intuitives Verständnis erinnern wir uns an die Ableitungsregel für den natürlichen Logarithmus

$$\frac{d\log(x)}{dx} = \frac{1}{x}$$

Dies können wir uns umgeschrieben denken als

$$d \log x = \frac{dx}{x}$$

und erinnern uns, dass wir $d(\log x)$ als eine infinitesimal kleine Änderung von $\log(x)$ interpretieren können. Ähnlich können wir uns dx als eine infinitesimal kleine Änderung von x vorstellen, weshalb dx/x eine (infinitesimal kleine) *relative* Änderung von x darstellt.

In Analogie dazu würden wir erwarten, dass für diskrete Fälle *näherungsweise* ('$\approx$') gilt

$$\log(x + \Delta x) - \log(x) \approx \frac{\Delta x}{x}$$

Bemerkung: Wenn wir zwei konkrete Punkte x_0 und x_1 vor Augen haben können wir für $(x + \Delta x)$ auch schreiben $[x_0 + (x_1 - x_0)] = x_1$, d.h. für den obigen Ausdruck $\log(x_1) - \log(x_0) \approx (x_1 - x_0)/x_0$. □

Dieser Zusammenhang gilt tatsächlich, wenn $\Delta x/x$ 'relativ' klein ist (siehe 'Exkurs: Logarithmische Differenz und relative Änderungsraten', Seite 97).

Wir halten also fest: die Differenz einer logarithmierten Variable misst näherungsweise die relative Änderung der ursprünglichen Variable.

Wenn man eine *relative Änderung* mit 100 multipliziert erhält man die prozentuelle Änderung.

Man beachte, dass dies eine Eigenschaft der logarithmischen Funktionsform ist, und aus dieser Eigenschaft folgt, wie wir gleich noch ausführlicher erläutern werden, dass nicht wie bei der linearen Funktionsform $\widehat{y} = b_1 + b_2 x$ die Steigung $d\widehat{y}/dx$ über den gesamten Funktionsverlauf konstant ist, sondern dass bei einer log-log Funktionsform $\widehat{\log(y)} = b_1 + b_2 \log(x)$ das Verhältnis der *relativen Änderungen* $(d\widehat{y}/\widehat{y})/(dx/x) = b_2$ über den gesamten Funktionsverlauf konstant ist!

3. Durch Logarithmierung werden kleine Zahlenwerte gespreizt, große Zahlenwerte gestaucht, siehe Abbildung 2.30. Dies führt manchmal dazu, dass der Einfluss extremer Beobachtungen auf die Schätzung reduziert wird, oder dass schiefe Verteilungen symmetrischer werden. Ein klassisches Beispiel dazu sind Einkommensdaten, vgl. Abbildung 2.31.

2.8.2 Interpretation der Koeffizienten logarithmierter Variablen

Für OLS Schätzungen selbst spielt es keine unmittelbare Rolle, ob die Variablen logarithmiert wurden oder nicht. Allerdings ändert sich dadurch die Interpretation der Koeffizienten, je nach dem, ob nur die abhängige Variable, nur die erklärende Variable oder beide logarithmiert wurden.

$$
\begin{aligned}
\log(0) &= -\infty \\
\log(0.000001) &= -13.816 \\
\log(0.01) &= -4.605 \\
\log(0.1) &= -2.303 \\
\log(1) &= 0 \\
\log(10) &= 2.303 \\
\log(100) &= 4.605 \\
\log(1000) &= 6.908 \\
\log(1000000) &= 13.816
\end{aligned}
$$

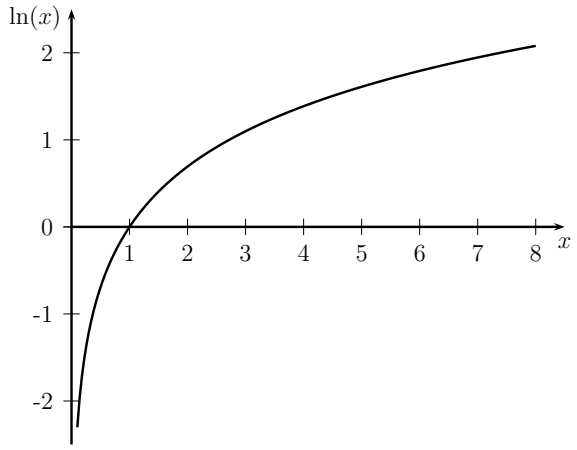


Abbildung 2.30: Durch Logarithmierung werden kleine Zahlenwerte gespreizt, große Zahlenwerte gestaucht.

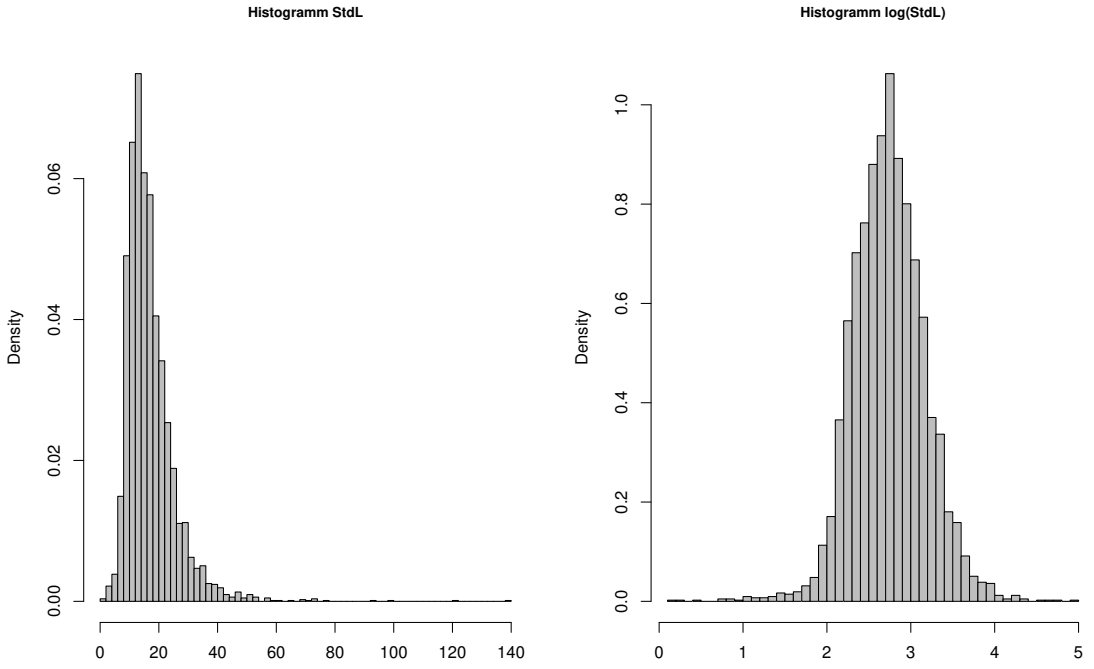


Abbildung 2.31: Histogramme des Brutto-Stundenlohns (StdL) und des Logarithmus vom Brutto-Stundenlohn (Stundenlöhne < 1 und > 200 wurden entfernt); log(Stundenlöhne) sind symmetrischer verteilt! Quelle: EU-Silc 2018

Exkurs: Logarithmische Differenz und relative Änderungsraten

Wir haben im Text behauptet, dass die logarithmische Differenz einer Variable ungefähr gleich der relativen Änderung dieser Variable ist

$$\log(x + \Delta x) - \log(x) \approx \frac{\Delta x}{x} \qquad \text{für kleine } \frac{\Delta x}{x}$$

Dies kann allgemein mit Hilfe einer Taylor Expansion gezeigt werden. Mit einer Taylor Reihenentwicklung können nichtlineare (differenzierbare ...) Funktionen in der Umgebung bestimmter Punkte durch Potenzreihen dargestellt werden. Insbesondere gilt

$$\log(1 + x) = \sum_{n=1}^{\infty} (-1)^{(n+1)} \frac{x^n}{n} = x - \frac{x^2}{2} + \frac{x^3}{3} - \ldots \qquad \text{für } |x| < 1$$

bzw. für $x = \frac{\Delta y}{y}$

$$\log\left(1 + \frac{\Delta y}{y}\right) = \frac{\Delta y}{y} - \frac{1}{2}\left(\frac{\Delta y}{y}\right)^2 + \frac{1}{3}\left(\frac{\Delta y}{y}\right)^3 - \ldots$$

Die linke Seite kann auch als logarithmische Differenz geschrieben werden, da

$$\log\left(1 + \frac{\Delta y}{y}\right) = \log\left(\frac{y + \Delta y}{y}\right) = \log(y + \Delta y) - \log(y)$$

Daraus folgt

$$\log(y + \Delta y) - \log(y) \approx \frac{\Delta y}{y}$$

da für kleine $\Delta y/y$ die Folgeterme $-1/2\,(\Delta y/y)^2 + 1/3\,(\Delta y/y)^3 - \ldots$ der Reihe sehr klein werden und für praktische Zwecke oft vernachlässigbar sind.

Für das Verständnis des Folgenden sind nur zwei Fakten wichtig, *erstens*, dass wie vorhin betont eine logarithmische Differenz ungefähr gleich einer relativen Änderung ist, d.h. $d\,\log(y) = \frac{dy}{y}$ oder für diskrete Änderungen

$$\Delta \log(y) := \log(y + \Delta y) - \log(y) \approx \frac{\Delta y}{y}$$

und *zweitens*, dass der marginale Effekt meist als ein einfacher Differenzenquotient geschrieben werden kann, d.h. für $\widehat{y} = b_1 + b_2 x$ ist der marginale Effekt

$$b_2 = \frac{d\widehat{y}}{dx} = \frac{\Delta\widehat{y}}{\Delta x}$$

wobei das zweite '=' Zeichen nur für *lineare* Funktionsformen exakt gilt, aber wir würden hoffen, dass dies auch für nichtlineare Funktionsformen zumindest für kleine Änderungen näherungsweise gelten sollte.

Wie vorhin schon erwähnt wird die nichtlineare Exponentialfunktion $\widehat{y} = b_0 x_2^{b_2}$ durch logarithmieren linear in den Parametern, d.h.

$$\widehat{\log(y)} = b_1 + b_2 \log(x_2) \quad (\text{mit } b_1 = \log(b_0))$$

Der marginale Effekt von $\log(x_2)$ ist wie üblich der Differenzenquotient

$$b_2 = \frac{d\,\widehat{\log(y)}}{d\,\log(x)} = \frac{\frac{d\widehat{y}}{\widehat{y}}}{\frac{dx}{x}} \qquad \text{bzw. diskret} \qquad b_2 = \frac{\Delta\widehat{\log(y)}}{\Delta \log(x)} \approx \frac{\frac{\Delta\widehat{y}}{\widehat{y}}}{\frac{\Delta x}{x}}$$

(beachte das '≈' Zeichen in der diskreten Darstellung.)

Wenn – wie in diesem Fall – die abhängige *und* die erklärende Variable logarithmiert wird spricht man von einem log-linearen oder besser log-log Modell. Wenn hingegen nur die abhängige *oder* nur die erklärende Variable logarithmiert wird spricht man von einem semi-log Modell, oder manchmal bei $\widehat{\log(y)} = b_1 + b_2 x$ von einem *log-level*, bzw. bei $\widehat{y} = b_1 + b_2 \log(x)$ von einem *level-log* Modell.

2.8.3 Log-log (bzw. log-lineare) Modelle

Wie schon erwähnt kann die Exponentialfunktion

$$y_i = b x_i^{b_2} \exp(e_i)$$

durch Logarithmierung linearisiert werden (vgl. Abbildung 2.32)

$$\log y_i = b_1 + b_2 \log x_i + e_i \qquad \text{mit } b_1 := \log b$$

Dieses Modell ist *linear in den Parametern* und kann deshalb ganz normal mit OLS geschätzt werden. Der *marginale Effekt*

$$b_2 = \frac{d\log(y)}{d\log(x)}$$

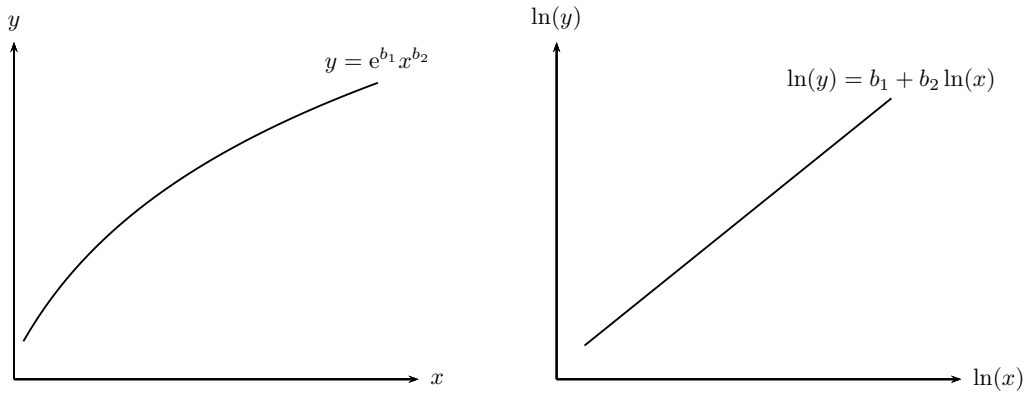


Abbildung 2.32: Log-log Modell mit logarithmisch und linear skalierter Skala

kann zwar wie üblich interpretiert werden, nämlich um wie viele Einheiten sich $\log(y)$ ändert, wenn $\log(x)$ um *eine Einheit* zunimmt, aber wie würden Sie dies einem Laien erklären? Was soll man sich unter einer Einheit von $\log(x)$ vorstellen?

Hier kommt uns die oben erwähnte Eigenschaft logarithmischer Funktionen zu Hilfe, dass die absolute Differenz zwischen zwei logarithmierten Werten näherungsweise der *relativen* Differenz der ursprünglichen Werte entspricht (d.h. $\Delta \log y \approx \Delta y/y$), also[22]

$$b_2 = \frac{d\log(y)}{d\log(x)} = \frac{\frac{dy}{y}}{\frac{dx}{x}} \approx \frac{\frac{\Delta y}{y} \times 100}{\frac{\Delta x}{x} \times 100} = \frac{\text{prozentuelle Änderung von } y}{\text{prozentuelle Änderung von } x} = \text{Elastizität}_{y,x}$$

wobei eine Elastizität als das *Verhältnis zweier relativer (bzw. prozentueller) Änderungen* definiert ist.

Deshalb können wir die Koeffizienten von log-log Modellen unmittelbar als Elastizitäten interpretieren.

> ⇒ Der Koeffizient in einem log-log Modell gibt an, um wie viele *Prozent* sich die abhängige Variable y (ceteris paribus) ändert, wenn die erklärende Variable x *um ein Prozent zunimmt*.

wobei die ceteris paribus Klausel natürlich nur für multiple Regressionen in Bezug auf die anderen berücksichtigten Regressoren gilt.

[22] Erinnern wir uns, dass $de^x/dx = e^x$ und $d\log(x)/dx = 1/x$.

Wir können nun entweder $y = \exp(b_1 + b_2 \log(x))$ nach x differenzieren, indem wir die Kettenregel anwenden

$$\frac{dy}{dx} = b_2 \frac{1}{x} \underbrace{[\exp(b_1 + b_2 \log(x))]}_{y} \quad \Rightarrow \quad b_2 = \frac{dy}{dx}\frac{x}{y} = \frac{\frac{dy}{y}}{\frac{dx}{x}}$$

oder wir können alternativ $\log(y) = b_1 + b_2 \log(x)$ total differenzieren

$$\frac{1}{y}\, dy = b_2 \frac{1}{x}\, dx \quad \Rightarrow \quad b_2 = \frac{dy}{dx}\frac{x}{y} = \frac{\frac{dy}{y}}{\frac{dx}{x}}$$

Für infinitesimale Änderungen gilt dieser Zusammenhang also exakt.

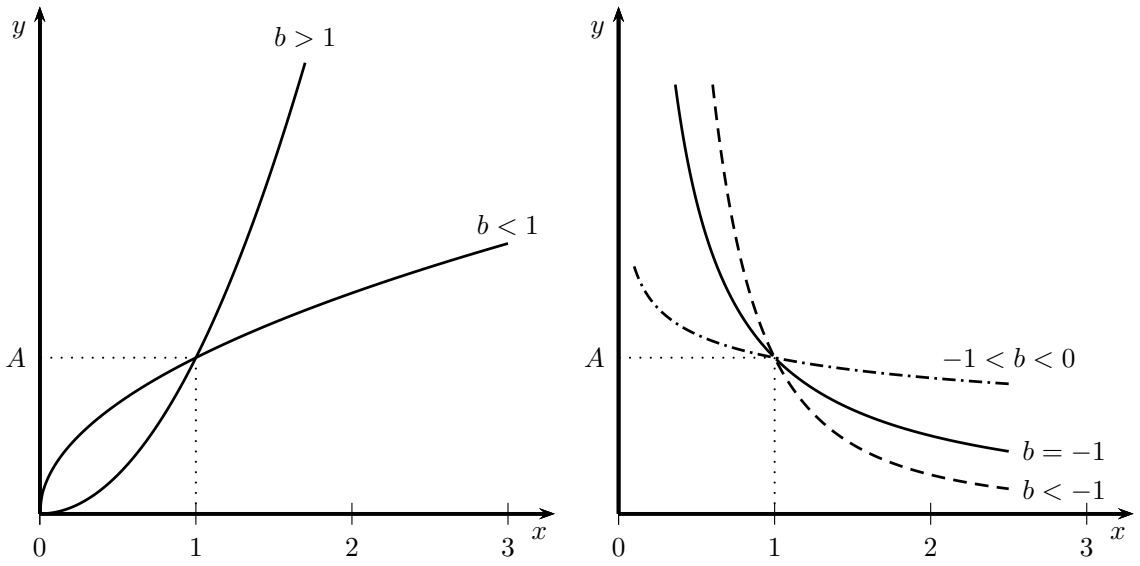


Abbildung 2.33: Log-log (bzw. log-lineare) Modelle: Verläufe der Exponentialfunktion $y = Ax^b$ für unterschiedliche b.

Beispiel: Angenommen wir haben die Funktion

$$\log(y) = 1 + 0.2\log(x)$$

Wie ändert sich y, wenn x um *ein Prozent* zunimmt? Aufgrund der obigen Diskussion erwarten wir, dass y um 0.2 *Prozent* zunehmen wird. Stimmt das wirklich? Wir können dies einfach überprüfen, indem wir obige Funktion umschreiben zu $y = \exp[1 + 0.2\log(x)]$ und für x zwei Werte einsetzen, z.B. 5 und 5.05.

x	$(\Delta x)/x$	$\%\Delta x$	$y = \exp(1 + 0.2\log(x))$	$(\Delta y)/y$	$\%\Delta y$
5.00			3.750494		
5.05	$\frac{5.05-5}{5} = 0.01$	1%	3.757965	0.001992	$\approx 0.2\%$

Wir sehen, dass dies nicht exakt gilt, da eine Änderung um ein Prozent keine infinitesimal kleine Änderung ist, aber für praktische Zwecke ist diese Näherung in den allermeisten Fällen mehr als ausreichend. □

Abbildung 2.33 zeigt mögliche Verläufe für positive (links) und negative b (rechts) der Exponentialfunktion $y = Ax^b$, die durch Logarithmierung linear wird $\log(y) = \log(A) + b\log(x)$.

Beispiel: **?** schätzten für die USA von 1899–1922 folgende Produktionsfunktion

$$\begin{array}{rcccc} \log(Q_i) & = & -0.177 & + & 0.807\log(L_i) \\ & & (0.434) & & (0.145) \end{array} \quad + \quad \begin{array}{c} 0.233\log(K_i) + e_i \\ (0.064) \end{array}$$

$$R^2 = 0.957, \quad n = 24$$

(Standardfehler in Klammern)

Beide Steigungskoeffizienten weisen das erwartete Vorzeichen auf und sind signifikant von Null verschieden. Der Koeffizient von $\log(L_i)$ gibt die Ausbringungselastizität des Faktors Arbeit an, d.h., falls der Arbeitseinsatz um 1% erhöht wird erwarten wir ceteris paribus eine Erhöhung des Outputs um 0.807%. Analog, wenn der Kapitaleinsatz um 1% steigt erwarten wir ceteris paribus eine Zunahme des Outputs um 0.233%.

