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 gehend erlaubt uns die Regressionsanalyse 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, die uns erlaubt die Koeffizienten einer linearen Regression zu berechnen, und uns mit der Interpretation der Ergebnisse befassen, bevor wir die Technik auf mehr als zwei Variablen verallgemeinern und ein paar wichtige Spezialfälle untersuchen.

[0]

Wir werden später sehen, dass wir all dies als Voraussetzung für die stochastische Regressionsanalyse benötigen, die wir im nächsten Kapitel diskutieren werden.

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 der Index k die Anzahl der erklärenden 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 y und x 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.

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 mit dem 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 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

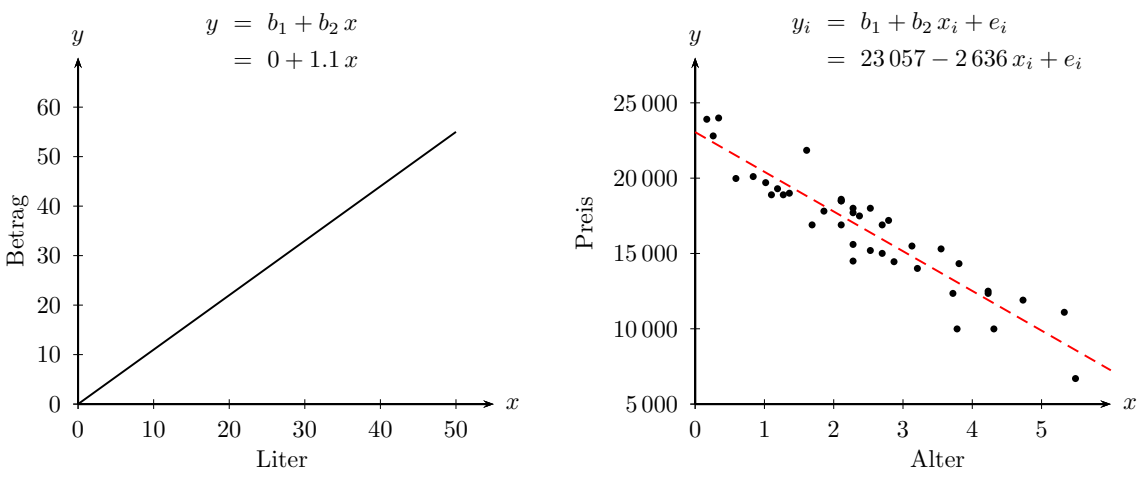


Abbildung 2.1: Linkes Panel: ein exakter Zusammenhang zwischen getankten Litern und zu bezahlendem Betrag für einen Preis $b_2 = 1.1$ Euro. Rechtes Panel: ein 'ungefährer' Zusammenhang zwischen dem Alter von Gebrauchtautos und deren Preis.

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 erkannbar, 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 '≈' Zeichens (*'ist ungefähr'*) schreiben $y \approx b_1 + b_2 x$, aber mit '≈' 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.

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

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	Obs.	Preis	Alter	AlterJ	km
1	10000	3.78	4	188000	21	15000	2.70	3	51500
2	21850	1.61	2	25900	22	18500	2.11	2	25880
3	14500	2.28	2	83300	23	18500	2.11	2	19230
4	11100	5.33	5	120300	24	12350	3.72	4	75000
5	6700	5.49	5	142000	25	16900	2.70	3	22000
6	24000	0.34	0	5500	26	18000	2.28	2	35000
7	10000	4.31	4	100500	27	18890	1.27	1	22500
8	16900	1.69	2	31000	28	20100	0.84	1	18000
9	18000	2.53	3	23000	29	19700	1.02	1	12600
10	15300	3.55	4	73000	30	17500	2.37	2	35900
11	19980	0.59	1	1500	31	19300	1.19	1	5000
12	15600	2.28	2	21700	32	15500	3.13	3	39000
13	17200	2.79	3	27570	33	14000	3.21	3	56400
14	18890	1.10	1	13181	34	16900	2.11	2	55000
15	23900	0.17	0	1800	35	17700	2.28	2	25100
16	14320	3.81	4	67210	36	12500	4.23	4	59200
17	11900	4.73	5	73900	37	19000	1.36	1	19000
18	15200	2.53	3	27000	38	22800	0.26	0	5000
19	14450	2.87	3	90000	39	12350	4.23	4	73000
20	18600	2.11	2	27000	40	17800	1.86	2	35000

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. 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]

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.