Achtung: Die Beziehung $y = \alpha x^{\beta}$ kann ökonometrisch auf verschiedene Arten modelliert werden, d.h. die Störterme können unterschiedlich in das Modell eingehen

$$\begin{array}{llllll} 1) & y_i = \alpha x_i^{\beta} \exp(e_i) & \Rightarrow & \log y_i = \log\alpha + \beta \log x_i + e_i \\ 2) & y_i = \alpha x_i^{\beta} e_i & \Rightarrow & \log y_i = \log\alpha + \beta \log x_i + \log e_i \\ 3) & y_i = \alpha x_i^{\beta} + e_i & \Rightarrow & \log y_i = \log(\alpha x_i^{\beta} + e_i) \end{array}$$

Nur die erste Gleichung kann unmittelbar mittels OLS geschätzt werden!

Die zweite Gleichung kann zwar geschätzt werden, aber dies hat meist unerwünschte Implikationen für die Residuen e_i (nicht zu verwechseln mit der Eulerschen Konstanten e), denn wenn $\log(e_i)$ normalverteilt ist, d.h. $\log(e_i) \sim \mathrm{N}(0, \sigma^2)$, dann sind die Residuen e_i log-normalverteilt mit einem positiven Erwartungswert $\exp(\sigma^2/2)$ (siehe Exkurs Seite 101)! Dies hat insbesondere auch Implikationen für Prognosen, wenn die gefitteten Werte $\widehat{\log(y)}$ geschätzt wurden, wir aber an $\widehat{y}$ interessiert sind, siehe z.B. **?**, 204ff.

Die dritte Gleichung $\log y_i = \log(\alpha x_i^{\beta} + e_i)$ ist schließlich nicht linear in den Parametern, da $\log(A + B) \neq \log A + \log B$, und kann deshalb nicht einfach mit OLS geschätzt werden!

2.8.4 Log-level (bzw. log-lin) Modelle

Beim log-level Modell wird nur die abhängige Variable y logarithmiert, aber nicht die erklärende x Variable

$$\widehat{\log(y)} = b_1 + b_2 x$$

Mögliche Funktionsverläufe für einen positiven und negativen Steigungskoeffizienten sind in Abbildung 2.34 dargestellt.

Der marginale Effekt von x auf $\widehat{\log(y)}$ ist wie üblich die Ableitung $\frac{d\widehat{\log(y)}}{dx} = b_2$, aber dies ist etwas schwierig zu interpretieren, da sich unter einer Änderung von $\widehat{\log(y)}$ vermutlich nicht viele etwas vorstellen können.

Wir können natürlich den übliche marginalen Effekt dy/dx von $d\log(y) = b_1 + b_2 x$ berechnen

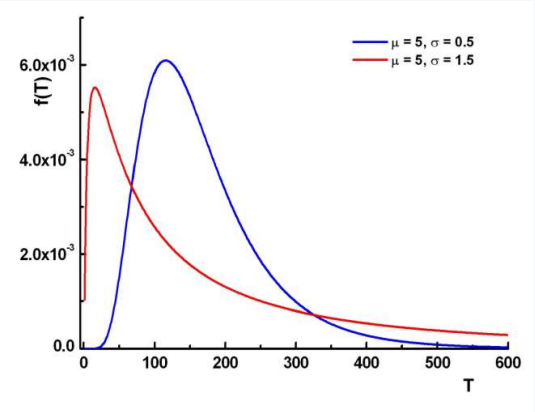
$$\frac{d\log(y)}{dx} = \frac{1}{y}\frac{dy}{dx} = b_2 \quad \Rightarrow \quad \frac{dy}{dx} = b_2 y = b_2(b_1 + b_2 x)$$

aber offensichtlich ist die Größe dieses üblichen marginalen Effekts nicht konstant, sondern hängt von der Ausprägung von x ab! Deshalb ist die Angabe dieses marginalen Effekts bei log-level Modellen ziemlich unüblich.

Exkurs: Die Log-Normalverteilung (Logarithmische Normalverteilung)

Eine Zufallsvariable, deren natürlicher Logarithmus normalverteilt ist, ist log-normalverteilt.

Das heißt, wenn $\log(X) \sim \mathrm{N}(\mu, \sigma^2)$), dann ist X log-normalverteilt.
Anders herum, wenn eine Zufallsvariable Y normalverteilt ist, dann ist $X = \exp(Y)$ log-normalverteilt.
Die Log-Normalverteilung ist rechtsschief und kann nur positive Werte annehmen.



Eine log-normale Verteilung wird häufig zur Modellierung von Zufallsvariablen herangezogen, die man sich als das (multiplikative) Produkt vieler kleiner unabhängiger Faktoren vorstellen kann.
Erwartungswert und Varianz sind

$$\begin{aligned} \mathrm{E}(X) &= \exp\left(\mu + \frac{\sigma^2}{2}\right) \\ \mathrm{var}(X) &= \exp(2\mu + 2\sigma^2)[1 - \exp(-\sigma^2)] \\ \mathrm{Median}(X) &= \exp(\mu) \end{aligned}$$

Für $\mu = 0$ ist der Mittelwert $\exp(\sigma^2/2)$ und die Varianz $\exp(\sigma^2)(\exp(\sigma^2 - 1))$.

□

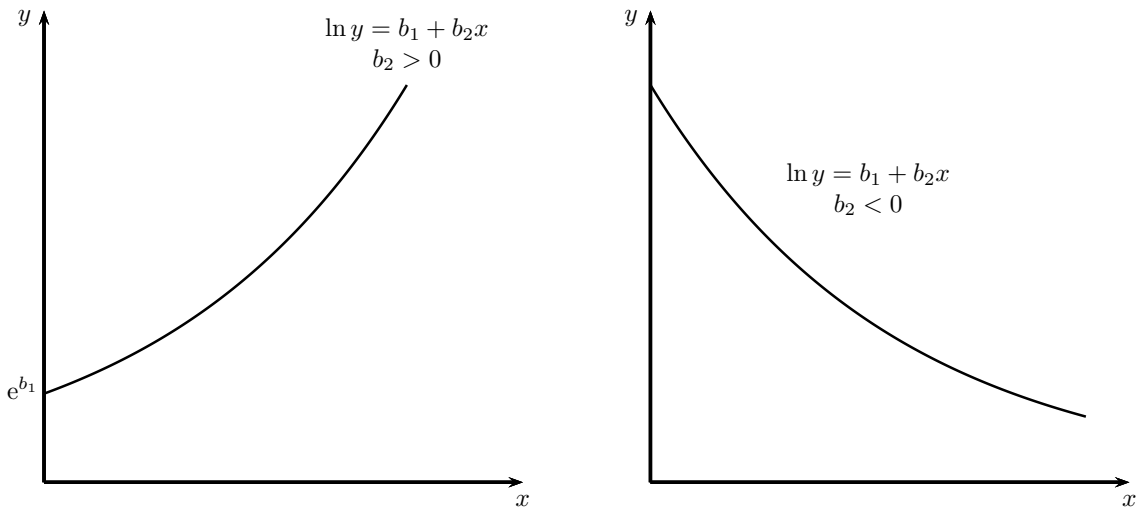


Abbildung 2.34: Log-level Modelle: $\log y = \alpha + \beta x$

Aber wenn wir uns wieder erinnern, dass für infinitesimal kleine Änderungen gilt $d\log(y) = dy/y$, können wir schreiben[23]

$$b_2 = \frac{d\log(y)}{dx} = \frac{\frac{dy}{y}}{dx} \approx \frac{\frac{\Delta y}{y}}{\Delta x}$$

wobei – wie wir gleich noch zeigen werden – die Approximation nur für *'kleine'* Δx und *kleine* b_2 hinreichend genau ist.

Aber auch dies ist schwierig zu kommunizieren, wenn x um eine kleine Einheit zunimmt, ändert sich $\Delta(y)/y$ um b_2? Viel einfacher geht es, wenn wir die linke *und* rechte Seite mit 100 multiplizieren, dann erhalten wir eine *prozentuelle* Änderung von y.

$$100 \times b_2 \approx \frac{\frac{\Delta y}{y} \times 100}{\Delta x}$$

oder in Worten:

> $\Rightarrow$ Wenn $\log(y) = b_1 + b_2 x$ und wenn x um *eine kleine Einheit* zunimmt, ändert sich y (ceteris paribus) *ungefähr* um $100 \times b_2$ *Prozent*.

Für die log-level Funktionsformen gilt diese Interpretation für *alle* x, d.h. der so ausgedrückte marginale Effekt hängt nicht von der Ausprägung von x ab! Dies erleichtert die Interpretation ganz erheblich.

Dieser Ausdruck wird manchmal auch eine *Semi-Elastizität* genannt, da der Zusammenhang zwischen einer *prozentuellen* Änderung von y und einer *absoluten* Änderung von x beschrieben wird.

Allerdings gilt dies für diskrete Änderungen nur *ungefähr*. Wie Abbildung 2.35 zeigt führt eine Zunahme von x um *eine Einheit* (d.h. $\Delta x = 1$) zu einer Änderung von y um Δy Einheiten, die Steigung der Tangente dy/dx im Ausgangspunkt beschreibt dies nur *näherungsweise*.

Für stark gekrümmte Kurven (d.h. große b_2) und große Δx ist diese Approximation weniger genau.

Aber diese Ungenauigkeit kann leicht korrigiert werden, um die Auswirkungen solcher diskreter Änderungen von x auf y zu bestimmen bilden wir Differenzen. Für $\log(y) = b_1 + b_2 x$ erhalten wir durch einfache Umformungen

$$\begin{aligned}
\Delta\log(y) := \log(y + \Delta y) - \log(y) &= b_1 + b_2(x + \Delta x) - (b_1 + b_2 x) \\
\log\left(\frac{y + \Delta y}{y}\right) &= b_2 \Delta x \\
1 + \frac{\Delta y}{y} &= \exp(b_2 \Delta x) \\
\frac{\Delta y}{y} &= \exp(b_2 \Delta x) - 1 \\
\left(\frac{\Delta y}{y}\right) \times 100 &= [\exp(b_2 \Delta x) - 1] \times 100
\end{aligned}$$

[23] Indem wir z.B. $\log(y) = b_1 + b_2 x$ total differenzieren $\frac{1}{y} dy = b_2 dx$ und nach b_2 lösen, oder alternativ indem wir $y = \exp(b_1 + b_2 x)$ nach x differenzieren und umschreiben.

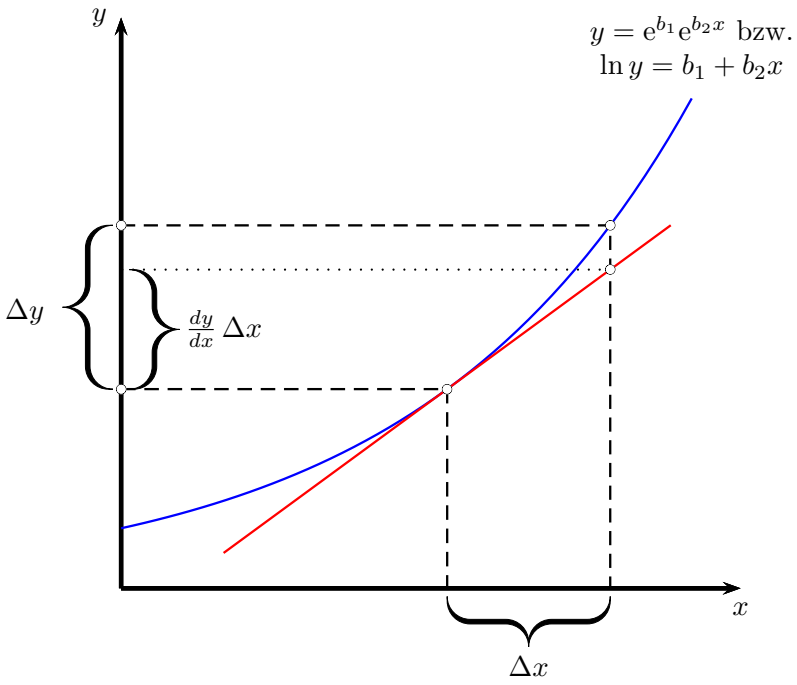


Abbildung 2.35: Auswirkung einer diskreten Änderung von x (d.h. Δx) auf y.

bzw.

$$\%\Delta y := \left(\frac{\Delta y}{y}\right) \times 100 = [\exp(b_2) - 1] \times 100 \qquad \text{für } \Delta x = 1$$

d.h. wenn x um *eine Einheit* zunimmt ändert sich y (c.p.) um $[\exp(b_2) - 1] \times 100$ *Prozent*.

Für kleine b_2 (z.B. $b_2 < 0.1$) gilt $[\exp(b_2) - 1] \approx b_2$ (vgl. Abbildung 2.36), deshalb wird auf diese Korrektur für kleine b_2, z.B. für $b_2 < 0.1$, häufig verzichtet.

b_2	$\exp(b_2) - 1$
−0.50	−0.393
−0.30	−0.259
−0.10	−0.095
−0.01	−0.010
0.01	0.010
0.05	0.051
0.10	0.105
0.30	0.350
0.50	0.649

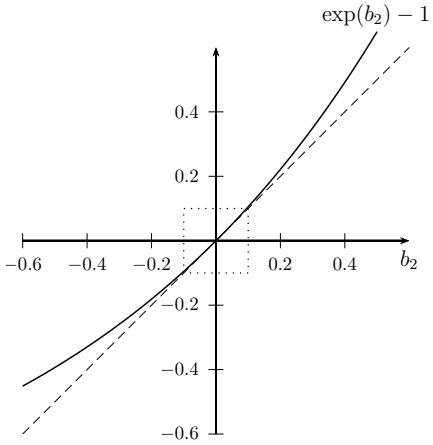


Abbildung 2.36: Für kleine Werte von b_2 (z.B. $|b_2| \leq 0.1$) ist der Unterschied zwischen b_2 und $\exp(b_2) - 1$ oft vernachlässigbar.

Log-level Modelle mit Dummy Variablen Bei der Interpretation der Koeffizienten von Dummy Variablen in log-level Modellen ist zu beachten, dass sich Dummy Variablen per Definition nicht infinitesimal ändern können, deshalb muss wieder die Differenz zwischen beiden Ausprägungen der Dummy Variable gebildet werden. Sei d eine Dummy Variable und $\log(y_i) = b_1 + b_2 d_i$, dann sind die Differenzen

$$\begin{aligned}
\Delta \log(y_i) := [\log(y_i)|d_i = 1] - [\log(y_i)|d_i = 0] &= b_1 + b_2 \times 1 - b_1 - b_2 \times 0 \\
\log\left(\frac{y_i|d_i = 1}{y_i|d_i = 0}\right) &= b_2 \\
\frac{y_i|d_i = 1}{y_i|d_i = 0} - 1 &= \exp(b_2) - 1 \\
\left[\frac{(y_i|d_i = 1) - (y_i|d_i = 0)}{(y_i|d_i = 0)}\right] \times 100 &= [\exp(b_2) - 1] \times 100
\end{aligned}$$

d.h. um den prozentuellen Unterschied in $\widehat{y}$ zwischen den beiden durch die Dummy-Variable definierten Kategorien zu ermitteln müssen wir wieder die gleiche Korrektur anwenden (wenn b_2 'groß' ist):

> $\Rightarrow$ Der durchschnittliche *prozentuelle* Unterschied zwischen den zwei durch eine Dummy Variable definierten Kategorien ist
>
> $$[\exp(b_2) - 1] \times 100$$

Für $|b_2| < 0.1$ gilt wieder $[\exp(b_2) - 1] \approx b_2$.

Beispiel 1: Auf Grundlage von EU-Silc Daten für Österreich (2018) wurde folgende Lohngleichung geschätzt, wobei 'StdL' den Stundenlohn unselbständig Beschäftigter bezeichnet, 'potBildg' ist die potentielle Ausbildungszeit in Jahren (Alter bei Berufseinstieg minus 6), 'Erf' die Berufserfahrung in Jahren und 'weibl' ist eine Dummy Variable.

$$\begin{array}{rccccccc}
\log(\text{StdL}) = & 2.087 & + & 0.038\ \text{potBildg} & + & 0.011\ \text{Erf} & - & 0.139\ \text{weibl} \\
& (0.03)^{***} & & (0.002)^{***} & & (0.001)^{***} & & (0.013)^{***}
\end{array}$$

$$R^2 = 0.161, \qquad n = 4104$$

Nach dieser Schätzung nimmt der Stundenlohn ceteris paribus mit jedem Jahr potentieller Bildung um 3.8% zu (genauer: $(\exp(0.038) - 1)100 = 3.87\%$). Aufgrund der log-level Funktionsform gilt dies für alle Bildungsniveaus.

Frauen verdienen nach dieser Schätzung im Durchschnitt und ceteris paribus um ca. 13 *Prozent* weniger als Männer, weil $(\exp(-0.139) - 1) * 100 = -13\%$.

Beispiel 2: Betrachten wir die Funktion

$$\log(y) = 1 + 0.2x$$

Wie ändert sich y, wenn x um *eine Einheit* von 5 auf 6 zunimmt? Aufgrund der obigen Diskussion erwarten wir, dass y um $[(\exp(0.2) - 1] \times 100\% = 22.14$ *Prozent* zunehmen wird.

Dies können wir wieder einfach numerisch zeigen, indem wir obige Funktion umschreiben zu $y = \exp(1 + 0.2x)$ und für x zwei Werte einsetzen, z.B. 5 und 6.

x	Δx	$y = \exp(1 + 0.2x)$	$(\Delta y)/y$	$\%\Delta y$
5		7.3890561		
6	1	9.0250135	0.2214	22.14%

Da in diesem Fall der Koeffizient $b_2 = 0.2$ relativ groß ist müssen wir die Korrektur $[(\exp(0.2) - 1]100\% = 22.14\%$ durchführen, d.h. die Zunahme von x um *eine Einheit* führt zu einer Zunahme von y um 22.14 *Prozent*.

□

Beispiel 3: Berechnung von durchschnittlichen Wachstumsraten mittels OLS Mit Hilfe einer einfachen log-level Regression auf den Trend[24] kann eine durchschnittliche Wachstumsrate berechnet werden.

Wenn g die diskrete Wachstumsrate einer Variable y ist gilt

$$y_t = y_0(1 + g)^t \quad \Rightarrow \quad \log y_t = \log y_0 + \log(1 + g) \times t$$

Dieser Zusammenhang sollte für jede Periode gelten. Um die diskrete Wachstumsrate g zu schätzen können wir deshalb t durch eine Trendvariable Trend $= 1, 2, 3, \ldots, T$ ersetzen. Wenn wir mit y_0 den Wert von y in der Ausgangsperiode bezeichnen ist

$$\log y_t = \underbrace{\log y_0}_{b_1} + \underbrace{\log(1 + g)}_{b_2} \times \text{Trend}_t$$

Wir können also einfach

$$\log y_t = b_1 + b_2 \text{Trend}_t + e_t$$

schätzen und aus $b_2 = \log(1 + g)$ die durchschnittliche diskrete Wachstumsrate g berechnen, denn aus

$$b_2 = \log(1 + g) \quad \text{folgt} \quad g = \exp(b_2) - 1$$

Die prozentuelle durchschnittliche diskrete Wachstumsrate ist deshalb

$$\boxed{g\% := g \times 100 = [\exp(b_2) - 1] \times 100}$$

Wenn b_2 sehr klein ist (z.B. kleiner als 0.1) wird sich b_2 nur geringfügig von $\exp(b_2) - 1$ unterscheiden, bei größeren Werten sollte die Korrektur aber durchgeführt werden.

[24] Eine Trendvariable nimmt mit jeder Beobachtung um eine Einheit zu, z.B. Trend $= 1, 2, 3, \ldots, T$.

Beispiel: Für das reale BIP pro Kopf Chinas wurde folgende Regression geschätzt (Datenquelle: World dataBank, WDI; Dependent variable: GDP, PPP, constant 2005 international dollar.)

$$\begin{array}{rcccc} \log(\text{GDPpc}) & = & -\ 160.669 & + & 0.084\ \text{Trend} \\ & & (2.71)^{***} & & (0.001)^{***} \end{array}$$

$$R^2 = 0.994, \qquad n = 25 \quad (1995 - 2019)$$
(Standardfehler in Klammern)

Nach dieser Schätzung hat das reale pro Kopf Einkommen Chinas im Zeitraum 1995 – 2019 im Durchschnitt *jährlich* um $(\exp(0.084) - 1) \times 100 \approx 8.81$ *Prozent* zugenommen (das R- und Stata Programm zur Berechnung dieser Wachstumsraten finden Sie im Appendix, Seite 148).

Hinweis: Wir könnten uns fragen, in welcher Zeitspanne sich das Einkommen verdoppelt, wenn es mit einer natürlichen Wachstumsrate r wächst. Dazu müssen wir nur

$$2Y_0 = Y_0 \mathrm{e}^{rt}$$

nach t lösen. Logaritmieren gibt $\log(2) = rt$ oder $t = \log(2)/r \approx 0.7/r$. Wenn wir Zähler und Nenner mit 100 multiplizieren erhalten wir eine prozentuelle Wachstumsrate, also[25]

$$\text{Verdoppelungszeit} \approx \frac{70}{r\%}$$

Mit einer Wachstumsrate von 10% würde sich das Einkommen also ungefähr alle sieben Jahre verdoppeln!

Beispiel 4: Eine spezielle Spezifikation ist

$$\log(y) = b_1 + b_2 \frac{1}{x}$$

Dieses log-inverse Modell erlaubt die Modellierung zuerst zunehmender und dann abnehmender marginaler Effekte, vgl. Abbildung 2.37. Ein solches Modell könnte zum Beispiel verwendet werden, um Verkaufsumsätze in Abhängigkeit von Werbeausgaben zu erklären. Die S-förmige Funktionsform erlaubt zuerst zunehmende Grenzerträge von Werbeausgaben, und ab dem Wendepunkt bei $b_2/2$ abnehmende Grenzerträge, und nähert sich schließlich asymptotisch einem horizontalen Verlauf an.

[25] Natürlich gilt dies näherungsweise auch für eine diskrete Wachstumsrate g. Wir lösen $2Y_0 = Y_0(1+g)^t$ nach der Verdoppelungszeit t und erhalten $t = \log(2)/\log(1+g)$. Für kleine g gilt $\log(1+g) \approx g$, da wie in Exkurs Seite 97 gezeigt

$$\log(x + \Delta x) - \log(x) = \log\left(\frac{x + \Delta x}{x}\right) = \log\left(1 + \frac{\Delta x}{x}\right) \approx \frac{\Delta x}{x}$$

Für $g := \Delta x/x$ folgt daher $\log(1+g) \approx g$, d.h. dieser Zusammenhang gilt *näherungsweise* auch für diskrete Wachstumsraten.

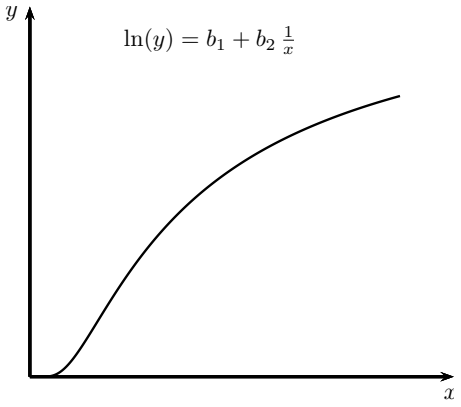


Abbildung 2.37: Log-Reziproke Transformationen

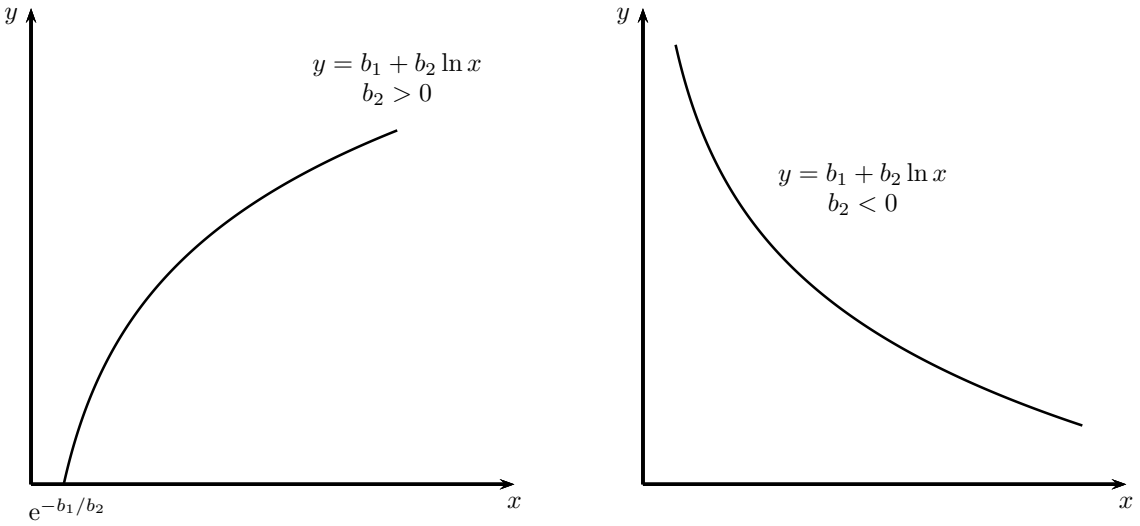


Abbildung 2.38: Lin-log Modell: $y = b_1 + b_2 \log x$

2.8.5 Level-log (bzw. lin-log) Modelle

Beim level-log Modell wird nicht die abhängige, sondern die erklärende Variable logarithmiert. Eine grafische Abbildung des level-log Modells

$$y_i = b_1 + b_2 \log x_i$$

findet sich in Abbildung 2.38.

Wir können dies wieder in Änderungen anschreiben[26]

$$\Delta y = b_2 \Delta \log(x) \quad \Rightarrow \quad b_2 = \frac{\Delta y}{\Delta \log(x)} \approx \frac{\Delta y}{\frac{\Delta x}{x}} = \frac{\text{absolute Änderung von } y}{\textit{relative} \text{ Änderung von } x}$$

[26]Für infinitesimale Änderungen differenzieren wir $y = b_1 + b_2 \log(x)$ total und erhalten $dy = b_2 \frac{1}{x} dx$ woraus folgt

$$b_2 = \frac{dy}{dx/x}$$

Um eine prozentuelle Änderung von x zu erhalten müssen wir in diesem Fall die linke und die rechte Seite durch 100 *dividieren*

$$\frac{b_2}{100} \approx \frac{\Delta y}{\frac{\Delta x}{x} \times 100} = \frac{\text{absolute Änderung von } y}{\textit{prozentuelle} \text{ Änderung von } x}$$

d.h. eine Zunahme von x um *ein Prozent* führt ceteris paribus zu einer absoluten Änderung von y um $0.01 \times b_2$ *Einheiten*.

Beispiel 1: Angenommen wir haben die Funktion

$$y = 1 + 0.2 \log(x)$$

Wie ändert sich y, wenn x um *ein Prozent* zunimmt? Aufgrund der obigen Diskussion erwarten wir, dass y um 0.002 *Einheiten* zunehmen wird.

Wir überprüfen dies wieder, indem wir in obige Funktion zwei Werte für x einsetzen, z.B. 5 und 5.05.

x	$(\Delta x)/x$	$\%\Delta x$	$y = 1 + 0.2 \log(x)$	Δy
5.00			1.3218876	
5.05	$\frac{5.05-5}{5} = 0.01$	1%	1.3238777	0.00199

Da eine Änderung um ein Prozent keine infinitesimal kleine Änderung ist gilt dies nicht exakt, aber für praktische Zwecke ist diese Näherung meist ausreichend. □

Beispiel 2: Abbildung 2.39 zeigt den Zusammenhang zwischen Lebenserwartung (LE; Life expectancy at birth, total (years)) und pro Kopf Einkommen (GNIpC; GNI per capita, PPP (current international $)) für einen Länderquerschnitt und das Jahr 2018. Das linke Panel zeigt die level-level Abbildung, offensichtlich beschreibt eine lineare Regression diesen Zusammenhang deutlich schlechter als eine level-log Regression. Das rechte Panel zeigt den Zusammenhang, wenn auf der x-Achse das logarithmierte pro Kopf Einkommen aufgetragen wird.

Tabelle 2.15 zeigt die dazugehörigen Regressionen.

Interpretation: Wenn das pro Kopf Einkommen um *ein Prozent* zunimmt erwarten wir auf Grundlage dieser Schätzung (ceteris paribus) eine Verlängerung der Lebenserwartung um ca. 0.057 *Jahre*, das sind ca. 21 Tage.

Achtung: wir dürfen dieses Ergebnis nicht kausal interpretieren, da sowohl die Lebenserwartung als auch das pro Kopf Einkommen von vielen weiteren Variablen abhängen, und damit ziemlich sicher ein *'omitted variable bias'* vorliegt!!!

2.8.6 Wann logarithmieren?

Logarithmische Transformationen werden in der Ökonometrie häufig angewandt, vor allem für Größen, die in Geldbeträgen oder einem anderen Niveau gemessen werden, z.B. BIP, Bevölkerung oder Fläche. Insbesondere makroökonomische Variablen,

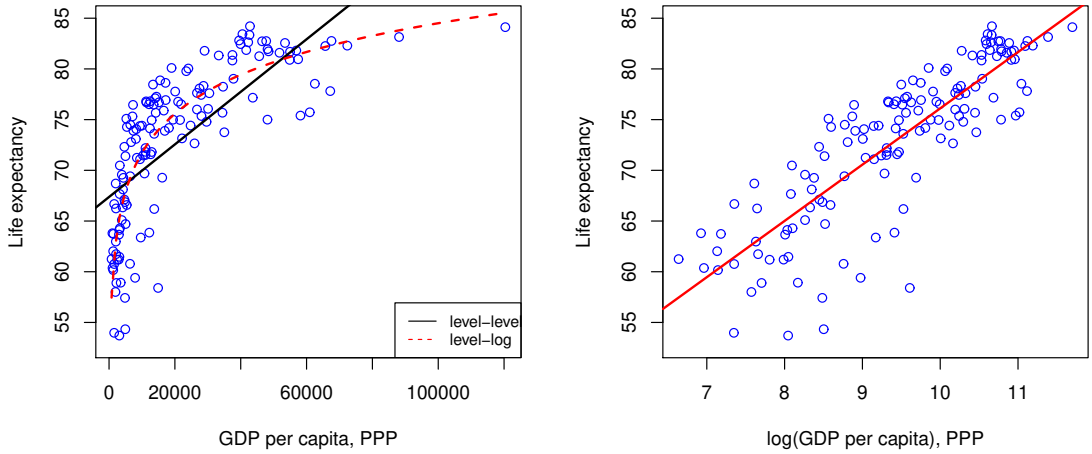


Abbildung 2.39: Lebenserwartung (Life expectancy at birth, total (years)) vs. pro Kopf Einkommen (GNI per capita, PPP, constant 2011 international $), 2018; Datenquelle: World Bank, WDI, `https://databank.worldbank.org/`; (Den R Code zur Erzeugung dieser Abbildung finden Sie im Appendix, Seite 149 erzeugt.)

von denen man plausibel annehmen kann, dass sie langfristig ungefähr exponentiell wachsen, werden häufig logarithmiert, denn wie wir gesehen haben nimmt der Logarithmus solcher Variablen linear zu.

Als einfache Richtgröße können Sie überlegen, ob Sie bei Änderungen einer Variable eher an *prozentuelle* oder an Änderungen um einen *Absolutbetrag* denken.

Variablen, die eine Zeitdimension haben, werden hingegen eher selten logarithmiert, z.B. Alter, Berufserfahrung oder Bildungsjahre.

Darüber hinaus bietet die Logarithmierung manchmal Vorteile, wenn die Verteilung von y schief ist oder die Varianzen von e_i nicht konstant sind (Heteroskedastizität), denn häufig erfüllt die Verteilung einer logarithmierten Variable die Annahmen des Regressionsmodells besser die Verteilung einer nicht logarithmierten Variable. Da durch die Logarithmierung große Zahlenwerte 'gestaucht' werden, sind logarithmisch spezifizierte Regressionen häufig auch weniger anfällig gegenüber Ausreißern (*'outliers'*).

Ein weiterer Grund besteht darin, dass die Standardabweichung vieler ökonomischer Variablen ungefähr proportional zum Niveau dieser Variablen ist, deshalb ist die Standardabweichung der logarithmierten Variablen näherungsweise konstant. In anderen Worten, das Logarithmieren ökonomischer Variablen führt häufig zu einer Art 'Stabilisierung' der Standardfehler der Koeffizienten.

Ein großer Vorteil logarithmischer Transformationen besteht darin, dass die Steigungskoeffizienten als Elastizitäten, bzw. Semi-Elastizitäten, interpretiert werden können, und deshalb unabhängig von der ursprünglichen Maßeinheit sind. Das sollte aber nicht dazu verleiten unbedacht eine logarithmische Funktionsform zu wählen.

Tabelle 2.15: Zusammenhang zwischen Lebenserwartung und pro Kopf Einkommen in einem Länderquerschnitt (in der level-level Spezifikation wurde das GDPpc in 1000 $ gemessen. Wie in Abbildung 2.39 ersichtlich ist liefert die level-level Spezifikation eine sehr schlechte Anpassung an die Daten.

	Dependent variable:	
	LifeExp	
	(1)	(2)
Constant	19.303***	66.820***
	(2.936)	(0.724)
log(GDPpc)	5.683***	
	(0.314)	
GDPpc/1000		0.266***
		(0.026)
Observations	149	149
R^2	0.691	0.425
Note:	*p<0.1; **p<0.05; ***p<0.01	

Das entscheidende Argument für die Wahl der Funktionsform sollte sein, welches Modell mit der Theorie konsistent ist und die Daten besser abbildet. Prinzipiell existieren auch Tests für die Wahl der Funktionsform (z.B. Test von MacKinnon, White & Davidson), doch haben diese oft keine sehr große Power. Manchmal reicht auch schon ein Blick auf einen Scatterplot oder das Histogramm der geschätzten Residuen um zu erkennen, welche Funktionsform eher angebracht ist.

Prozent versus Prozent*punkte*: Vorsicht ist geboten, wenn *Wachstumsraten* und/oder *Anteile* in Regressionen verwendet werden. Für die Interpretation der Koeffizienten solcher Regressionen ist es wesentlich zwischen Prozent und Prozent*punkten* zu unterscheiden. Wenn z.B. die Arbeitslosenrate von 5% auf 6% steigt, so ist dies eine Zunahme um einen Prozent*punkt*, aber eine Zunahme von 20 Prozent ($= (6-5)/5 \times 100 = 20$) gegenüber dem ursprünglichen Niveau. Man beachte, dass die Differenz der Logarithmen näherungsweise einer relativen Änderung entspricht, z.B. $\log(6) - \log(5) = 0.1823 \approx 0.2$ (beachte, dass $(\exp(0.1823) - 1) \times 100 = 20$).

Angenommen in der Regression $\widehat{\log(y)} = b_1 + b_2 A$ sei A ein Anteil, d.h. eine Zahl zwischen Null und Eins ($0 \leq A \leq 1$), es handelt sich also um ein übliches log-level Modell. Deshalb gibt $100b_2$ näherungsweise an, um wie viel *Prozent* sich $\widehat{y}$ ändert, wenn A um *eine Einheit* zunimmt. Aber was soll man sich unter 'einer Einheit' eines Anteils vorstellen? Wenn wir A mit 100 multiplizieren erhalten wir Prozent, bzw. eine Zunahme um einen *Prozentpunkt*, denn im log-level Modell $\widehat{\log(y)} = b_1 + b_2 A$ ist

$$b_2 \approx \frac{\frac{\Delta \widehat{y}}{\widehat{y}}}{\Delta A} = \frac{\frac{\Delta \widehat{y}}{\widehat{y}} \times 100}{\Delta A \times 100} = \frac{\text{prozentuelle Änderung von } \widehat{y}}{\text{Zunahme von } A \text{ um einen Prozent}punkt}$$

Deshalb gibt b_2 näherungsweise an, um wie viel Prozent sich $\widehat{y}$ ändert, wenn A um *einen Prozentpunkt* zunimmt (wie immer in log-level Modellen erhalten wir einen etwas genaueren Wert mit $[\exp(b_2) - 1]$, und diese Korrektur sollte zumindest durchgeführt werden, wenn $b_2 > 0.1$).

In einem log-log Modell $\widehat{\log(y)} = b_1 + b_2 \log(A)$ wird b_2 wieder wie üblich als Elastizität interpretiert, das heißt, um wieviel Prozent sich $\widehat{y}$ ändert, wenn der Anteil A um *ein Prozent* zunimmt.

Beobachtungen mit Nullen: Wie schon mehrfach betont ist der Logarithmus von Null und negativen Werten nicht definiert, deshalb dürfen Variablen, die negative Werte oder den Wert Null enthalten (oder solche Werte annehmen können), nicht logarithmiert werden!

Ein bisschen eine Glaubensfrage ist, was man tun soll, wenn eine Variable y nur einige wenige Nullen enthält. Falls man aus irgendwelchen Gründen y trotzdem logarithmieren möchte, und es wirklich nur wenige Nullen sind, die von der inhaltlichen Interpretation her auch keinen großen Stellenwert besitzen, empfehlen manche Autoren einfach $\log(1 + y)$ anstelle von $\log(y)$ zu verwenden. Die übliche Prozent-Interpretation bleibt dabei oft zumindest näherungsweise erhalten, mit Ausnahme von Änderungen in der Nähe von $y = 0$, wo sie nicht einmal definiert ist (vgl. **?**, 199). Generell empfiehlt sich in solchen Fällen aber die Anwendung geeigneterer Schätzverfahren, z.B. Tobit oder Poisson Modellen.

Wiederholung

1. Partieller Effekt im linearen Modell $\widehat{y} = b_1 + b_2 x_2 + \cdots + b_h x_h + \cdots + b_k x_k$:

$$b_h = \left.\frac{\Delta \widehat{y}}{\Delta x_h}\right|_{\text{ceteris paribus}}$$

   mit ceteris paribus meinen wir, dass *alle anderen* x-Variablen konstant angenommen werden ($\Delta x_1 = \ldots = \Delta x_{h-1} = \Delta x_{h+1} = \ldots = \Delta x_k = 0$)

2. Partieller Effekt im log-log Modell $\widehat{\log y} = b_1 + b_2 x_2 + \cdots + b_h \log x_h + \cdots + b_k x_k$

$$b_h = \left.\frac{\Delta \widehat{\log y}}{\Delta \log x_h}\right|_{\text{c.p.}} \approx \frac{\frac{\Delta \widehat{y}}{\widehat{y}}}{\frac{\Delta x_h}{x_h}} = \frac{\frac{\Delta \widehat{y}}{\widehat{y}} \times 100}{\frac{\Delta x_h}{x_h} \times 100} \qquad \text{(Elastizität)}$$

3. Partieller Effekt im log-level Modell $\widehat{\log y} = b_1 + b_2 x_2 + \cdots + b_h x_h + \cdots + b_k x_k$

$$b_h = \left.\frac{\Delta \widehat{\log y}}{\Delta x_h}\right|_{\text{c.p.}} \approx \frac{\frac{\Delta \widehat{y}}{\widehat{y}}}{\Delta x_h} \quad \text{oder} \quad 100 \times b_h \approx \frac{\frac{\Delta \widehat{y}}{\widehat{y}} \times 100}{\Delta x_h}$$

4. Partieller Effekt im level-log Modell $\widehat{y} = b_1 + b_2 x_2 + \cdots + b_h \log x_h + \cdots + b_k x_k$

$$b_h = \left.\frac{\Delta \widehat{y}}{\Delta \log x_h}\right|_{\text{c.p.}} \approx \frac{\Delta \widehat{y}}{\frac{\Delta x_h}{x_h}} \quad \text{oder} \quad 0.01 \times b_h = \frac{\Delta \widehat{y}}{\frac{\Delta x_h}{x_h} \times 100}$$

Zusammenfassung: Interpretation der Koeffizienten in Log-Modellen

	Spezifikation	Interpretation
I.	$\log(y_i) = b_1 + b_2 \log(x_i) + e_i$	Eine Zunahme von x um *ein Prozent* geht einher mit einer Änderung von y um b_2 *Prozent*, d.h. b_2 kann unmittelbar als Elastizität interpretiert werden.
II.	$\log(y_i) = b_1 + b_2 x_i + e_i$	Eine Zunahme von x um *eine Einheit* (z.B. einen Euro) geht einher mit einer Änderung von y um *ungefähr* $100 \times b_2$ *Prozent* (wenn $\lvert b_2 \rvert < 0.1$), oder genauer, zu einer Änderung von $[\exp(b_2) - 1] \times 100$ *Prozent*.
III.	$y_i = b_1 + b_2 \log(x_i) + e_i$	Eine Zunahme von x um *ein Prozent* geht einher mit einer Änderung von y um $0.01 \times b_2$ *Einheiten* (z.B. Euro).