Bei der Behauptung, dass die beiden Koeffizienten b_1 und b_2 nicht beobachtungsspezifisch seien, handelt es sich genau genommen um eine *Annahme*. Wie wir gleich zeigen werden benötigen wir diese Annahme, um die Koeffizienten überhaupt aus den Daten berechnen zu können.

Im Autobeispiel approximiert die Geradengleichung die Beobachtungen relativ gut, aber es ist auch klar, dass diese Approximation nur für einen bestimmten Bereich

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

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 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 '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

[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'*)

'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 englische 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 *unsystematischen 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.*

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

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

die systematische Komponente $\widehat{y}$ beschreibt also den Teil von y, der mit der erklärenden Variable x 'zusammenhängt'.

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 gleich sehen, 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.

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.

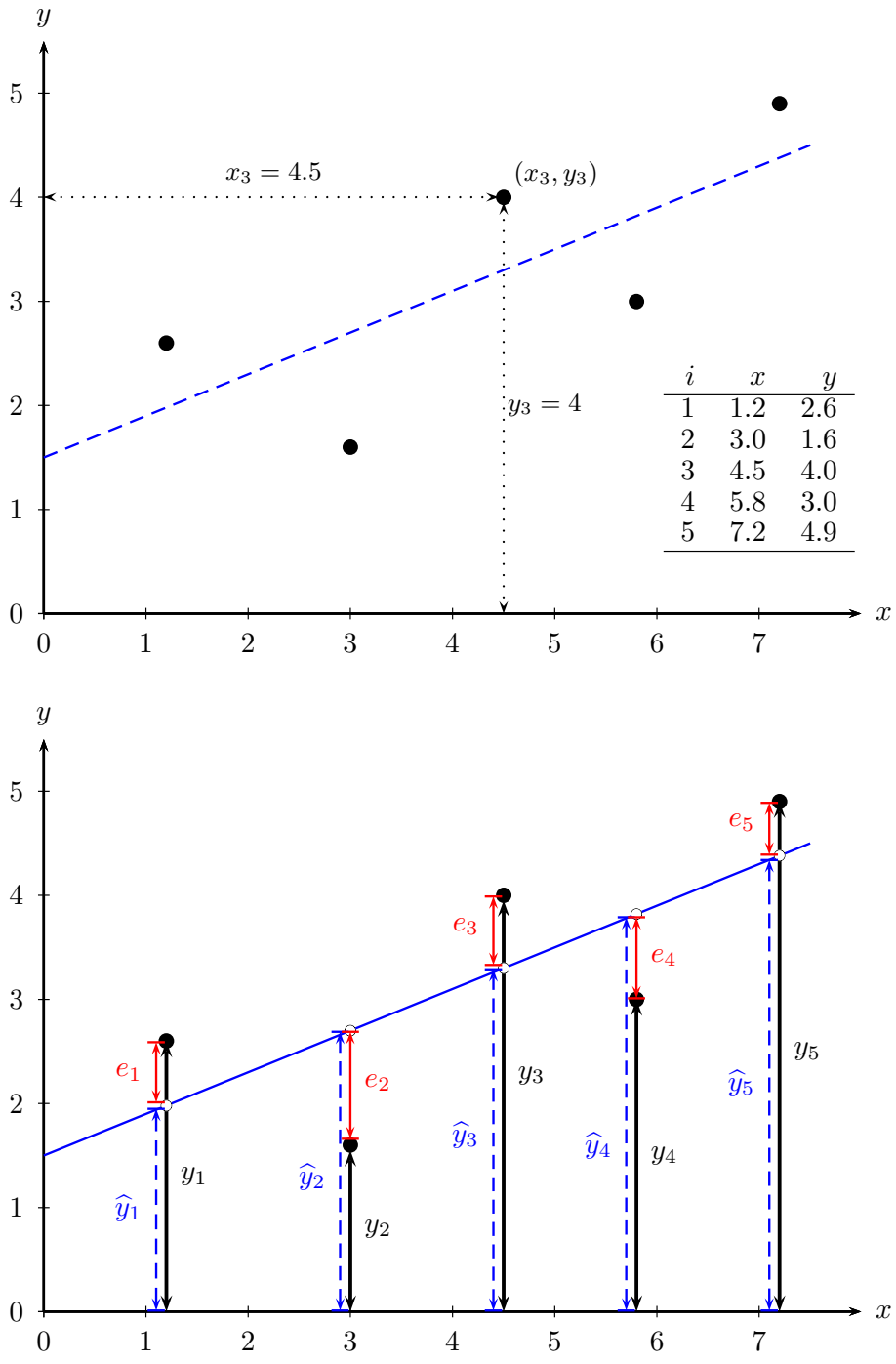


Abbildung 2.2: Zerlegung von y_i in eine systematische Komponente $\widehat{y}_i$ und in ein unsystematisches Residuum e_i (für $i = 1, \ldots, 5$). [local, www]

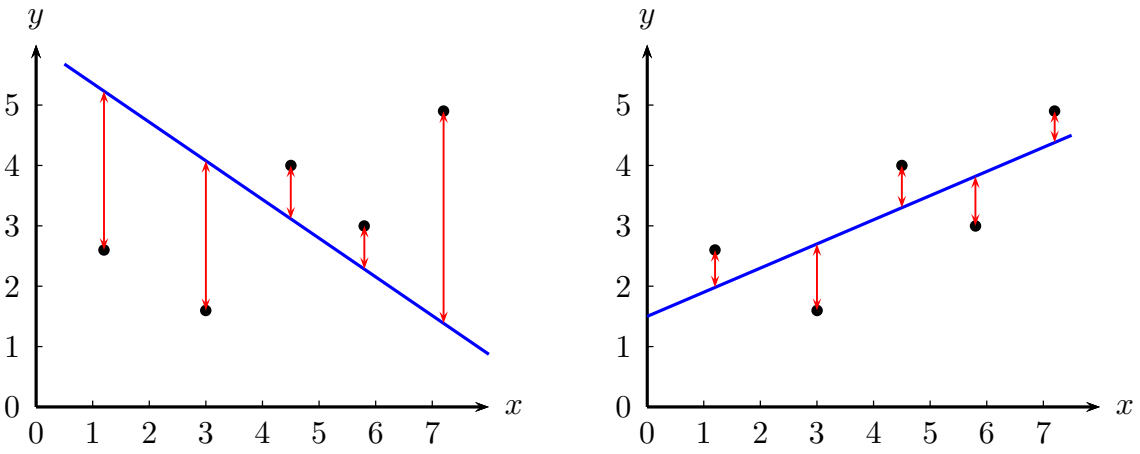


Abbildung 2.3: Die Summe der Abweichungen $\sum_i e_i = \sum_i (y_i - \widehat{y}_i)$ hat in beiden Abbildungen den gleichen Wert, da sich positive und negative Werte aufheben.

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 schwieriger zu lösen, zum anderen werden damit große Abweichungen nicht überproportional stärker gewichtet als kleine Abweichungen. Tatsächlich sind die meisten Menschen 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, die die kleinste Summe der Quadratflächen liefert, 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