(für multiple Regressionen gilt jeweils die *ceteris paribus* Annahme)

2.9 Quadratische Modelle

Lineare Funktionsformen sind praktisch und einfach zu interpretieren, aber leider ist die Realität nicht immer so einfach, manchmal sind die marginalen Effekte nicht konstant, sondern hängen vom Niveau ab.

Zum Beispiel wissen wir aus der Mikroökonomik, dass die kurzfristige Durchschnittskostenfunktion einer Unternehmung häufig einen U-förmigen Verlauf hat.

Eine sehr einfache – wenngleich ziemlich restriktive – Methode zur Modellierung solcher Nichtlinearitäten besteht in der Verwendung von Polynomen, wobei man sich in den meisten Fällen auf quadratische Funktionsformen beschränkt.

$$\widehat{y}_i = b_1 + b_2 x + b_3 x^2$$

Man beachte, dass diese Funktion linear in den Parametern b_1 und b_2 ist, und deshalb einfach mit OLS geschätzt werden kann, aber *nichtlinear in der erklärenden Variable* x ist.

Der *marginalen Effekt* kann wie üblich als Ableitung berechnet werden

$$\frac{d\widehat{y}}{dx} = b_2 + 2b_3 x$$

Offensichtlich ist der marginale Effekt für dieses quadratische Modell nicht konstant, sondern hängt vom Niveau von x ab. Dies ist auch in Abbildung 2.40 ersichtlich;

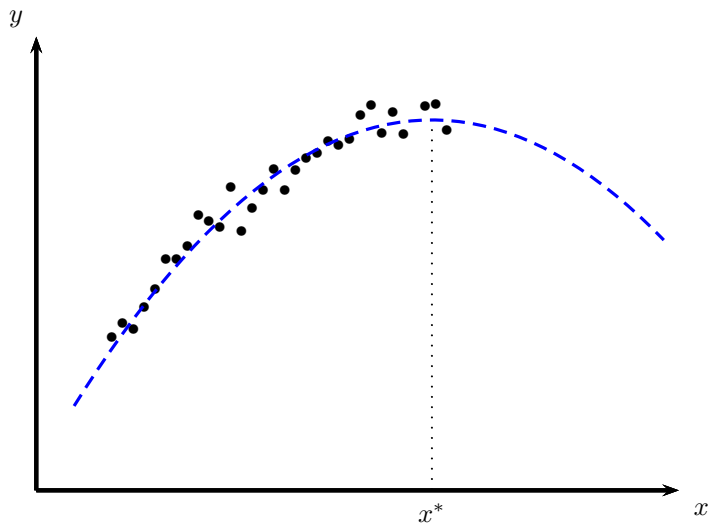


Abbildung 2.40: Quadratische Modelle $\widehat{y} = b_1 + b_2 x + b_3 x^2$ unterstellen einen symmetrischen Verlauf, deshalb können sie in der Stichprobe zwar einen sehr guten Fit haben, aber sehr schlechte Prognosen liefern. In diesem Beispiel ist $b_2 > 0$ und $b_3 < 0$.

dort ist $b_2 > 0$ und $b_3 < 0$, deshalb ist die Steigung bei sehr kleinem x groß, nimmt mit zunehmenden x ab, und wird schließlich negativ.

Achtung: da $d\widehat{y}/dx = b_2 + 2b_3 x$ misst b_2 den marginalen Effekt nur im Punkt $x = 0$, und dieser ist nur selten von Interesse. Meist ist es zweckmäßiger, den marginalen Effekt im Mittelwert der x oder in einem anderen gut interpretierbaren Punkt (z.B. Quartile) anzugeben. Sollte genügend Raum zur Verfügung stehen kann man ihn auch grafisch darstellen.

Achtung: Sollte der Einfluss von x auf $\widehat{y}$ statistisch getestet werden, muss die *gemeinsame* Signifikanz von b_2 *und* b_3 mittels F-Statistik getestet werden; mehr dazu später.

Quadratische Funktionen haben ein Maximum oder Minimum, welches einfach berechnet werden kann, indem man die Ableitung $d\widehat{y}/dx = b_2 + 2b_3 x$ gleich Null setzt und nach x löst

$$x^* = \frac{-b_2}{2b_3}$$

Bevor dieser Extremwert x^* interpretiert wird sollte man allerdings überprüfen, welcher Anteil der Beobachtungen links bzw. rechts vom Extremwert liegen, vergleiche Abbildung 2.40.

Quadratische Log-level Modelle: Angenommen, wir möchten den marginalen Effekt von x im Modell

$$\widehat{\log(y)} = b_1 + b_2 x + b_3 x^2$$

berechnen.

Der übliche marginale Effekt von x ist in diesem Fall nicht konstant, sondern hängt von der Ausprägung von x ab

$$\frac{d\,\widehat{\log(y)}}{d\,x} = b_2 + 2b_3x \approx \frac{\frac{\Delta\widehat{y}}{\widehat{y}}}{\Delta x}$$

Wenn wir angeben wollen, um wie viele *Prozent* sich $\widehat{y}$ ändert, wenn x um *eine Einheit* zunimmt, könnten wir dies z.B. für den Mittelwert $\bar{x}$ berechnen

$$\%\Delta\widehat{y} := \frac{\frac{\Delta\widehat{y}}{\widehat{y}} \times 100}{\Delta x} \approx (\exp(b_2 + 2b_3\bar{x}) - 1) \times 100$$

Allerdings gilt auch dies nicht exakt, wenn wir für diskrete Änderungen Differenzen bilden erhalten wir [27]

$$\begin{aligned}\widehat{\Delta\log(y)} := \log(\widehat{y} + \Delta\widehat{y}) - \widehat{\log(y)} &= b_1 + b_2(x + \Delta x) + b_3(x + \Delta x)^2 - \\ &\quad b_1 - b_2x - b_3x^2 \\ &= b_2\Delta x + b_3[2x\Delta x + (\Delta x)^2]\end{aligned}$$

und für $\Delta x = 1$

$$\log\left(1 + \frac{\Delta\widehat{y}}{\widehat{y}}\right) = b_2 + b_3(2x + 1) = b_2 + 2b_3(x + 0.5) \approx \frac{\Delta\widehat{y}}{\widehat{y}}$$

und

$$\%\Delta\widehat{y} := \frac{\Delta \log \widehat{y}}{\Delta x} \times 100 \approx \frac{\frac{\Delta\widehat{y}}{\widehat{y}} \times 100}{\Delta x} = [\exp(b_2 + 2b_3(x + 0.5)) - 1] \times 100$$

Wie wichtig diese Korrektur ist hängt wieder von den konkreten Zahlenwerten ab.

Beispiel: für österreichische EU-Silc Daten (2018) erhalten wir[28]

$$\begin{aligned}\log(\text{StdL}) = \underset{(0.035)^{***}}{1.871} + \underset{(0.002)^{***}}{0.037\ \text{potBildg}} + \underset{(0.002)^{***}}{0.028\ \text{Erf}} - \underset{(0.00005)^{***}}{0.000366\ \text{Erf}^2}\end{aligned}$$

$$R^2 = 0.15, \qquad n = 4104$$

Demnach nimmt der mittlere Stundenlohn (StdL) mit jedem zusätzlichen Jahr potentieller Bildung (potBildg) ceteris paribus um ca. 3.7% zu.

$$\frac{\Delta\widehat{\log(\text{StdL})}}{\Delta\text{potBildg}} = 0.037$$

oder etwas genauer

$$(\exp(0.037) - 1) * 100 = 3.77\%$$

[27] Dank an Klaus Nowotny für den Hinweis!

[28] Diese Art von Lohngleichungen werden zu Ehren von Jacob Mincer (1922 – 2006, einer der Begründer der modernen empirischen Arbeitsmarktökonomik) häufig Mincer-Einkommensgleichungen genannt. Für eine Übersicht zur Schätzung und Messung des Einflusses von Humankapital auf das Einkommen siehe z.B. die frei verfügbaren Beiträge von **?** und **?**.

Hingegen ist der durchschnittliche Effekt der Berufserfahrung (Erf) nicht konstant

$$\frac{\Delta\widehat{\log(\text{StdL})}}{\Delta\text{Erf}} = 0.028 - 2 * 0.000366\text{Erf}$$

er nimmt zuerst zu, erreicht nach ca. 38 Jahren ein Maximum, da[29]

$$\text{Erf}^{\max} = \frac{b_3}{-2b_4} = \frac{0.028}{-2(-0.000366)} = 38.25$$

und nimmt nachher ab.

Der marginale Effekt der Berufserfahrung für einen Anfänger (d.h. für Erf = 0) ist ceteris paribus

$$\%\Delta\widehat{\log(\text{StdL})} = [\exp(0.028 - 2 \times 0.000366(0 + 0.5)) - 1]100 = 2.8$$

d.h. im ersten Berufsjahr nimmt der mittlere Stundenlohn ceteris paribus um ca. 2.8% zu, ...

$$\%\Delta\widehat{\log(\text{StdL})} = [\exp(0.028 - 2 \times 0.000366(20 + 0.5)) - 1]100 = 1.31$$

nach 20 Jahren (d.h. Erf = 20) nimmt der gefittete Stundenlohn mit einem weiteren Jahr Berufserfahrung ceteris paribus nur noch um ca. 1.31 Prozent zu.

2.10 Interaktions-Modelle

Als erklärende Variablen können auch *Produkte* einzelner erklärender Variablen verwendet werden, z.B. im folgenden Modell das Produkt von x_2 und x_3

$$\widehat{y} = b_1 + b_2x_2 + b_3x_3 + b_4(x_2x_3)$$

In diesen Modellen werden x_2 und x_3 Hauptterme oder Haupteffekte (*'main terms'*) genannt und das Produkt x_2x_3 wird als Interaktionsterm (*'interaction term'*) bezeichnet.

Wenn in einem Modell Interaktionsterme berücksichtigt werden sollten auf jeden Fall auch die Haupteffekte berücksichtigt werden, da sonst die Gefahr einer Fehlspezifikation aufgrund fehlender relevanter Variablen extrem groß ist (der Interaktionsterm x_2x_3 ist praktisch immer mit x_2 und x_3 korreliert (vgl. **?**)).

Der marginale Effekt von x_2 hängt vom Niveau von x_3 ab,

$$\frac{\partial\widehat{y}}{\partial x_2} = b_2 + b_4x_3$$

d.h. die ceteris paribus Auswirkung einer Änderung von x_2 auf $\widehat{y}$ hängt auch vom absoluten Wert von x_3 an der betreffenden Stelle ab.[30]

[29] Wir setzen die erste Ableitung nach Erf gleich Null und lösen nach Erf.

[30] Wenn später getestet werden soll, ob x_2 einen Effekt auf y hat, darf nicht der einfache t-Test für den Koeffizienten von x_2 herangezogen werden, sondern es muss z.B. mit einem F-Test die simultane Nullhypothese $\beta_2 = 0$ und $\beta_4 = 0$ getestet werden.

Man kann die Abhängigkeit des marginalen Effekts von x_2 vom Wert von x_3 z.B. grafisch darstellen, indem man in einer Grafik $\partial y/\partial x_2$ gegen x_3 aufträgt.

Achtung: Der Koeffizient b_2 misst den marginalen Effekt von x_2 nur im Punkt $x_3 = 0$!

Analog gilt für eine Änderung von x_3, wenn x_2 konstant gehalten wird,

$$\frac{\partial \widehat{y}}{\partial x_3} = b_3 + b_4 x_2$$

Aus der Tatsache, dass der Koeffizient des Interaktionsterms im linearen Modell

$$\widehat{y} = b_1 + b_2 x_2 + b_3 x_3 + b_4 (x_2 x_3)$$

einfach die zweite Ableitung ist

$$b_4 = \frac{\frac{\partial \widehat{y}}{\partial x_2}}{\partial x_3} = \frac{\frac{\partial \widehat{y}}{\partial x_3}}{\partial x_2} = \frac{\partial^2 \widehat{y}}{\partial x_2 \partial x_3}$$

folgt eine weitere wichtige Beobachtung:

> Der Koeffizient des Interaktionsterms b_4 gibt an, wie sich der marginale Effekt von x_2 ändert, wenn x_3 um eine (infinitesimale) Einheit zunimmt.

Man beachte, dass die Funktionsform eine Symmetrie der marginalen Effekte erzwingt (dies folgt aus der Symmetrie der zweiten Ableitungen, vgl. Young's Theorem), d.h. b_4 kann auch als Änderung des marginalen Effekts von x_3 interpretiert werden, wenn x_2 um eine Einheit zunimmt.

Beispiel: Die Wirksamkeit von Entwicklungshilfe ist seit jeher sehr umstritten, einfache Regressionen von Indikatoren für die empfangene Entwicklungshilfe auf die Wachstumsrate der Empfängerländer zeigten regelmäßig keinen Zusammenhang oder lieferten widersprüchliche Ergebnisse.

In einem sehr einflussreichen Paper zur Entwicklungshilfe argumentierten **?** *"We find that aid has a positive impact on growth in developing countries with good fiscal, monetary, and trade policies but has little effect in the presence of poor policies."* Ihre Aussage begründeten Sie mit einer etwas komplexeren OLS Regression, die (stark) vereinfacht folgendermaßen aussieht

$$G = b_1 + b_2 A + b_3 P + b_4 (A \times P) + \cdots + e$$

dabei sind G (*growth*) die Wachstumsrate der Empfängerländer, A (*aid*) ein Indikator für die empfangene Entwicklungshilfe, und P (*policy*) ist ein Indikator für 'gute' Wirtschaftspolitik, beruhend auf makroökonomischen Variablen wie z.B. Inflationsrate, Budgetdefizit, etc.

Das Argument von **?**, das Entwicklungshilfe nur in Ländern mit 'guter' Wirtschaftspolitik wirkt, beruhte auf dem Koeffizienten der Interaktionsvariable, der in ihrem Modell positiv und statistisch signifikant von Null verschieden war. Das bedeutet,

dass Entwicklungshilfe in Ländern mit 'guter' Wirtschaftspolitik positive Auswirkungen auf die Wachstumsrate hat

$$\frac{\partial \widehat{G}}{\partial A} = b_2 + b_4 P$$

Dieses Papier war politisch extrem einflussreich, da es gut in das Weltbild vieler betroffener Akteure passte, aber es wurde bald gezeigt, dass dieses Resultat sehr stark von der Länderauswahl und der Zeitperiode abhängig war. Für eine größere Zahl von Ländern und spätere Jahre konnte das Resultat häufig nicht reproduziert werden (**?**).

2.10.1 Alternative Parametrisierung von Interaktionsmodellen*

Durch eine einfache 'Reparametrisierung' kann ein alternatives Interaktionsmodell geschätzt werden, dessen Koeffizienten direkt den marginalen Effekt im Mittelwert der betreffenden Variable messen.

Beginnen wir mit dem einfachen Interaktionsmodell

$$\widehat{y} = b_1 + b_2 x_2 + b_3 x_3 + b_4 x_2 x_3 \tag{2.19}$$

Wenn uns z.B. der marginale Effekt von x_2 gemessen im *Mittelwert von* x_3 interessiert können wir diesen einfach berechnen

$$\frac{\partial \widehat{y}}{\partial x_2} = b_2 + b_4 \bar{x}_3 \tag{2.20}$$

Wir werden nun zeigen, dass wir diesen marginalen Effekt *im Mittelwert der anderen Variablen* noch einfacher durch eine einfache Variablentransformation schätzen können.

Angenommen wir schätzen anstelle von Gleichung (2.19) ein reparametrisiertes Modell

$$y = a_1 + a_2 x_2 + a_3 x_3 + a_4 (x_2 - \bar{x}_2)(x_3 - \bar{x}_3) \tag{2.21}$$

wobei $\bar{x}_2$ und $\bar{x}_3$ wie üblich die Mittelwerte bezeichnen, dann misst a_2 den marginalen Effekt von x_2 *im Mittelwert von* x_3, d.h. gibt exakt den gleichen Wert, den wir aus Gleichung (2.20) erhalten!

Warum dies so ist kann einfach gezeigt werden. Wenn wir den Interaktionsterm ausmultiplizieren erhalten wir

$$\begin{aligned} y &= a_1 + a_2 x_2 + a_3 x_3 + a_4 [x_2 x_3 - x_2 \bar{x}_3 - x_3 \bar{x}_2 + \bar{x}_2 \bar{x}_3] \\ &= a_1 + a_4 \bar{x}_2 \bar{x}_3 + (a_2 - a_4 \bar{x}_3) x_2 + (a_3 - a_4 \bar{x}_2) x_3 + a_4 x_2 x_3 \end{aligned}$$

Aus einem Vergleich mit Gleichung (2.19) sehen wir sofort, dass $(a_2 - a_4 \bar{x}_3) = b_2$, $(a_3 - a_4 \bar{x}_2) = b_3$ und $a_4 = b_4$ sein muss.

Was haben wir damit gewonnen? Aus $(a_2 - a_4 \bar{x}_3) = b_2$ und $a_4 = b_4$ folgt

$$a_2 = b_2 + b_4 \bar{x}_3$$

dies ist aber exakt der marginale Effekt von x_2 gemessen im *Mittelwert von* x_3 den wir aus Gleichung (2.20) erhalten haben!

Wenn wir also anstelle von Gleichung (2.19) die reparametrisierte Gleichung (2.21) schätzen misst der Koeffizienten a_2 direkt den marginalen Effekt von x_2 *im Mittelwert von* x_3, mit dem dazugehörigen korrekten Standardfehler des marginalen Effekts im Mittelwert. Die dazugehörige t-Statistik kann also unmittelbar verwendet werden um zu testen, inwieweit der marginale Effekt *im Mittelwert von* x_3 von Null verschieden ist. Analoges gilt für a_3.

Dies ist offensichtlich interessanter als die Schätzung von b_2 aus Gleichung (2.19), denn dieser Koeffizient misst den marginalen Effekt nur im Punkt $x_3 = 0$, der kaum je relevant sein dürfte.

Allerdings ändert dies nichts an der Nicht-Linearität des Zusammenhangs, der marginale Effekt von x_2 ist für jeden Wert von x_3 unterschiedlich!

Zusammenfassend einige abschließende Empfehlungen für den Umgang mit Interaktionseffekten von **?**:

- Interpretieren Sie die Koeffizienten von Interaktionsvariablen nie als unbedingte marginale Effekte!
- Falls Sie an den marginalen Effekten interessiert sind berechnen Sie diese für interessierende Werte der anderen Variablen, oder stellen sie diese grafisch dar.
- Berechnen Sie den marginalen Effekte für relevante Werte der Variablen, z.B. im Mittelwert oder Median. Eine einfache Möglichkeit dazu bietet die oben erläuterte alternative Parametrisierung.
- Für Signifikanztest: testen Sie die *gemeinsame* Signifikanz der entsprechenden Koeffizienten! (z.B. mit Hilfe eines F-Tests, kommt später im Kapitel über Hypothesentests im multiplen Regressionsmodell).

Für eine neuere und ausführliche Diskussion von multiplikativen Interaktionseffekten mit praktischen Hinweisen siehe **?**.

Beispiel: Tabelle 2.16 zeigt zwei Schätzungen für Lohngleichungen mit Interaktionseffekten. Spalte (1) von Tabelle 2.16 zeigt den unmittelbaren Interaktionseffekt wie in Gleichung (2.19), Spalte (2) die Schätzung für das reparametrisierte Modell wie in Gleichung (2.21).

Alle Koeffizienten weisen das erwartete Vorzeichen auf und sind hoch signifikant von Null verschieden.