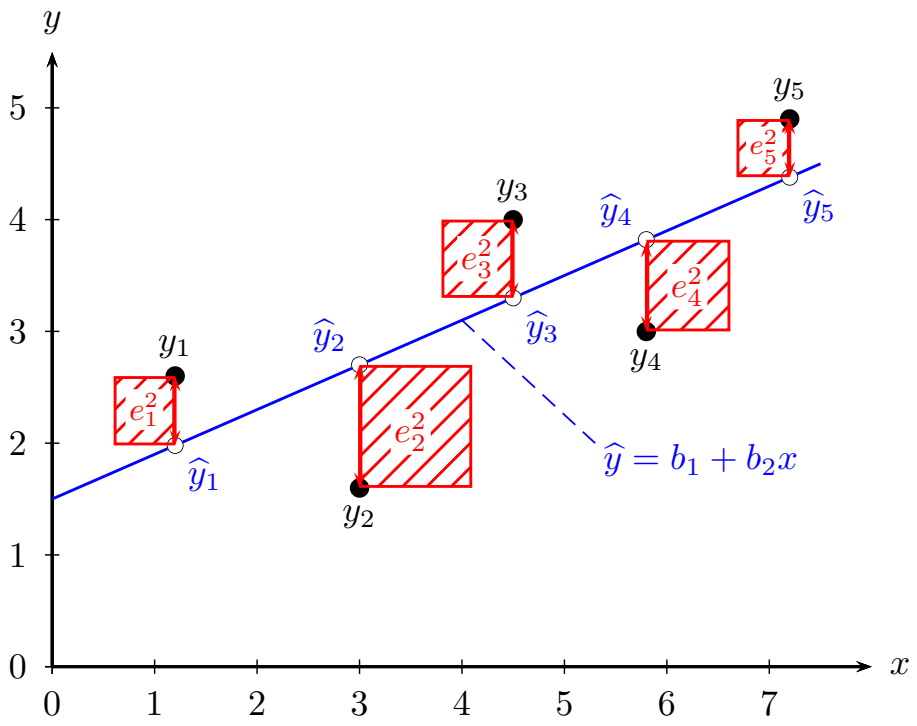


Abbildung 2.4: Nach der OLS Methode werden b_1 und b_2 derart gewählt, dass die *Summe der quadrierten Abweichungen* möglichst klein wird, d.h., die Gesamtfläche der schraffierten Quadrate wird minimiert.

erlauben auch die Grenzen dieses Ansatzes zu verstehen, und in einem weiteren Schritt geeignete Maßnahmen zu ergreifen, wenn die Annahmen verletzt sind, denn 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 folgendermaßen darstellen

$$\begin{aligned} y_i &= \underbrace{(b_1 + b_2 x_i)}_{\widehat{y}} + e_i \qquad \text{bzw.} \\ e_i &= y_i - b_1 - b_2 x_i \end{aligned}$$

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 der 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 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)

$$\boxed{\begin{aligned} \sum_i e_i &= 0 \\ \sum_i x_i e_i &= 0 \end{aligned}}$$

Diese zwei Bedingungen sind von größter Bedeutung, 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

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

[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] $\sum_i x_i e_i = \sum_i e_i(x_i - \bar{x} + \bar{x}) = \sum_i e_i(x_i - \bar{x}) + \bar{x}\sum_i e_i = \sum_i e_i(x_i - \bar{x}) = \sum_i (e_i - \bar{e})(x_i - \bar{x}) = 0.$

können, und dass $\sum_i b_1 = nb_1$, weil b_1 eine Konstante ist

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

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

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]

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

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

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

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

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

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

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

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

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

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

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

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

□

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

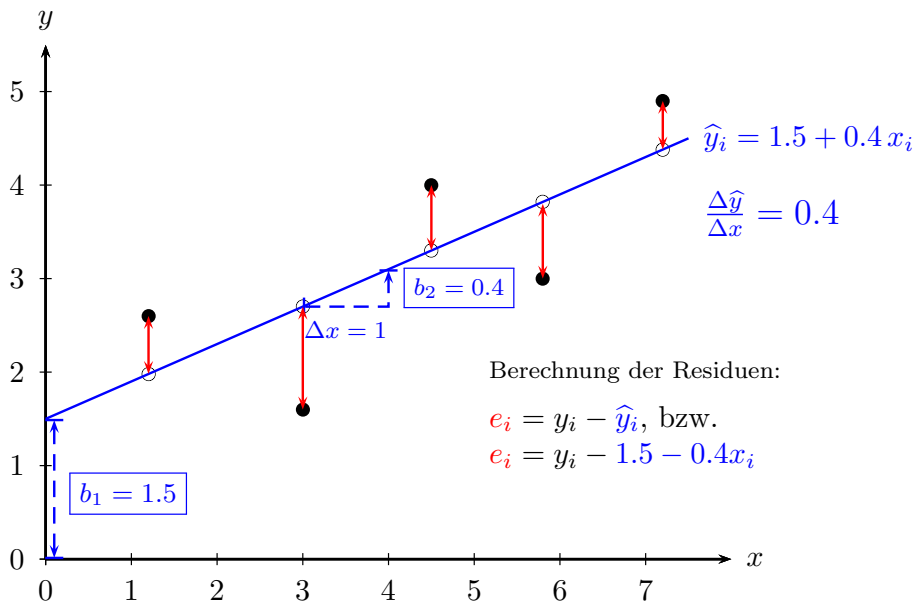


Abbildung 2.5: Beispiel

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 gefittete 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.

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

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 12).

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 Interpretation des deskriptiven bivariaten Regressionsmodells

"Es ist schon alles gesagt, nur noch nicht von allen."
(Karl Valentin, 1882–1948)

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 die die gefitteten Werte $\widehat{y}$ und die Koeffizienten interpretiert werden können. Diese Einsichten werden im nächsten Abschnitt über das multiplen Regressionsmodell erweitert, und liefern uns auch 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 nichtsystematische Komponente, die Residuen e, zerlegt.

Für die Interpretation interessieren wir uns naheliegenderweise ausschließlich für die *systematische* Komponente

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

oder für das frühere Gebrauchtautobeispiel $\widehat{\text{Preis}}_i = 23\,057 - 2\,636\,\text{Alter}_i$ (siehe Abbildung 2.1, Seite 3), 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 'erklärt'.

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

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

Tabelle 2.3: Autopreise nach gerundetem Alter. $\bar{y}$ bezeichnet das arithmetische Mittel nach Altersklassen und $\widehat{y}$ die gefitteten Werte der Regression $\widehat{y}_i = 22\,709 - 2\,517x_i$.

	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

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

Dazu kommen wir dazu nochmals auf das Beispiel mit den Gebrauchtautos zurück, aber wir wenden einen Trick an: wir runden die erklärende Variable 'Alter' auf ganze Jahre. Damit wird aus der stetigen Variable 'Alter' eine diskrete Variable, 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 4).

Tabelle 2.3 zeigt die gleichen Beobachtungen wie Tabelle 2.1, aber gruppiert nach 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.

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

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

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

Wenn wir dies für alle Altersstufen machen erhalten wir die *bedingte Mittelwertfunk-*

tion, jeder Altersstufe 'AlterJ' wird ein bedingter Mittelwert zugeordnet

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

mit $j \in \{0, 1, 2, \ldots, 5\}$.

Zeile $\overline{y}$ in Tabelle 2.3 zeigt ebenfalls diese bedingte Mittelwertfunktion.

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, jeweils 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 $\Delta\overline{y}$ $(:= \overline{y}_t - \overline{y}_{t-1}$, mit $t = 1, \ldots, 5)$ 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

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

Für die Berechnung dieser Werte benötigen wir lediglich die zwei OLS Koeffizienten b_1 und b_2, wir erreichen also einen noch größere 'Informationsverdichtung', die allerdings auf Kosten der Genauigkeit geht.

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

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.

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;