Spalte (1) von Tabelle 2.16 zeigt die Schätzung für die Gleichung

$$\widehat{\log(\text{StdL})} = b_1 + b_2\,\text{Bildg} + b_3\,\text{Erf} + b_4\,\text{Erf}^2 + b_5\,\text{Bildg} \times \text{Erf}$$

Der marginale Effekt von Bildg ist

$$\frac{\partial \widehat{\log(\text{StdL})}}{\partial \text{Bildg}} = b_2 + b_5 \text{Erf}$$

Tabelle 2.16: Lohngleichung für Österreich mit Interaktionseffekten (Daten: EU-Silc 2018)

Dependent Var.: log(StdL)	(1)	(2)
Constant	2.31315***	1.88338***
	(0.05849)	(0.03429)
Bildg	0.00851**	0.04287***
	(0.00342)	(0.00176)
Erf	0.00024	0.01909***
	(0.00379)	(0.00252)
Erf^2	−0.00015***	−0.00015***
	(0.00005)	(0.00005)
(Bildg × Erf)	0.00151***	
	(0.00016)	
$(\text{Bildg} - \overline{\text{Bildg}}) \times (\text{Erf} - \overline{\text{Erf}})$		0.00151***
		(0.00016)
Observations	4,104	4,104
R^2	0.16744	0.16744
Adjusted R^2	0.16663	0.16663
F Statistic (df = 4; 4099)	206.09320***	206.09320***
Note:		*p<0.1; **p<0.05; ***p<0.01

Diese Funktion ist im linken Panel von Abbildung 2.41 dargestellt. Wenn wir die geschätzten Koeffizienten aus Tabelle 2.16 und den Mittelwert für die Berufserfahrung $\overline{\text{Erf}} = 22.79$ einsetzen erhalten wir

$$\frac{\partial \widehat{\log(\text{StdL})}}{\partial \text{Bildg}} = 0.00851 + 0.00151 \times 22.79 = 0.04287$$

also genau den Wert des Koeffizienten von Bildg in Spalte (2). Das ist natürlich nicht überraschend, denn dies haben haben wir im vorhergehenden Abschnitt über die Reparametrisierung allgemein bewiesen. Für jemanden mit 22.79 Jahren Berufserfahrung erwarten wir also, dass ein *weiteres Jahr an Ausbildung* ungefähr einen um 4.287% höheren Stundenlohn bringt.

Die gute Nachricht ist, dass der Wert der einmal erworbenen Bildung mit der Berufserfahrung zunimmt, der marginale Effekt der Bildung steigt mit der Erfahrung.

Ebenfalls durch eine einfache partielle Ableitung können wir den marginalen Effekt für die Berufserfahrung 'Erf' berechnen

$$\frac{\partial \widehat{\log(\text{StdL})}}{\partial \text{Erf}} = b_3 + 2b_4\,\text{Erf} + b_5\text{Bildg}$$

In diesem Fall hängt der marginale Effekt der Berufserfahrung auf den Stundenlohn vom Level des Ausbildungsjahre *und* von der Berufserfahrung ab!

Offensichtlich unterscheidet sich dieser marginale Effekt je nach Bildung und Erfahrung.

Aber wir können mit Hilfe der Koeffizienten aus Spalte (1) aus Tabelle 2.16 die marginalen Effekte für verschiedene Werte von für Bildung und Erfahrung berechnen, z.B. den marginalen Effekt in den Mittelwerten ($\overline{\text{Bildg}} = 12.51, \overline{\text{Erf}} = 22.79$)

$$\frac{\partial \widehat{\log(\text{StdL})}}{\partial \text{Erf}} = 0.0002371 + 2 \times (-0.000154) \times 22.79 + 0.00151 \times 12.51 \approx 0.0121$$

Das heißt, jemand mit durchschnittlicher Bildung und Erfahrung kann bei einem zusätzlichen Jahr Erfahrung mit einem Lohnanstieg von 1.21% rechnen (natürlich sprechen wir von einer linearen Approximation ...).

Für jemand mit durchschnittlicher Bildung und Erf = 0 erhalten wir einen marginalen Effekt der Erfahrung von 0.0191, den Wert aus Spalte (2) aus Tabelle 2.16.

So können wir für jede Bildg – Erf Kombination einen marginalen Effekt berechnen, aber dies ist nicht sehr zweckmäßig. Man kann sich entweder auf interessierende Kombinationen beschränken, oder die Effekte grafisch zeigen, siehe Abbildung 2.41, aber diese Möglichkeit wird für so einfache Modelle eher selten genützt.

Das rechte Panel in Abbildung 2.41 zeigt die Abhängigkeit des marginalen Effekts von Erfahrung (Erf) vom Stundenlohn für unterschiedliche Ausbildungsdauer Bildg und für drei Levels von Erf (nach Quartilen) ist im rechten Panel von Abbildung 2.41. Offensichtlich nimmt der marginale Effekt der Berufserfahrung ceteris paribus mit der Ausbildungsdauer zu, und ist ceteris paribus umso höher, je geringer die Berufserfahrung ist.

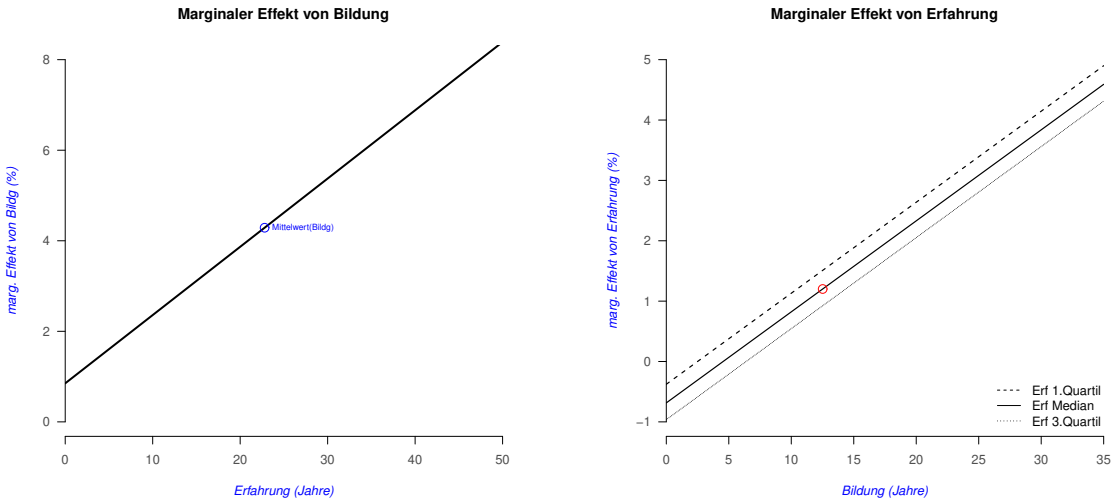


Abbildung 2.41: Marginale Effekte mit Interaktionseffekten für Lohngleichung.

Wenn wir für Erf anstelle von Null den Median von 20 Jahren einsetzen erhalten wir den Wert 0.01565, das heißt, für jemanden mit 13.8 Jahren Ausbildung erwarten wir, dass der Stundenlohn um ca. 1.565% steigt, wenn die Berufserfahrung von 20 Jahren auf 21 Jahre zunimmt. Dieser Punkt ist in der Grafik eingezeichnet.

Vorsicht mit nicht-linearen Funktionsformen:

- Mit polynomischen Modellen (z.B. quadratischen oder kubischen Modellen) kann man zwar manchmal einen sehr guten Fit in der Stichprobe erreichen, aber für Prognosen sind sie meistens ziemlich unbrauchbar, da die Funktionsform *'out of sample'* häufig extreme Verläufe erzwingt.

- Das Bestimmtheitsmaß R^2 darf nur für den Vergleich von Modellen verwendet werden, in denen die abhängige Variable y nicht transformiert wurde *und* wenn beide Modelle die gleiche Anzahl erklärender Variablen haben.

  Dies ist einfach zu erkennen: das Bestimmtheitsmaß ist definiert als Anteil der durch die x erklärten Streuung an der gesamten Streuung von y. Falls y transformiert wird (also z.B. $\log(y)$ anstelle von y verwendet wird), ändert sich natürlich auch die Streuung, und damit das R^2

  $$R_y^2 = 1 - \frac{\sum_i e_i^2}{\sum_i (y_i - \overline{y})^2} \quad \Leftrightarrow \quad R_{\log(y)}^2 = 1 - \frac{\sum_i e_i^2}{\sum_i [\log(y_i) - \overline{\log(y)}]^2}$$

  Dies gilt auch für das korrigierte Bestimmtheitsmaß $\bar{R}^2$, welches bloß einen Vergleich von Modellen mit einer *unterschiedliche* Anzahl erklärender Variablen ermöglicht.

- Grundsätzlich sollte weder das normale R^2 noch das korrigierte Bestimmtheitsmaß $\bar{R}^2$ für die Einschätzung der Qualität einer Schätzung überinterpretiert werden, da sie nur die Anpassung in der Stichprobe beschreiben. Viel wichtiger ist z.B., ob die geschätzten Koeffizienten die theoretischen Erwartungen erfüllen und signifikant von Null verschieden sind. Im Kapitel über *Spezifikation* werden wir einige weitere Kriterien und Tests kennen lernen.

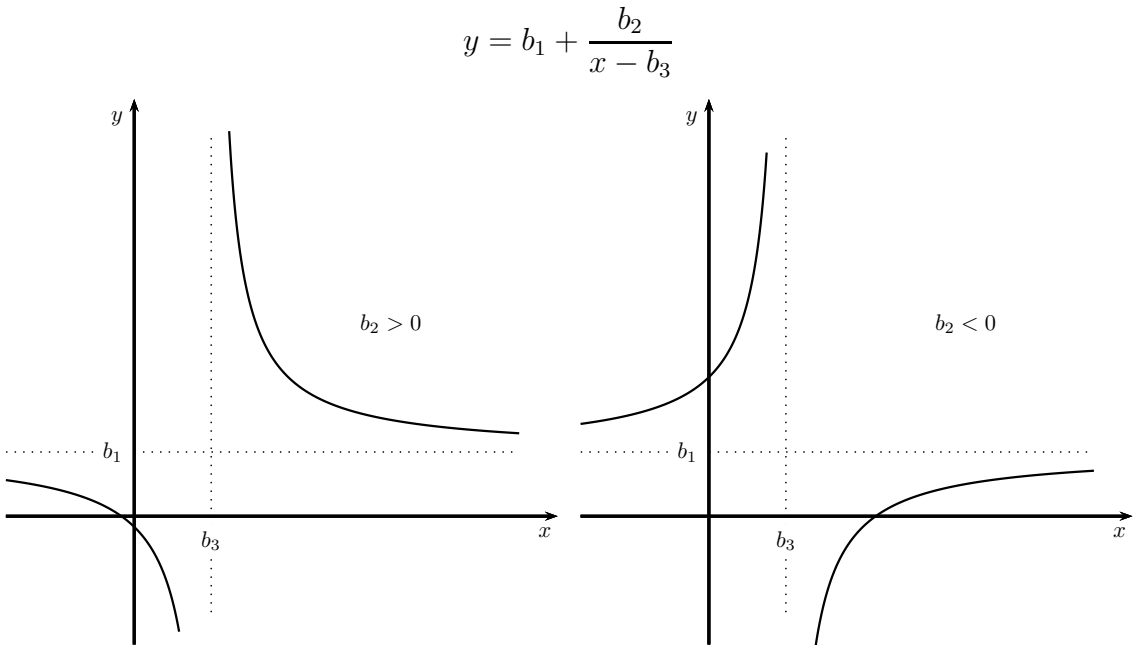


Abbildung 2.42: Reziproke Transformationen

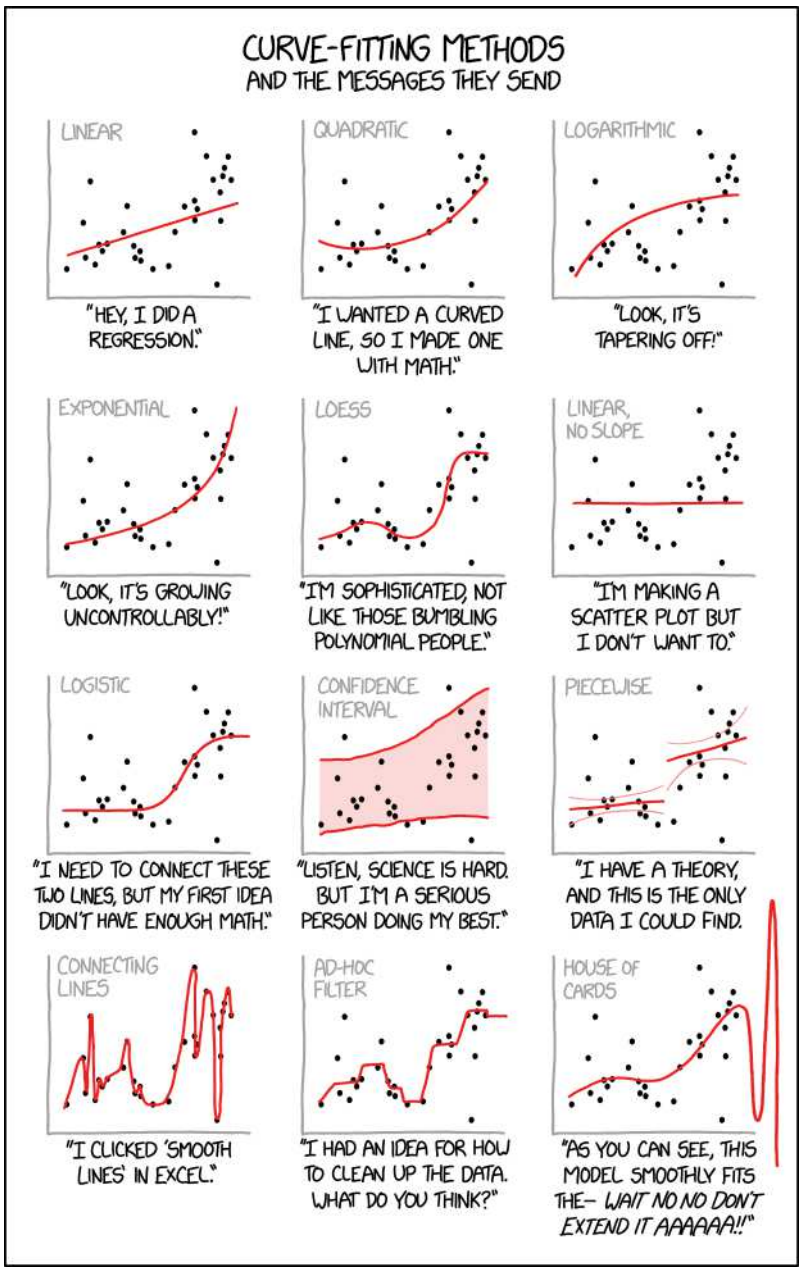
2.11 Reziproke Transformationen

Eine andere Funktionsform, die z.B. für die Schätzung von Phillips-Kurven herangezogen werden kann, sind reziproke Transformationen, man verwendet einfach den Kehrwert der Variablen, siehe Abbildung 2.42.

$$y = b_1 + b_2 \frac{1}{x}$$

Übersicht:

Modell	**Gleichung**	**Steigung** $(= \frac{dy}{dx})$	**Elastizität** $(= \frac{dy}{dx}\frac{x}{y})$
Linear	$y = \alpha + \beta x$	β	$\beta(x/y)$
Log-log	$\log y = \alpha + \beta \log x$	$\beta(y/x)$	β
Log-level	$\log y = \alpha + \beta x$	$\beta(y)$	$\beta(x)$
Level-log	$y = \alpha + \beta \log x$	$\beta(1/x)$	$\beta(1/y)$
Reziprok	$y = \alpha + \beta(1/x)$	$-\beta(1/x^2)$	$-\beta(1/xy)$



Quelle: XKCD, `https://xkcd.com/2048/`

2.12 Diverses

2.12.1 Mittelwerttransformationen

Es gibt eine spezielle Datentransformation, die in der Ökonometrie häufig angewandt wird und die sich später oft als nützlich erweisen wird, nämlich die Mittelwerttransformation.

Die Mittelwerttransformation besteht einfach darin, dass von jeder einzelnen Beobachtung x_i einer Datenreihe der Mittelwert der selben Datenreihe $\bar{x}$ subtrahiert wird.

Die resultierende Datenreihe besteht einfach aus Abweichungen vom Mittelwert, daher der Name Mittelwerttransformation. Wir werden eine derart transformierte Beobachtung (bzw. Datenreihe) im Folgenden mit zwei Punkten über dem betreffenden Variablennamen kennzeichnen, also z.B.

$$\ddot{x}_i := x_i - \bar{x}$$

Abbildung 2.43 zeigt eine grafische Interpretation dieser Mittelwerttransformation. Durch diese Transformation "Subtraktion des Mittelwertes" werden die Koordinaten der so transformierten Variable im Verhältnis zu einem neuen Koordinatensystem gemessen, dessen neuer Nullpunkt im Mittelwert der ursprünglichen Variablen $(\bar{x}, \bar{y})$ liegt. Gewissermaßen bewirkt die Subtraktion des Mittelwertes also eine Verschiebung des Koordinatensystems, so dass der neue Nullpunkt in den Mittelwert der Daten verschoben wird.

Solche mittelwerttransformierte Daten werden uns wiederholt begegnen, und sind uns auch schon begegnet; zum Beispiel wird Gleichung (2.8) für b_2 aus den mittelwertransformierten Variablen x und y gebildet, d.h.

$$b_2 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} := \frac{\sum_i \ddot{x}_i \ddot{y}_i}{\sum_i \ddot{x}_i^2}$$

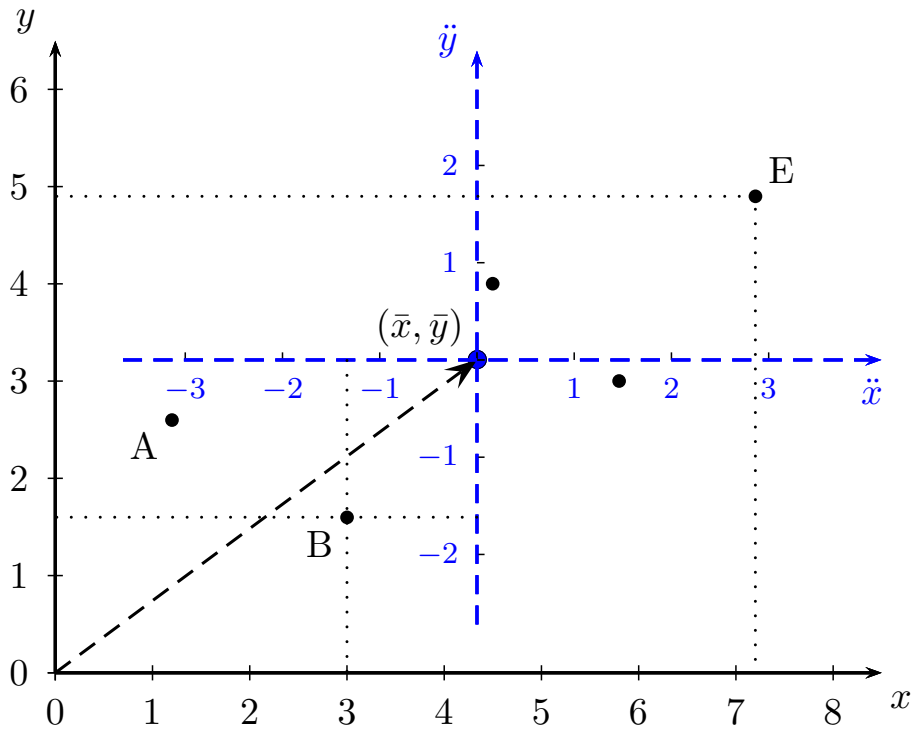
Man beachte auch, dass der Mittelwert einer mittelwertransformierten Variablen stets Null ist, denn

$$\bar{\ddot{y}} := \frac{1}{n}\sum_i \ddot{y}_i = \frac{1}{n}\sum_i (y_i - \bar{y}) = \frac{1}{n}\sum_i y_i - \frac{1}{n}\sum_i \bar{y} = \bar{y} - \frac{1}{n} n\bar{y} = \bar{y} - \bar{y} = 0$$

Daraus folgt zum Beispiel auch, dass

$$b_2 = \frac{\mathrm{cov}(y,x)}{\mathrm{var}(x)} = \frac{\mathrm{cov}(\ddot{y},\ddot{x})}{\mathrm{var}(\ddot{x})}$$

Deshalb spielt es für die Berechnung des Steigungskoeffizienten b_2 keine Rolle, ob man die ursprünglichen Datenreihen oder mittelwerttransformierte Datenreihen verwendet, die OLS-Methode liefert in beiden Fällen das gleiche Ergebnis für den Steigungskoeffizienten.



Daten:

	y	x	$\ddot{y}$	$\ddot{x}$
A	2.60	1.20	−0.62	−3.14
B	1.60	3.00	−1.62	−1.34
C	4.00	4.50	0.78	0.16
D	3.00	5.80	−0.22	1.46
E	4.90	7.20	1.68	2.86
Mittelwert:	3.22	4.34	0.00	0.00

Abbildung 2.43: Datentransformation, Subtraktion des Mittelwertes. Die Koordinaten des Punktes B im ursprünglichen Koordinatensystem sind $(3.0, 1.6)$; wenn der Mittelwert subtrahiert wird erhält man die Koordinaten in Bezug auf ein neues Koordinatensystem, dessen Ursprung im Mittelwert der Beobachtungen $(\bar{x}, \bar{y})$ liegt, für Punkt B also $(-1.34, -1.62)$. [local,www]

Allerdings kann aus den mittelwerttransformierten Datenreihen das Interzept b_1 nicht mehr unmittelbar berechnet werden, denn dies fällt bei der Mittelwerttransformation raus

$$\begin{array}{rcll} y_i & = & b_1 + b_2 x_i + e_i & \\ \bar{y} & = & b_1 + b_2 \bar{x} + \bar{e} & /- \\ \hline y_i - \bar{y} & = & b_1 - b_1 + b_2(x_i - \bar{x}) + e_i - \bar{e} & \\ \ddot{y}_i & = & b_2 \ddot{x}_i + \ddot{e}_i & \end{array}$$

Dies sollte nicht erstaunen, denn wie wir vorhin gesehen haben entspricht die Mittelwerttransformation grafisch einer Verschiebung des Nullpunkts des Koordinatensystems in den Mittelwert der Variablen, und dort in das Interzept per Definition Null.

Aber selbstverständlich kann das Interzept aus den nicht-transformierten Daten mit $b_1 = \bar{y} - b_2 \bar{x}$ einfach wieder berechnet werden.

Übungsbeispiel: Mit Hilfe der mittelwerttransformierten Daten können wir den Zusammenhang $y_i = b_1 + b_2 x_i + e_i$ kürzer schreiben $\ddot{y}_i = b_2 \ddot{x}_i + e_i$, denn der OLS Schätzer b_2 ist tatsächlich in beiden Fällen der selbe.

Wir können zur Übung den OLS Schätzer für das mittelwerttransformierte Modell herleiten. Die Residuen sind $e_i = \ddot{y}_i - b_2 \ddot{x}_i$, deshalb ist das Minimierungsproblem

$$\min_{b_2} \sum_i e_i^2 = \min_{b_2} \sum_i (\ddot{y}_i - b_2 \ddot{x}_i)^2$$

Die Bedingung erster Ordnung ist

$$\frac{d \sum_i e_i^2}{d\, b_2} = -2 \sum_i (\ddot{y}_i - b_2 \ddot{x}_i)(-\ddot{x}_i) = 0$$

Daraus folgt $\sum_i \ddot{y}_i \ddot{x}_i = b_2 (\sum_i \ddot{x}_i)^2$ oder

$$b_2 = \frac{\sum_i \ddot{x}_i \ddot{y}_i}{\sum_i \ddot{x}_i^2} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{\mathrm{cov}(x, y)}{\mathrm{var}(x)}$$

Das Interzept kann wieder wie üblich mit $b_1 = \bar{y} - b_2 \bar{x}$ berechnet werden. Wie wir bereits erwähnt haben kann dieses Ergebnis als ein Spezialfall des Frisch-Waugh-Lovell Theorems interpretiert werden.

2.12.2 Skalierung

Manchmal ist es erforderlich die Skalierung von Variablen zu ändern, zum Beispiel wenn Preise in Dollar statt in Euro, die Zeit in Monaten statt in Jahren oder die Entfernungen in Kilometern statt in Metern angegeben werden soll.

In solchen Fällen werden abhängige und/oder erklärende Variablen mit einer konstanten Zahl multipliziert (z.B. dem Wechselkurs). Welche Konsequenzen hat dies für Schätzungen? Müssen wir die Schätzung mit den linear transformierten Daten neu durchführen, oder können wir die Resultate einfach aus den alten Schätzungen berechnen?

Es zeigt sich, dass es in solchen Fällen genügt, die Koeffizienten und Standardfehler die Schätzungen ebenfalls linear zu transformieren, die Schätzungen müssen *nicht* neu durchgeführt werden.

Erinnern wir uns, dass die Koeffizienten im einfachen linearen Modell den marginalen Effekt messen, d.h. um wie viele Einheiten sich $\widehat{y}$ ändert, wenn x ceteris paribus um eine Einheit zunimmt. Wenn wir nun eine erklärende x Variable mit einer Konstanten c multiplizieren misst der neue Koeffizient um wie viele Einheiten sich $\widehat{y}$ ändert, wenn die erklärende Variable um 'eine *neue* Einheit' $(= cx)$ zunimmt; wir brauchen also lediglich den ursprünglichen Koeffizienten durch c dividieren

$$\widehat{y} = b_1 + (\underbrace{\frac{1}{c} b_2}_{b_2^*})(cx) + e$$

wobei b_2^* den Koeffizienten der skalierten Gleichung bezeichnet.

Kehren wir zu unserem alten Beispiel mit den Gebrauchtautos zurück, die erste Spalte von Tabelle 2.17 zeigt die ursprüngliche Schätzung, die zweite Spalte zeigt was passiert, wenn wir das Alter nicht in Jahren, sondern in Monaten angeben, d.h. wir multiplizieren das Alter in Jahren mit 12. Dies ist eine einfache lineare Transformation und wir sehen, dass sich das Interzept, der Koeffizient von km sowie das Bestimmtheitsmaß R^2 dadurch nicht ändert.

Tabelle 2.17: Skalierung:

	Dependent variable:		
	Preis in €		Preis in 1000 €
	(1)	(2)	(3)
Constant	$22,649.880^{***}$	$22,649.880^{***}$	22.650^{***}
	(411.870)	(411.870)	(0.412)
Alter in Jahren	$-1,896.264^{***}$		
	(235.215)		
Alter in Monaten		-158.022^{***}	-0.158^{***}
		(19.601)	(0.020)
km	-0.031^{***}	-0.031^{***}	-0.00003^{***}
	(0.008)	(0.008)	(0.00001)
Observations	40	40	40
R^2	0.907	0.907	0.907
Note:			*p<0.1; **p<0.05; ***p<0.01

Der Koeffizient der zweiten Gleichung gibt an, dass der gefittete Preis ceteris paribus um 158.022 Euro abnimmt, wenn das Auto um *einen Monat* älter wird.

Wie vorhin behauptet können wir den Koeffizienten und den Standardfehler (die Zahl in Klammern unter dem Koeffizienten) auch einfach aus der ursprünglichen

Gleichung berechnen: für den Koeffizienten $-1\,896.264/12 = -158.022$ und für den Standardfehler $235.215/12 = 19.601$.

Wenn die Skalierung der *abhängigen Variable* geändert wird, d.h. wenn y mit einer Konstanten d multipliziert wird, muss auch die rechte Seite der Gleichung mit d multipliziert werden, d.h. für $y^* := dy$

$$y^* = dy = \underbrace{db_1}_{b_1^*} + \underbrace{db_2}_{b_2^*} x + \underbrace{de}_{e^*}$$

Die dritte Spalte von Tabelle 2.17 zeigt die Schätzung, wenn der Preis anstelle von Euro in Einheiten von tausend *Euro* gemessen wird (d.h. $d = 1/1000 = 0.001$). Wenn das Alter um ein *Monat* zunimmt sinkt der Preis ceteris paribus um 0.158 *tausend* Euro (= 158.022 Euro).

Etwas allgemeiner kann man dies für das bivariate Modell zeigen, indem wir für zwei beliebige Konstante $c, d > 0$ ein skaliertes Modell mit

$$y^* := dy \quad \text{und} \quad x^* := cx$$

definieren und die Schätzfunktionen des skalierten Modells $y^* = b_1^* + b_2^* x^* + e^*$ mit dem ursprünglichen Modell $y = b_1 + b_2 x + e$ vergleichen.

Für die OLS Schätzfunktionen der Koeffizienten sind

$$b_2^* = \frac{\sum \ddot{x}_i^* \ddot{y}_i^*}{\sum \ddot{x}_i^{*2}} = \frac{\sum (c\,\ddot{x}_i)(d\,\ddot{y}_i)}{\sum (c\,\ddot{x}_i)^2} = \frac{dc \sum \ddot{x}_i \ddot{y}_i}{c^2 \sum \ddot{x}_i^2} = \frac{d \sum \ddot{x}_i \ddot{y}_i}{c \sum \ddot{x}_i^2} = \frac{d}{c}\, b_2$$

$$b_1^* = \bar{y}^* - b_2^* \bar{x}^* = d\bar{y} - b_2^* c\bar{x} = d\bar{y} - \left(\frac{d}{c}\, b_2\right) c\bar{x} = db_1$$

mit $\ddot{y}_i := y_i - \bar{y}$ und $\ddot{x}_i := x_i - \bar{x}$.

Ebenso einfach kann man erkennen, dass das Bestimmtheitsmaß durch eine Skalierung nicht beeinflusst wird

$$R^{*2} = 1 - \frac{\sum e^{*2}}{\sum \ddot{y}^{*2}} = 1 - \frac{\sum (de)^2}{\sum (d\ddot{y})^2} = 1 - \frac{d^2 \sum e^2}{d^2 \sum \ddot{y}^2} = R^2$$

Die Standardfehler ($\widehat{\text{se}}(b_h)$ mit $h = 1, 2$) werden wir zwar erst in einem späteren Kapitel diskutieren und berechnen, aber es sei gleich hier vorausgeschickt, dass die Standardfehler bei einer Skalierung von x oder y ebenso einfach angepasst werden können (die folgenden Formeln für die Standardfehler werden wir erst später herleiten)

$$s^{*2} = \frac{\sum e_i^{*2}}{n-2} = \frac{\sum (de_i)^2}{n-2} = d^2\, s^2$$

$$\widehat{\text{se}}(b_2^*) = \sqrt{\frac{s^{*2}}{\sum \ddot{x}_i^{*2}}} = \sqrt{\frac{d^2 s^2}{c^2 \sum \ddot{x}_i^2}} = \frac{d}{c}\, \widehat{\text{se}}(b_2)$$

$$\widehat{\text{se}}(b_1^*) = \sqrt{\frac{s^{*2} \sum x_i^{*2}}{n \sum \ddot{x}_i^{*2}}} = \sqrt{\frac{d^2 s^2\, c^2 \sum x_i^2}{n\, c^2 \sum \ddot{x}_i^2}} = d\, \widehat{\text{se}}(b_1)$$

Damit wurde gezeigt, dass für $y_i^* := dy_i$ und $x_i^* := cx_i$ gilt

$$\begin{aligned} b_1^* &= db_1 \\ b_2^* &= \frac{d}{c}\, b_2 \\ s^{2*} &= d^2\, s^2 \\ \widehat{\mathrm{se}}(b_1^*) &= d\, \widehat{\mathrm{se}}(b_1) \\ \widehat{\mathrm{se}}(b_2^*) &= \left(\frac{d}{c}\right) \widehat{\mathrm{se}}(b_2) \\ R^{2*} &= R^2 \end{aligned}$$

Dies gilt allgemeiner auch für das multiple Regressionsmodell.

Übung: Zeigen Sie für das bivariate Regressionsmodell, dass die Addition einer Konstanten zur abhängigen und/oder erklärenden Variable sich nur auf das Interzept auswirkt, aber keine Auswirkung auf den Steigungskoeffizienten hat.

Gilt dies auch für $y_i^* := d_1 + d_2 y_i$ und $x_i^* := c_1 + c_2 x_i$?

2.12.3 Standardisierte (Beta-) Koeffizienten

Wie wir gerade gesehen haben hängt der Wert der Regressionskoeffizienten sowie deren Standardfehler von den Maßeinheiten ab, in denen die Variablen gemessen wurden (aber nicht das R^2).

In manchen Anwendungen haben die Variablen keine natürlichen Dimensionen, z.B. sind bei psychologischen Tests die Einheiten oft willkürlich angenommen.

Deshalb werden in solchen Fällen manchmal die Variablen z-transformiert (standardisiert) bevor die Regression geschätzt wird; das heißt, der Mittelwert wird von allen Ausprägungen subtrahiert und die resultierenden Werte durch die Standardabweichung der Variable dividiert.

Durch diese Standardisierung ist die neue Einheiten die Standardabweichungen der Variable. Diese Standardisierung ist natürlich nur eine Skalierung, dies ändert also nichts an den Zusammenhängen, aber da nun alle Variablen auf der gleichen Skala gemessen sind können dadurch die Koeffizienten der einzelnen Variablen sinnvoll miteinander verglichen werden.

Solche Koeffizienten einer Regression mit z-transformierten Variablen werden 'Standardisierte Koeffizienten' oder 'Beta-Koeffizienten' genannt.

Wenn wir die Abweichungen von dem Stichprobenmittelwert wieder mit zwei Punkten über der Variable kennzeichnen erhalten wir die z-transformierten Variablen durch Division durch deren Standardabweichung

$$\ddot{y}_i^z := \frac{y_i - \bar{y}}{s_y} := \frac{\ddot{y}_i}{s_y}, \quad \ddot{x}_{i1}^z := \frac{\ddot{x}_{i1}}{s_{x_1}}, \quad \ddot{x}_{i2}^z := \frac{\ddot{x}_{i2}}{s_{x_2}}, \quad \ldots, \quad \ddot{x}_{ik}^z := \frac{\ddot{x}_{k,i}}{s_{x_k}}$$

Für das ursprüngliche Modell

$$y_i = b_1 + b_2 x_{i2} + b_3 x_{i3} + \ldots + b_k x_{ik} + e_i$$

erhalten wir nach der z-Transformation

$$\ddot{y}_i^z = b_2^z \ddot{x}_{i2}^z + b_3^z \ddot{x}_{i3}^z + \ldots + b_k^z \ddot{x}_{ik}^z + \tilde{e}_i$$

Man beachte, dass durch Subtraktion der Mittelwerte das Interzept wegfällt.

Die Unabhängigkeit von Maßeinheiten erlaubt nun einen unmittelbaren Vergleich der Koeffizienten b_h^z $(h = 1, 2, \ldots, k)$ untereinander, das heißt, wenn z.B. die Variable x_h ceteris paribus um *eine Standardabweichung* zunimmt, erwarten wir eine Änderung der abhängigen Variable y um b_h^z *Standardabweichungen*.

Da die meisten ökonomischen Variablen in gut interpretierbaren Einheiten gemessen werden und weil diese 'Beta-Koeffizienten' b_h^z den Einfluss einzelner erklärender Variablen auf die abhängige Variable y nicht besser isolieren können als die üblichen Koeffizienten b_h werden sie in der Ökonometrie eher selten verwendet.

2.12.4 Verdrehte Regression (*'Reverse Regression'*)

Wir haben bisher die Quadratsumme der Residuen der Gleichung $y_i = b_1 + b_2 x_i + e_i$ minimiert, also das Quadrat der vertikalen Abstände zwischen y_i und $\widehat{y}_i$, weil wir y mit Hilfe der x Variable 'erklären' wollen. In manchen Fällen ist die Wirkungsrichtung aber nicht klar, so können wir z.B. bei dem Zusammenhang zwischen Körpergröße x und Gewicht y in beide Richtungen argumentieren.

Ad hoc würden viele erwarten, dass es keine Rolle spielt ob wir y auf x oder x auf y regressieren, also

$$y_i = b_1 + b_2 x_i + e_i \qquad \longleftrightarrow \qquad x_i = b_1^* + b_2^* y_i + e_i^*$$

denn $y = b_1 + b_2 x + e$ kann natürlich umgeschrieben werden zu

$$x = -\frac{b_1}{b_2} + \frac{1}{b_2}\, y - \frac{1}{b_2}\, e$$

Man könnte irrtümlich vermuten, dass $b_1^* = -b_1/b_2$ und $b_2^* = 1/b_2$ sein sollte, aber dem ist nicht so! Die Umformungen sind natürlich korrekt, aber diese sind *nicht* die OLS Schätzer.

Die OLS Schätzer der 'verdrehten' Regression sind

$$b_2^* = \frac{\mathrm{cov}(x, y)}{\mathrm{var}(y)}, \qquad b_1^* = \bar{x} - b_2^* \bar{y}$$

Abbildung 2.44 zeigt, dass im Fall der verdrehten Regression die Quadratsummen der horizontalen Abstände minimiert werden. Zu Vergleichszwecken ist auch die direkte Regression $\widehat{y}_i = b_1 + b_2 x_i$ strichliert eingezeichnet.

2.12.5 Historisches

Die tatsächlichen Ursprünge der OLS Methode sind bis heute nicht restlos geklärt. Sicher ist nur, dass sie zuerst für astronomische Anwendungen entwickelt wurde, und zwar um aus einer Reihe ungenauer Messungen das wahrscheinlichste Ergebnis für eine neue Messung zu berechnen, und dass sie erstmals 1805 vom französischen Mathematiker Adrien-Marie Legendre (1752-1833) im Anhang eines Werkes zur Berechnung von Kometenbahnen[31] publiziert wurde. Legendre suchte nach einer

[31] "Nouvelles méthodes pour la détermination des orbites des comètes." Paris 1805, Anhang: "Sur la Méthode des moindres quarrés", S. 72-80 .

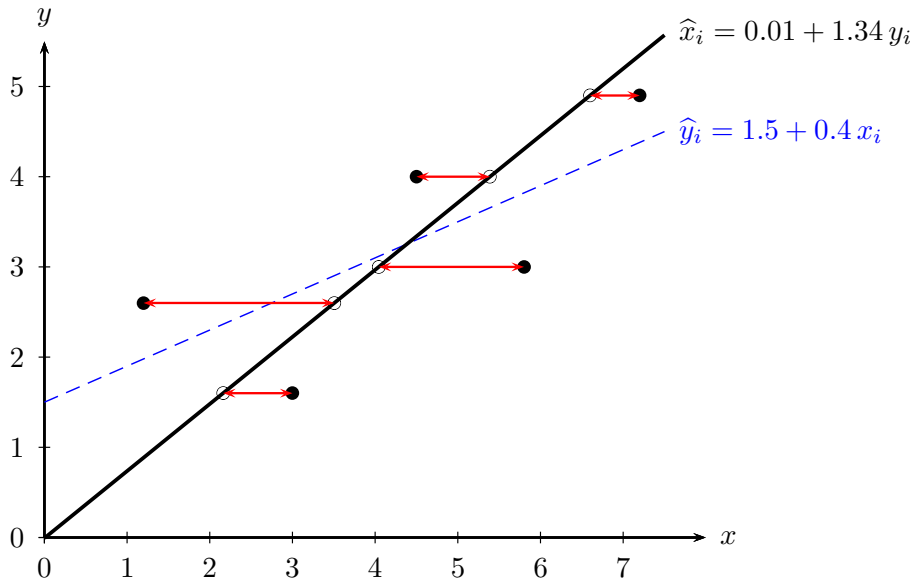


Abbildung 2.44: *'Reverse Regression'*: die Regression von x auf y ($\hat{x}_i = b_1^* + b_2{}^* y_i$), sowie strichliert die normale Regression $\widehat{y}_i = b_1 + b_2 x_i$.

Methode, wie ein Gleichungssystem mit mehr Gleichungen als Unbekannten gelöst werden könnte, und zeigte, dass die 'Methode der Kleinsten Quadrate' (*"Méthode des moindres quarrés"*) zu einem Gleichungssystem führt, das mit *'gewöhnlichen'* Methoden gelöst werden kann, daher die Bezeichnung OLS (*'Ordinary Least Squares'*).