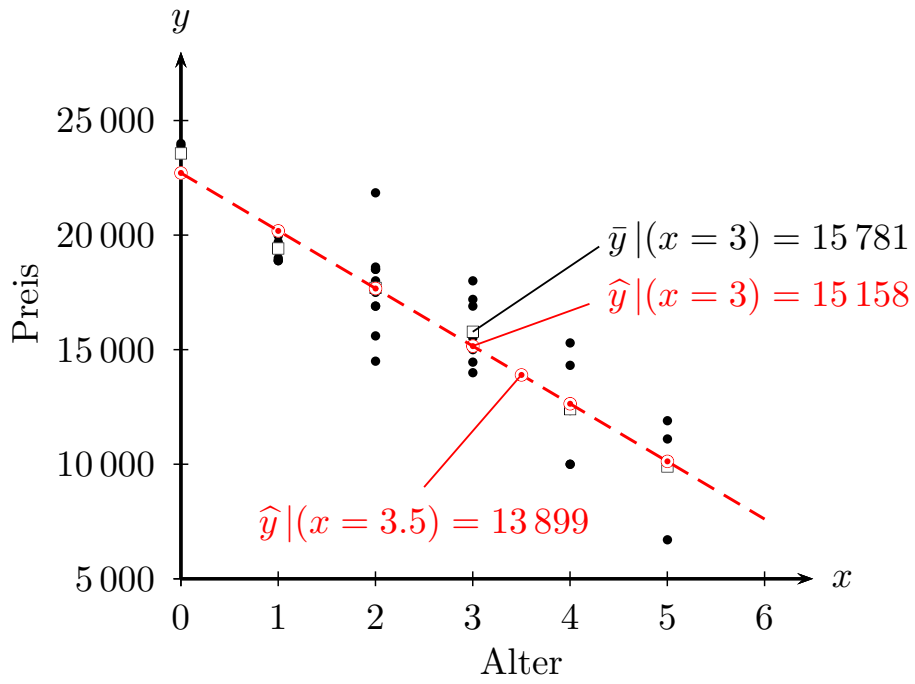


Abbildung 2.6: Deskriptive Regression als lineare Approximation an die 'bedingte Mittelwertfunktion'. (• Beobachtungen; □ bedingte Mittelwerte; ⊙ lineare Approximation).

und sie dient hier 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. Mit Hilfe der linearen Funktion $\widehat{y} = b_1 + b_2 x$ können wir $\widehat{y}$ für beliebige x zu 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 gar kein 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 in diesem Fall diese Interpretation auf der angenommenen linearen Funktionsform beruht, die diese Interpolation ermöglichte.

Diese Intuition bleibt auch dann gültig, wenn wir überhaupt keine wiederholten y-Beobachtungen für Ausprägungen der x-Variable haben, wie z.B. im ursprünglichen Beispiel aus Abbildung 2.1 (Seite 3).

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

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

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 einer

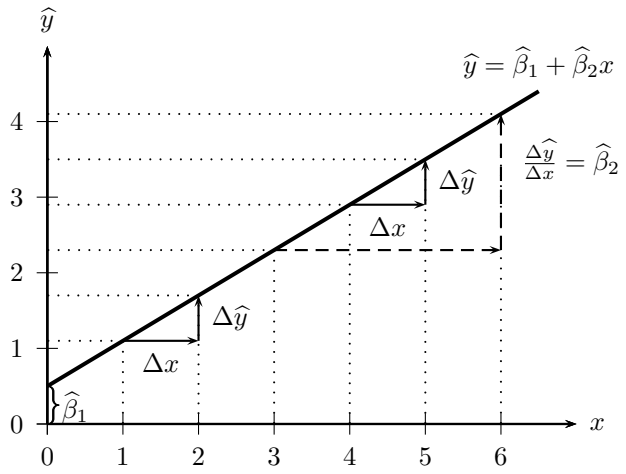


Abbildung 2.7: Lineare Funktion $\widehat{y} = b_1 + b_2 x = 0.5 + 0.6x$. Eine Zunahme von x um eine Einheit geht einher mit einer Änderung von $\widehat{y}$ um $+0.6$ Einheiten.

Änderung des 'mittleren' Preises oder Durchschnittspreises sprechen, aber es ist wichtig im Kopf zu behalten, dass wir in der linearen Regressionsanalyse jeweils von linearen Approximationen sprechen.

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[7] 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 meist 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 (siehe 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.

[7] 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.

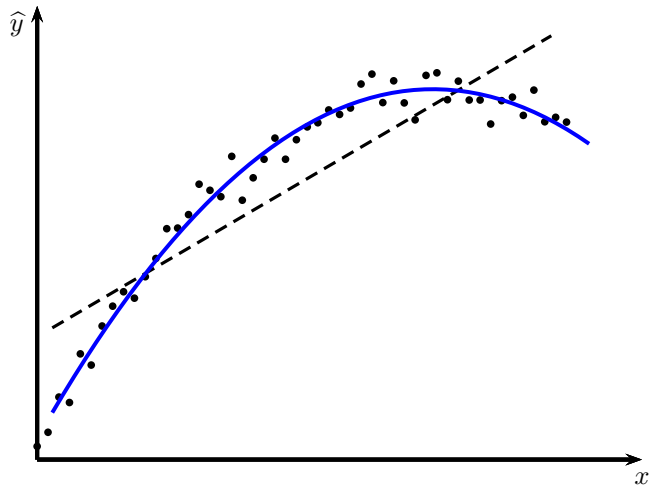


Abbildung 2.8: Eine lineare Funktion $\widehat{y} = b_1 + b_2 x$ kann einen sehr schlechten Fit liefern, wenn der tatsächliche Zusammenhang nicht-linear ist. Offensichtlich würde in diesem Fall eine nicht-lineare Funktion wie die strichlierte Linie einen deutlich besseren Fit liefern, aber für nicht-lineare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

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.