Es gilt aber als sehr wahrscheinlich, dass Carl Friedrich Gauss (1777-1855) die Grundlagen der OLS Methode bereits 1795 im Alter von 18 Jahren entwickelte. Vermutlich trug die Anwendung dieser Methode auch wesentlich zum frühen Ruhm von Gauss bei, denn sie erlaubte es ihm 1801 aus einer Reihe fehlerbehafteter Messungen ziemlich genau den Ort zu berechnen, an dem der kurz vorher entdeckte Zwergplanet Ceres wieder hinter der Sonne hervorkommen würde. Als Gauss die Methode 1809 schließlich publizierte nahm er die Entdeckung der OLS Methode für sich in Anspruch, was zu einem Streit über die Urheberschaft zwischen Gauss und dem um 25 Jahre älteren Legendre führte (vgl. **?**).

Die Bezeichnung *Regression* ist deutlich jünger und geht auf Francis Galton (1822 – 1911) zurück, einen Cousin von Charles Darwin. Galton war wie viele seiner Zeitgenossen – und insbesondere auch viele der frühen Pioniere der Statistik – besorgt, dass die Verbreitung negativ bewerteter Erbanlagen Großbritannien langfristig große Probleme bereiten würde, und wurde so zu einem Begründer der *Eugenik*, die nach Möglichkeiten suchte, den Anteil positiv bewerteter Erbanlagen zu vergrößern. Galton fand, dass in einer Regression der Körpergröße von Kindern auf die Körpergröße der Eltern der Regressionskoeffizient durchgehend kleiner als Eins war, dass also überdurchschnittlich große Eltern tendenziell kleinere Kinder, und überdurchschnittlich kleine Eltern tendenziell größere Kinder hatten. **?** nannte dies "*Regression towards Mediocrity in Hereditary Status*".[32] Die der Analyse zugrunde liegende statistische Technik wurde in der Folge als 'Regression' bekannt.

[32] siehe `http://galton.org/essays/1880-1889/galton-1886-jaigi-regression-stature.pdf`

Vor dem Hintergrund des Burenkrieges (1899 – 1902), der einen Mangel an tauglichen Rekruten zutage brachte, wurden die Ergebnisse von Galton von den damaligen Eliten mit einiger Sorge zur Kenntnis genommen. Es führte zu Ängsten vor Degeneration und langfristigem Niedergang der imperialen Größe.

Es zeigte sich allerdings, dass Galtons Sorgen unbegründet waren, ein Regressionskoeffizient kleiner Eins ist durchaus mit einer über die Zeit stabilen Verteilung der Körpergrößen kompatibel.[33] Deshalb ging dieses Phänomen als *"Galton's Fallacy"* in die Literatur ein.

[33]Stellen Sie sich drei Personen vor mit der Körpergröße 160, 180 und 200cm. Angenommen, jeder dieser Personen hätte wieder drei Kinder, eines um 10% kleiner, eines gleich groß, und eines 10% größer. Dann wäre schon in der zweiten Generation das kleinste Kind vom 160cm Vater nur noch 144cm groß, das größte Kind vom 200cm Vater bereits 220cm. Über wenige Generationen hätten die kleinsten Personen die Größe von Ameisen, und die größten Personen wären wahre Monster! Eine solche Verteilung wäre über die Zeit offensichtlich nicht stabil!

Anhang A

Die wichtigsten statistischen Kennzahlen und deren Eigenschaften

Einfache statistische Kennzahlen, wie vor allem das arithmetische Mittel und Varianzen, spielen in der Ökonometrie eine zentrale Rolle. So kann die einfache Regressionsanalyse einfach als eine Erweiterung der Berechnung einfacher Mittelwerte auf bedingte Mittelwerte verstanden werden. Deshalb ist ein gutes Verständnis dieser einfachen Kennzahlen unabdingbare Voraussetzung für das Verständnis der Regressionsanalyse, und wie wir später sehen werden, für das Verständnis komplexerer Konzepte, wie z.B. bedingter Erwartungswerte oder des Gesetzes iterativer Erwartungen.

A.1 Arithmetisches Mittel

Die Definition des arithmetischen Mittels ist altbekannt

$$\overline{x} := \sum_{i=1}^{n} x_i \qquad \text{mit } i = 1, \ldots, n$$

Besitzen einige der n Beobachtungen den gleichen numerischen Wert können diese zusammengefasst werden

$$\bar{x} = \frac{1}{n}\left(\underbrace{x_1 + \cdots + x_1}_{n_1\text{-mal}} + \underbrace{x_2 + \cdots + x_2}_{n_2\text{-mal}} + \cdots + \underbrace{x_k + \cdots + x_k}_{n_k\text{-mal}}\right)$$

mit Häufigkeiten $n_1, n_2, \ldots, n_k$

$$\bar{x} = \frac{1}{n}\left(x_1 n_1 + \cdots + x_k n_k\right) = \frac{1}{n}\sum_{j=1}^{k} x_j n_j = \sum_{j=1}^{k} x_j \frac{n_j}{n}$$

mit $\sum_{j=1}^{k} n_j = n$ wobei k die *unterschiedlichen* Merkmalsausprägungen bezeichnet (wenn alle Merkmalsausprägungen unterschiedlich sind ist $k = n$).

Wenn wir die *relativen Häufigkeiten* (Anteile) n_j/n mit f_j (für *frequencies*) bezeichnen können wir schreiben

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{k} x_j n_j = \sum_{j=1}^{k} x_j \frac{n_j}{n} = \quad \text{mit } j = 1, \ldots, k$$

bzw. mit $f_j := \frac{n_j}{n}$ (*relative* Häufigkeiten)

$$\boxed{\bar{x} = \sum_{j=1}^{k} x_j f_j \qquad \text{mit } j = 1, \ldots, k}$$

Diese Schreibweise verdeutlicht, warum man vom *gewogenen* arithmetischen Mittel spricht (die relativen Häufigkeiten f_j sind die Gewichte). Außerdem wird uns diese Schreibweise später in der Stochastik bei den Diskussion von Erwartungswerten wieder begegnen.

A.1.1 4 Eigenschaften des arithmetischen Mittels

1. **Schwerpunkteigenschaft** Die Summe der Abweichungen der Einzelwerte vom arithmetischen Mittel $\bar{x}$ ist immer gleich Null:

   $$\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_i x_i - \sum_i \bar{x} = n\bar{x} - n\bar{x} = 0$$

   Warum?

   Aus $\bar{x} := \frac{1}{n}\sum_i x_i$ folgt $\sum_i x_i = n\bar{x}$, und $\sum_i \bar{x} = n\bar{x}$

   Deshalb kann man sich das arithmetische Mittel als *Schwerpunkt einer Verteilung* vorstellen.

2. Die Summe der quadrierten Abweichungen von $\bar{x}$ ist kleiner als von jedem beliebigen anderen Wert z

   $$\sum_{i=1}^{n}(x_i - \bar{x})^2 < \sum_{i=1}^{n}(x_i - z)^2 \quad \text{für } \bar{x} \neq z$$

   Warum? Wir subtrahieren und addieren $\bar{x}$, wodurch sich der Wert nicht verändert, und verwenden die binomische Formel $(a+b)^2 = a^2 + 2ab + b^2$

   $$\begin{aligned}
   \sum_i (x_i - z)^2 &= \sum_i (x_i - \bar{x} + \bar{x} - z)^2 = \sum_i \left[(x_i - \bar{x}) + (\bar{x} - z)\right]^2 \\
   &= \sum_i (x_i - \bar{x})^2 + 2(\bar{x} - z)\underbrace{\sum_i (x_i - \bar{x})}_{=0} + \sum_i (\bar{x} - z)^2 \\
   &= \sum_i (x_i - \bar{x})^2 + \sum_i (\bar{x} - z)^2
   \end{aligned}$$

   Da quadratische Ausdrücke nie negativ sein können und $\bar{x} \neq z$ folgt $\sum_i (\bar{x} - z)^2 > 0$. Deshalb muss $\sum_i (x_i - z)^2 > \sum_i (x_i - \bar{x})^2$ sein!

Da die Varianz definiert ist als $\text{var}(x) := \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ folgt daraus auch, dass der Mittelwert die Varianz 'minimiert' (wir werden später sehen, dass das *Gauss-Markov Theorem* in gewisser Weise als Verallgemeinerung dieser einfachen Eigenschaft gesehen werden kann).

3. **Translationsäquivarianz:** Lineare Transformationen spielen bei der Umrechnung von Währungen oder physikalischen Einheiten (z.B. zwischen Kilometer und Meilen oder Celsius und Fahrenheit) eine wichtige Rolle.

   Das arithmetische Mittel vollzieht solche lineare Transformationen nach, das heißt, ist der Mittelwert einer Variable bekannt, kann dieser unmittelbar in eine andere Einheit umgerechnet werden.

   Warum?

   Werden die Einzelwerte einer Variablen linear transformiert $x_i^* = a_1 + a_2 x_i$ gilt

   $$\bar{x}^* = a_1 + a_2\bar{x}$$

   folgt

   $$\begin{aligned}\bar{x}^* &= \frac{1}{n}\sum_i (a_1 + a_2 x_i) \\ &= \frac{1}{n}\left(na_1 + a_2\sum_i x_i\right) \\ &= a_1 + a_2\frac{1}{n}\sum_i x_i = a_1 + a_2\bar{x}\end{aligned}$$

4. **Gewichtetes arithmetisches Mittel über Mittelwerte von Teilgruppen:** Falls die arithmetischen Mittel mehrerer disjunkter Teilgruppen sowie deren Umfang bekannt ist, kann aus den Mittelwerten der Teilgruppen das arithmetische Mittel der Gesamtheit (*'grand mean'*) berechnet werden.

   Das *'grand mean'* ist einfach die mit den Anteilen (d.h. relativen Häufigkeiten) gewichtete Mittel der Mittelwerte der Teilgruppen.

   Wenn zum Beispiel von Teilnehmern eines Kurses nur die durchschnittliche Körpergröße von Frauen und Männern sowie deren Anzahl bekannt ist, kann daraus die durchschnittliche Körpergröße aller Teilnehmer berechnet werden indem die Einzelmittelwerte mit den Anteil von Frauen bzw. Männern gewichtet und addiert werden

   $$\bar{\bar{x}} = \sum_{j=1}^{k}\frac{n_j}{n}\bar{x}_j$$

   wobei k die Anzahl der Teilgruppen und n_j die Anzahl der Beobachtungen von Teilgruppe j bezeichnet.

   Warum?

   Für zwei Teilgruppen: sei $n_1, \bar{x}_1, n_2, \bar{x}_2$ mit $n_1 + n_2 = n$ bekannt. Wir summieren zuerst über beide Gruppen getrennt und multiplizieren mit n_1/n_1, bzw.

$n_2/n_2 = 1$

$$\begin{aligned}
\bar{x} &= \frac{1}{n}\left(\sum_{i=1}^{n_1} x_{1i} + \sum_{i=1}^{n_2} x_{2i}\right) \\
&= \frac{1}{n}\left(\frac{n_1}{n_1}\sum_{i=1}^{n_1} x_{1i} + \frac{n_2}{n_2}\sum_{i=1}^{n_2} x_{2i}\right) \\
&= \frac{n_1}{n}\left(\frac{1}{n_1}\sum_{i=1}^{n_1} x_{1i}\right) + \frac{n_2}{n}\left(\frac{1}{n_2}\sum_{i=1}^{n_2} x_{2i}\right) = \frac{n_1}{n}\,\bar{x}_1 + \frac{n_2}{n}\,\bar{x}_2
\end{aligned}$$

Also: die Summe der mit den Anteilen gewichteten Teilmittelwerte gibt den Gesamtmittelwert (*'grand mean'*). Das gilt selbstverständlich auch für eine beliebige Anzahl von Teilgruppen.

Eine analoge Eigenschaft in der Stochastik bildet das Gesetz der iterativen Erwartungen.

A.2 Varianzen

Die Varianz s^2 ist ein Streuungsmaß um den Mittelwert und ist definiert als die mittlere quadratische Abweichung vom arithmetischen Mittel $\bar{x}$.

$$\begin{aligned}
\mathrm{var}(x) := s^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2\right) - \bar{x}^2 := \overline{x^2} - \bar{x}^2
\end{aligned}$$

Wie das arithmetische Mittel ist die Varianz nur für metrisch skalierte Variablen sinnvoll interpretierbar.

Warum gilt die zweite Zeile obiger Definition??

$$\begin{aligned}
\mathrm{var}(x) &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \frac{1}{n}\sum_{i}\left(x_i^2 - 2\bar{x}x_i + \bar{x}^2\right) \\
&= \frac{1}{n}\left(\sum_{i} x_i^2 - 2\bar{x}\sum_{i} x_i + \sum_{i}\bar{x}^2\right) \\
&= \frac{1}{n}\left(\sum_{i} x_i^2 - 2\bar{x}n\bar{x} + n\bar{x}^2\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2\right) - \bar{x}^2 := \overline{x^2} - \bar{x}^2
\end{aligned}$$

weil $\sum_i x_i = n\bar{x}$ und $\sum_i \bar{x}^2 = n\bar{x}^2$
Diese zweite Form ist für manche Berechnungen deutlich anwendungsfreundlicher als die ursprüngliche Definition.

Varianz linear transformierter Daten

Sei $x_i^* = a_1 + a_2 x_i$ für $i = 1, \ldots, n$

$$\text{var}(x^*) = \frac{1}{n}\sum_{i=1}^{n}(x_i^* - \bar{x}^*)^2 = a_2^2\,\text{var}(x_i)$$

Warum? Wir haben bereits früher gezeigt, dass $\bar{x}^* = a_1 + a_2\bar{x}$.

$$\begin{aligned}\text{var}(x^*) &= \frac{1}{n}\sum_i (a_1 + a_2 x_i - a_1 - a_2\bar{x})^2 \\ &= \frac{1}{n}\sum_i (a_2[x_i - \bar{x}])^2 \\ &= a_2^2\frac{1}{n}\sum_i (x_i - \bar{x})^2 = a_2^2\,\text{var}(x_i)\end{aligned}$$

Daraus folgt, dass die Addition oder Subtraktion einer Konstante keinen Einfluss auf die Varianz hat, aber die Multiplikation mit einer einer Konstanten wirkt sich quadratisch auf die Varianz aus!

Zwei Arten der Varianz

In der Literatur wird unterschieden zwischen

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \text{versus} \quad s_s^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Die meisten Programme (wie z.B. R) berechnen die Varianz nach der 2. Formel, d.h. sie verwenden den Vorfaktor $1/(n-1)$. Die Verwendung des Vorfaktor $1/(n-1)$ statt $1/n$ ist angebracht, wenn die Varianz aus einer Stichprobe berechnet wird und als *Schätzung* für die Varianz der Grundgesamtheit dient. Der Grund dafür liegt im Konzept der später diskutierten *Erwartungstreue*.
Geht es hingegen nur um die Beschreibung gegebener Daten im Sinne der deskriptiven Statistik, dann ist die linke Formel (Vorfaktor $1/n$) angebracht.

A.2.1 Standardabweichung

Die Varianz ist manchmal schwierig zu interpretieren, wenn z.B. x in Euro gemessen wird, hat die Varianz die Dimension Euro^2.
Die *Standardabweichung* hat gegenüber der Varianz den Vorteil, dass sie in der gleichen Einheit wie die Beobachtungswerte gemessen wird.
Definition der **Standardabweichung**:

$$s = \sqrt{\text{var}(x)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

A.3 Zusammenhangsmaße für metrisch skalierte Merkmale

Das wichtigste Zusammenhangsmaß für metrisch skalierte Merkmale ist die empirische Kovarianz.

A.3.1 Kovarianz

Die Kovarianz ist eine (nicht standardisierte) Maßzahl für den Zusammenhang zwischen zwei metrisch skalierten statistischen Merkmalen x und y.

$$\operatorname{cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

mit $\bar{x} := \frac{1}{n}\sum_i x_i$ und $\bar{y} := \frac{1}{n}\sum_i y_i$

Kovarianz:

- Die Kovarianz ist positiv, wenn x und y tendenziell einen gleichgerichteten linearen Zusammenhang aufweisen, d.h. hohe Werte von x gehen mit hohen Werten von y einher und niedrige mit niedrigen.
- Die Kovarianz ist negativ, wenn x und y einen gegengerichteten linearen Zusammenhang aufweisen.
- Ist die Kovarianz Null, so besteht kein *linearer Zusammenhang* (es kann aber trotzdem oder ein nicht-linearer Zusammenhang bestehen, z.B. U-förmig).

Fünf Eigenschaften der empirischen Kovarianz

1) Symmetrie:

$$\boxed{\operatorname{cov}(x,y) = \operatorname{cov}(y,x)}$$

Warum?

$$\begin{aligned}\operatorname{cov}(x,y) &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \\ &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) = \operatorname{cov}(y,x)\end{aligned}$$

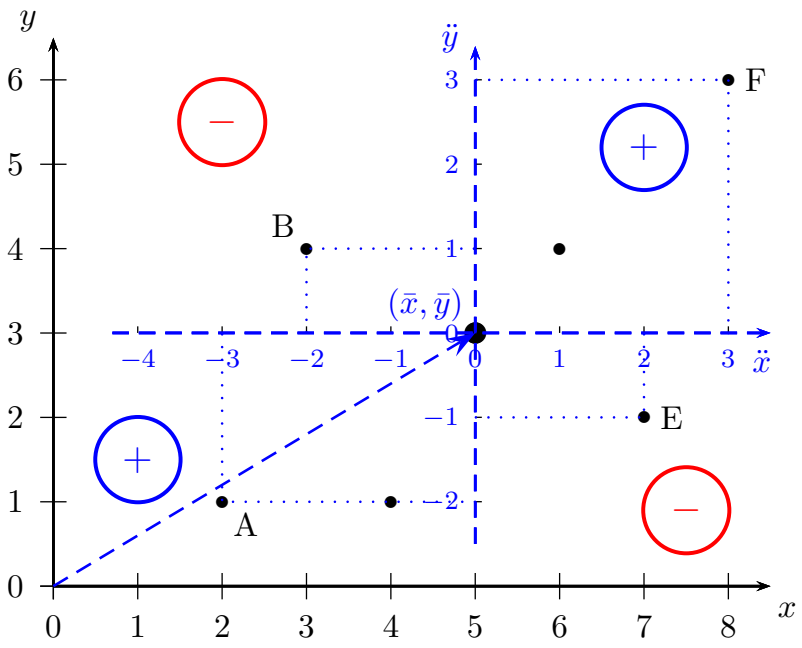
2) Konstante Faktoren können ausgeklammert werden: für $x, y \in \mathbb{R}^n$ und Zahlen $a, b \in \mathbb{R}$

$$\boxed{\operatorname{cov}(a_1 + a_2 x, b_1 + b_2 y) = a_2 b_2 \operatorname{cov}(x,y)}$$

Beispiel: Vorzeichen der Kovarianz:

	x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$
	2	1	-3	-2	6
	3	4	-2	1	-2
	4	1	-1	-2	2
	6	4	1	1	1
	7	2	2	-1	-2
	8	6	3	3	9
$\sum$	30	18	0	0	14

$$\text{cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \frac{14}{6} = 2.33$$



Warum?

$$\begin{aligned}
\text{cov}(a_1+a_2x, b_1+b_2y) &= \frac{1}{n}\sum_{i=1}^{n}\left[a_1+a_2x_i-(\overline{a_1+a_2x_i})\right]\left[b_1+b_2x_i-(\overline{b_1+b_2x_i})\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}(a_2x_i-a_2\bar{x})(b_2y_i-b_2\bar{y}) \\
&= \frac{1}{n}\sum_{i=1}^{n}a_2(x_i-\bar{x})b_2(y_i-\bar{y}) \\
&= a_2b_2\,\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) \\
&= a_2b_2\,\text{cov}(x,y)
\end{aligned}$$

3) Additivität: für $x, y, z \in \mathbb{R}^n$

$$\boxed{\text{cov}[x, (y+z)] = \text{cov}(x, y) + \text{cov}(x, z)}$$

Warum?

$$\begin{aligned}
\text{cov}[x, (y+z)] &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})[(y_i + z_i) - (\overline{y+z})] \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})[(y_i + z_i) - (\bar{y} + \bar{z})] \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})[(y_i - \bar{y}) + (z_i - \bar{z})] \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) + \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(z_i - \bar{z}) \\
&= \text{cov}(x, y) + \text{cov}(x, z)
\end{aligned}$$

4) Zusammenhang mit empirischer Varianz: für $x \in \mathbb{R}^n$

$$\boxed{\text{cov}(x, x) = \text{var}(x)}$$

Warum?

$$\begin{aligned}
\text{cov}(x, x) &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x}) \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \text{var}(x)
\end{aligned}$$

5) Empirische Varianz einer Summe: für $x, y \in \mathbb{R}^n$

$$\boxed{\text{var}(x+y) = \text{var}(x) + \text{var}(y) + 2\,\text{cov}(x, y)}$$

Warum?

$$\begin{aligned}
\text{var}(x+y) &= \frac{1}{n}\sum_{i=1}^{n}[(x_i + y_i) - (\bar{x} + \bar{y})]^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}[(x_i - \bar{x}) + (y_i - \bar{y})]^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 + \\
&\quad +2\,\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \\
&= \text{var}(x) + \text{var}(y) + 2\,\text{cov}(x, y)
\end{aligned}$$

Zwei Arten der Kovarianz

$$1) \quad \text{cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

versus

$$2) \quad \text{cov}_s(x,y) = \frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Die meisten Programme (wie z.B. R) berechnen die Kovarianz nach der 2. Formel, d.h. sie verwenden den Vorfaktor $1/(n-1)$

Die Anwendung des Vorfaktor $1/(n-1)$ statt $1/n$ ist angebracht, wenn die Kovarianz aus einer Stichprobe berechnet wird und als *Schätzung* für die Kovarianz der Grundgesamtheit dient.

A.3.2 Korrelationskoeffizient nach Bravais-Pearson

Kovarianzen haben einen Nachteil, sie hängen von den Maßeinheiten ab, in denen die Variablen gemessen werden! Um einen Zusammenhang vergleichbar zu machen, muss die Kovarianz normiert werden. Dies führt zu Korrelationskoeffizienten.

Generell versteht man unter Korrelationen eine Gruppe von statistischen Kennwerten, die den *"Zusammenhang"* zwischen zwei Variablen messen sollen. Der bekannteste Korrelationskoeffizient für metrisch skalierte Variablen ist der Korrelationskoeffizient nach Bravais-Pearson.

Korrelationskoeffizient nach Bravais-Pearson: Der Korrelationskoeffizient r ist ein dimensionsloses Maß für den Grad des linearen Zusammenhangs zwischen zwei *mindestens intervallskalierten* Merkmalen.

$$\boxed{\text{corr}(x,y) := r_{xy} = \frac{\text{cov}(x,y)}{\sqrt{\text{var}(x)}\sqrt{\text{var}(y)}}}$$

Eigenschaften des Korrelationskoeffizient nach Bravais-Pearson: für Datenvektoren $x, y \in \mathbb{R}^n$ und Zahlen $a, b, c, d \in \mathbb{R}$ gilt

1. $r_{x,y}$ kann nur Werte zwischen -1 und $+1$ annehmen

$$-1 \leq \text{corr}(x,y) \leq +1$$

2. $r_{x,y}$ ändert sich nicht bei einer linearen Transformation

$$\text{corr}(ax+b, cy+d) = \text{corr}(x,y)$$

3. Wenn der $\text{corr}(x,y) = 0$ sind die beiden Merkmale linear unabhängig (sie können aber trotzdem nicht-linear abhängig sein); wenn $|\,\text{corr}(x,y)| = 1$ sind die Merkmale exakt linear abhängig
   - $\text{corr}(x,y) = +1$ wenn $y = a + bx$
   - $\text{corr}(x,y) = -1$ wenn $y = a - bx$

Beispiel: mit $\ddot{x} := x - \bar{x}$, $\ddot{y} := y - \bar{y}$

	x	y	$\ddot{x}$	$\ddot{x}^2$	$\ddot{y}$	$\ddot{y}^2$	$\ddot{x}\ddot{y}$
	2	1	-3	9	-2	4	6
	3	4	-2	4	1	1	-2
	4	1	-1	1	-2	4	2
	6	4	1	1	1	1	1
	7	2	2	4	-1	1	-2
	8	6	3	9	3	9	9
$\sum$	30	18	0	28	0	20	14

$$r_{xy} = \frac{\mathrm{cov}(x,y)}{\sqrt{\mathrm{var}(x)}\sqrt{\mathrm{var}(y)}} = \frac{14}{\sqrt{28 \cdot 20}} = 0.591608$$

Übung: Zeigen Sie, dass der Korrelationskoeffizient

$$\begin{aligned} r &= \frac{\mathrm{cov}(x,y)}{\sqrt{\mathrm{var}(x)\,\mathrm{var}(y)}} \\ &= \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y}))}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \end{aligned}$$

alternativ berechnet werden kann als

$$r = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\sum_i x_i^2 - n\bar{x})(\sum_i y_i^2 - n\bar{y}^2)}}$$

Anhang B

Berechnung von durchschnittlichen Wachstumsraten

Durchschnittliche Wachstumsraten spielen in vielen Bereichen der Wirtschaftswissenschaften eine wichtige Rolle, deshalb wiederholen wir hier einige zentrale Konzepte. Die Wachstumsraten hauptsächlich auf Zeitreihen angewandt werden verwenden wir t als Beobachtungsindex, mit $t = 0, 1, 2, \ldots, T$.

B.1 Diskrete Wachstumsraten (i)

Unter der diskreten Wachstumsrate verstehen wir die relative Änderung einer Größe zwischen zwei Perioden

$$i = \frac{y_t - y_{t-1}}{y_{t-1}} := \frac{\Delta y_t}{y_{t-1}} = \frac{y_t}{y_{t-1}} - 1$$

Der Quotient y_t/y_{t-1} wird auch als Wachstumsfaktor bezeichnet.

Wenn eine Variable y mit einer konstanten diskreten Wachstumsrate i wächst nimmt sie in jeder Periode um iy_{t-1} Einheiten zu. Für $t = 1, 2, \ldots, T$

$$\begin{aligned} y_1 &= y_0 + iy_0 = y_0(1+i) \\ y_2 &= y_1(1+i) = y_0(1+i)(1+i) = y_0(1+i)^2 \\ &\vdots \\ y_T &= y_0(1+i)^T \end{aligned}$$

Sollte die durchschnittliche Wachstumsrate zwischen den Perioden 0 und T berechnet werden, so darf dazu *nicht* das arithmetische Mittel herangezogen werden!

Man kann sich einfach fragen, welche Wachstumsrate i führt vom Wert y_0 zu y_t. Dazu logarithmieren wir $y_T = y_0(1+i)^T$ und lösen nach i

$$\begin{aligned} \ln y_T &= \ln y_0 + T \ln(1+i) \\ \frac{1}{T}(\ln y_T - \ln y_0) &= \ln(1+i) \\ \frac{1}{T} \ln \left[\frac{y_T}{y_0}\right] &= \ln(1+i) \end{aligned}$$

$$\boxed{i = \exp\left(\frac{1}{T}\ln\left[\frac{y_T}{y_0}\right]\right) - 1}$$

T bezeichnet dabei die Anzahl der Perioden; sollte z.B. die durchschnittliche Wachstumsrate des BIP von 2005 – 2010 berechnet werden, und wird darunter die Periode vom 1.1.2005 – 31.12.2010 verstanden, also $T = 6$.

Um eine prozentuelle Wachstumsrate zu erhalten muss i mit 100 multipliziert werden.

Beispiel: Angenommen ein Wert hat zwischen 1990 und 2012 von 500 auf 2000 zugenommen, wie groß ist die durchschnittliche jährliche Wachstumsrate?

$$i = \exp\left(\frac{1}{23}\ln\left[\frac{2000}{500}\right]\right) = 0.062127177$$

die diskrete prozentuelle Wachstumsrate beträgt also ca. 6.2%. Wir können dies einfach überprüfen

$$500(1 + 0.062127177)^{23} = 2000$$

B.2 Stetiges Wachstum (r)

Der enge Zusammenhang zwischen Wachstumsraten und der Exponentialfunktion (bzw. dem Logarithmus) wird sofort klar, wenn man mehrere Verzinsungen pro Periode zulässt, und die Anzahl der Verzinsungen pro Periode gegen Unendlich gehen lässt.

Wenn die Verzinsung m Mal pro Periode erfolgt

$$\begin{aligned} y_t &= y_0\left(1 + \frac{r}{m}\right)^{mt} = y_0\left[\left(1 + \frac{r}{m}\right)^{m/r}\right]^{rt} \\ &= y_0\left[\left(1 + \frac{1}{w}\right)^{w}\right]^{rt} \qquad \text{mit } w := \frac{m}{r} \end{aligned}$$

Wir erinnern uns, dass die Eulersche Zahl als Grenzwert einer Folge dargestellt werden kann

$$\lim_{w\to\infty}\left(1 + \frac{1}{w}\right)^{w} = \mathrm{e} \approx 2.7182819$$

man beachte, dass $m \to \infty$ impliziert $w \to \infty$.

Also

$$\boxed{y_t = y_0\,\mathrm{e}^{rt} := y_0\exp(rt)}$$

Man beachte, dass in $y_t = y_0(1 + r/m)^{mt}$ die Zeit noch eine diskrete Variable ist, erst durch die Grenzwertbildung wird die Zeit zu einer stetigen Variable. Außerdem wird hier r als konstant angenommen, aber natürlich müssen tatsächliche Wachstumsraten nicht konstant sein.

Mit einer stetigen Wachstumsrate kann die Veränderung über die Zeit einfach als Ableitung nach der Zeit berechnet werden

$$\frac{dy_t}{dt} = \frac{d(y_0\mathrm{e}^{rt})}{dt} = ry_0\mathrm{e}^{rt}$$

Die relative Änderung ist deshalb

$$\frac{\frac{dy_t}{dt}}{y_t} = \frac{ry_0 e^{rt}}{y_0 e^{rt}} = r$$

Für sehr kurze Zeitperioden konvergiert die diskrete Wachstumsrate i gegen die stetige Wachstumsrate r

$$\begin{aligned} \lim_{\Delta t \to 0} i &= \lim_{\Delta t \to 0} \left\{ \frac{y_{t+\Delta t} - y_t}{\Delta t} \cdot \frac{1}{y_t} \right\} \\ &= \frac{dy}{dt} \frac{1}{y} = r \end{aligned}$$

Wenn eine Variable y exponentiell mit konstanter Wachstumsrate r wächst gilt $y_t = y_0 e^{rt}$, bzw. $\ln y_t = \ln y_0 + rt$. Daraus folgt durch Umschreiben

$$r = \frac{\ln y_t - \ln y_0}{t} = \frac{\Delta \ln y}{t}$$

Daraus folgt, dass die stetige Wachstumsrate zwischen zwei aneinandergrenzenden Perioden, d.h. wenn $\Delta t = 1$, einfach die logarithmische Differenz ist $r_{t,t+1} = \Delta \ln y_t$. Ebenso ist aus $\ln y_t = \ln y_0 + rt$ auch einfach ersichtlich, dass

$$\frac{d \ln y_t}{dt} = r$$

wenn y_0 konstant ist.

Diese Zusammenhänge werden in der empirischen Wirtschaftsforschung ausgiebig genützt, auch weil bei 'kleinen' Wachstumsraten (z.B. $r < 0.05$) die stetigen Wachstumsraten eine gute Annäherung an die diskreten Wachstumsraten darstellen.

B.3 Umrechnen zwischen stetigen und diskreten Wachstumsraten

Da zu jedem Zeitpunkt für eine beliebige diskrete Wachstumsrate i genau eine stetige Wachstumsrate r existiert, die zum gleichen Betrag y_t führt, kann einfach zwischen diskreten und stetigen Wachstumsraten umgerechnet werden

$$y_0(1+i)^t = y_0 \exp(rt)$$

folgt $\ln y_0 + t \ln(1+i) = \ln y_0 + rt$ bzw.

$$\boxed{r = \ln(1+i) \qquad \text{bzw.} \qquad i = \exp(r) - 1}$$

Wir haben vorhin gezeigt, dass die stetige Wachstumsrate zwischen zwei Perioden als logarithmische Differenz berechnet werden kann $r = \Delta \ln y / t$. Wenn man diese in eine diskrete Wachstumsrate umrechnet erhält man wieder $i = \exp(\Delta \ln y)/t - 1$.

Die prozentuellen Wachstumsraten erhält man wie üblich, indem man diese Wachstumsraten mit 100 multipliziert.

Anhang C

Beispiel Programme

C.1 Partielle Regression

R- und Stata-Programmcode zur Erzeugung von Abbildung 2.20 (Seite 56)
R:

```
rm(list=ls())
Auto <- read.csv2("http://www.hsto.info/econometrics/data/auto40.csv")

Auto$res_Preis <- resid(lm(Preis ~ Alter, data = Auto)) + mean(Auto$Preis)
Auto$res_km    <- resid(lm(km ~ Alter, data = Auto))    + mean(Auto$km)
# Grafik: partieller Scatterplot <- Alter
# x11(width = 400, height = 300)  ## nur Windows
par(mfrow=c(1,2),cex.main=0.85)
plot(Auto$km, Auto$Preis, main = "Bivariates Streudiagramm",
     xlab = "Kilometer", ylab = "Preis",
     xlim = c(0, max(Auto$km)), ylim = c(5000, max(Auto$Preis)))
abline(lm(Preis ~ km, data = Auto), lwd = 1.6, col = "blue")
points(mean(Auto$km), mean(Auto$Preis), pch = 22)
#
plot(Auto$res_km, Auto$res_Preis, main = "Partielles Streudiagramm",
     xlab = "Kilometer|Alter", ylab = "Preis|Alter",
     xlim = c(0, max(Auto$km)), ylim = c(5000, max(Auto$Preis)))
abline(lm(res_Preis ~ res_km, data = Auto), lwd = 1.6, col = "red")
points(mean(Auto$km), mean(Auto$Preis), pch = 22)
```

Stata:

```
clear all
insheet using http://www.hsto.info/econometrics/data/auto40.csv, delimiter(";")
destring alter, dpcomma replace  // Dezimalzeichen , durch . ersetzen
regress preis alter
predict res_preis, res
regress km alter
predict res_km, res
twoway (scatter preis km) (lfit preis km), ///
 title(Bivariate Regression) name(Graph1,replace) nodraw
```

```
twoway (scatter res_preis res_km) (lfit res_preis res_km), ///
 title(Partielle Regression) name(Graph2,replace) nodraw
graph combine Graph1 Graph2, cols(2)
```

C.2 R-Code für Tabelle 2.9 (Autopreise mit Alter-Dummies)

Tabelle 2.9 (Seite 71) wurde mit folgenden R-Code erzeugt:

```
# Autopreise, Alter auf Jahre gerundet
remove(list=ls())
d <- read.csv2("http://www.hsto.info/econometrics/data/auto40.csv")
attach(d)
AlterJ <- round(Alter,0)
AlterF <- as.factor(AlterJ)
eq1 <- lm(Preis ~ AlterJ)      # diskretes Alter
eq2 <- lm(Preis ~ AlterF)      # Dummies (Faktoren)
eq3 <- lm(Preis ~ AlterF - 1)  # Dummies ohne Interzept
library(stargazer)
stargazer(eq1,eq2,eq3, digits=2,intercept.bottom=FALSE,star.char="")
```

In Stata erhält man einen vergleichbaren Output mit[1]

```
clear all
insheet using http://www.hsto.info/econometrics/data/auto40.csv, ///
        delimiter(";")
destring alter, dpcomma replace
gen AlterJ = round(alter)

*** siehe http://repec.org/bocode/e/estout/index.html
*** ssc install estout, replace
eststo: regress preis AlterJ
eststo: regress preis i.AlterJ  // mit Dummies
eststo: regress preis i.AlterJ, nocons
esttab
```

C.2.1 Durchschnittliche jährliche Wachstumsrate von China

Das folgende kleine R Programm liest die aktuellen Daten mit Hilfe des `WDI` - packages direkt von der Weltbank ein und berechnet die durchschnittliche Wachstumsrate für den Zeitraum 1995 - 2019 (siehe Seite 107).

[1] Allerdings wird in Stata bei der Regression ohne Interzept (Option `nocons`) die erste Kategorie unterdrückt.

```
# Durchschnittliche Wachstumsrate des pro-Kopf Einkommens in China
# Daten: Weltbank, World Development Indicators (WDI)

rm(list = ls())
# install.packages(WDI)
# see e.g. https://cengel.github.io/gearup2016/worldbank.html
library(WDI)
WDIsearch(string = "GNI per capita, PPP", field = "name",
          short = TRUE) # search variable

# download data
china <- WDI(country = "CN",
             indicator = c("NY.GNP.PCAP.PP.KD"),
             start = 1995, end = 2019)
# data.frame china nach Jahr ansteigend sortieren
china <- china[order(china$year, decreasing = FALSE), ]

GDPpc <- china$NY.GNP.PCAP.PP.KD
Trend <- china$year

eq <- lm(log(GDPpc) ~ Trend)
b2 <- coef(eq)["Trend"]
WR <- (exp(b2) - 1)*100
# Ausgabe am Bildschirm
paste("Diskrete jährl. Wachtumsrate des chinesischen pro Kopf
      Einkommens von", min(Trend), "-", max(Trend), ":",
      round(WR ,2), "Prozent")  ## 8.81%
```

Das gleiche mit Stata und dem ado-File '`wbopendata`':

```
clear
* Einlesen von Weltbankdaten
* http://databank.worldbank.org/data/source/world-development-indicators
* NY.GNP.MKTP.PP.KD: Gross National Income, PPP (constant 2011 international $)

* ssc install wbopendata
wbopendata, indicator(NY.GNP.MKTP.PP.KD) country(CHN) clear long
rename ny_gnp_mktp_pp_kd GNI
drop if GNI == .
gen Trend = _n
gen lnGNI = log(GNI)

regress lnGNI Trend
quietly summarize year
display "Diskrete jährl. Wachstumsrate China " r(min) "-" r(max) ": "  ///
  round((exp(_b[Trend]) - 1)*100, 2) "%"
```

Lebenserwartung und pro Kopf Einkommen

Dieses Programm zur Erzeugung von Grafik 2.39 (Seite 110) ladet die Daten von den World Development Indicators (World Bank) und zeigt zwei Funktionsverläufe.

```
# WDI, Grafik: Life Exp. vs. log(GDP)
rm(list = ls())
# install.packages(WDI)
# see e.g. https://cengel.github.io/gearup2016/worldbank.html
library(WDI)

# download data
wdi_dat <- WDI(country = "all",
    indicator = c("NY.GNP.PCAP.PP.KD", "SP.DYN.LE00.IN"),
    start = 2018, end = 2018, extra = TRUE, cache = NULL)

# remove country aggregates
wdi_dat <- subset(wdi_dat, region != "Aggregates") # also removes NAs
# rename
names(wdi_dat)[names(wdi_dat) == "NY.GNP.PCAP.PP.KD"] <- "GDPpc"
names(wdi_dat)[names(wdi_dat) == "SP.DYN.LE00.IN"] <- "LifeExp"
eqlog <- lm(LifeExp ~ log(GDPpc), data = wdi_dat)

# Graph
x11(width=10,height=5) # only Windows, for Mac's: quartz()
par(mfrow = c(1,2)) # two graphs, 1 row, 2 columns
plot(wdi_dat$LifeExp ~ wdi_dat$GDPpc, col = "blue",
     ylab = "Life expectancy", xlab = "GDP per capita, PPP")
abline(lm(wdi_dat$LifeExp ~ wdi_dat$GDPpc), lwd = 2, col = "black")
curve(expr = (coef(eqlog)[1] + coef(eqlog)[2]*log(x)),
    add = TRUE, col = "red", lwd = 2, lty = 2)
legend("bottomright", legend = c("level-level", "level-log"),
    col=c("black", "red"), lty=1:2, cex=0.8)

plot(wdi_dat$LifeExp ~ log(wdi_dat$GDPpc), col = "blue",
     ylab = "Life expectancy", xlab = "log(GDP per capita), PPP")
abline(lm(wdi_dat$LifeExp ~ log(wdi_dat$GDPpc)), lwd = 2, col = "red")
```