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

Es wäre verlockend zu sagen, dass eine Zunahme des Alters um ein Jahr eine Veränderung des 'mittleren' Preises um $b_2 = 2\,517$ Euro *verursacht*, aber dies wäre falsch! 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 Rechtfertigung bedürfen.

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

In Beispiel mit den Gebrauchtautos wurden die bedingten Mittelwerte durch eine lineare Funktion sehr gut approximiert, aber dies muss aber 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

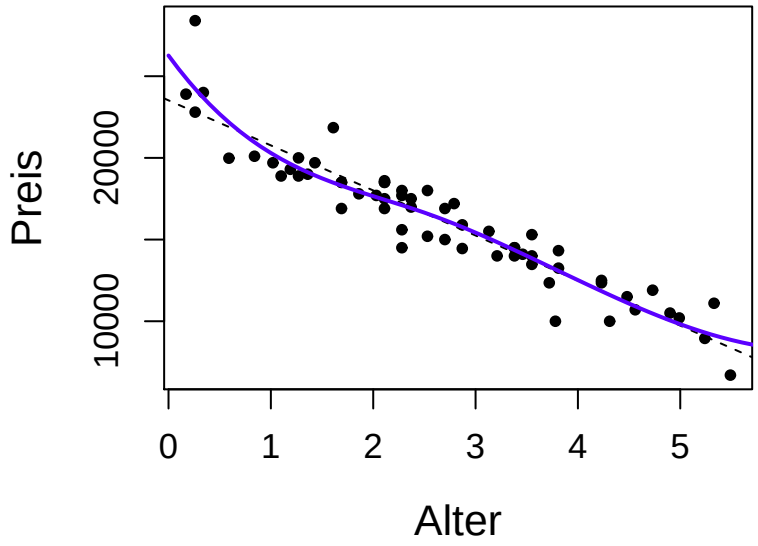


Abbildung 2.9: Spline-Funktion für die Preise von Gebrauchtautos

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

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

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 können nicht mehr so 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. Generell können wir festhalten

$$\text{Daten} + \text{Annahmen} \rightarrow \text{Schlussfolgerungen}$$

Es gibt keine Datenanalyse, die völlig ohne 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.

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

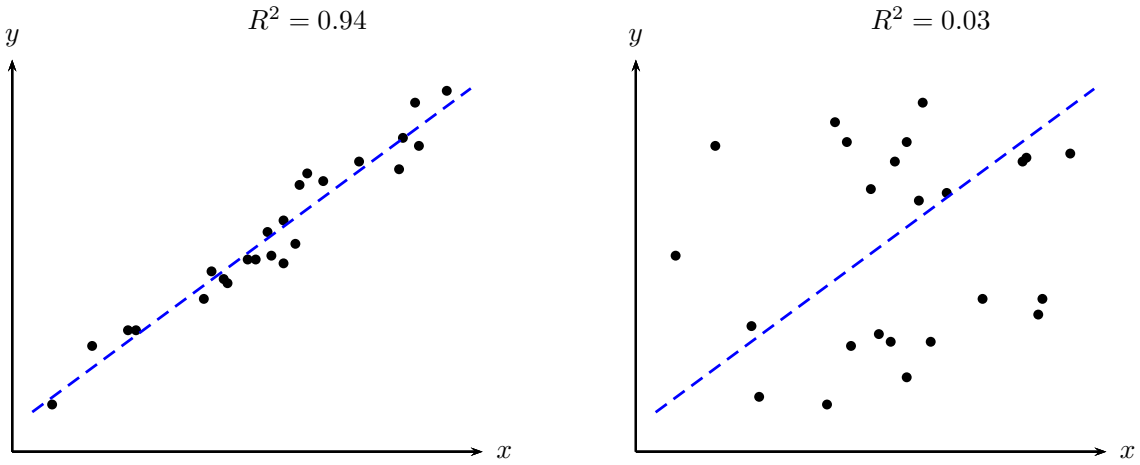


Abbildung 2.10: Der Zusammenhang zwischen zwei Variablen kann durch eine Regressionsgerade mehr oder weniger gut beschrieben werden.

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

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

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 $b_1 = \bar{y} - b_2 \bar{x}$ verwendet.

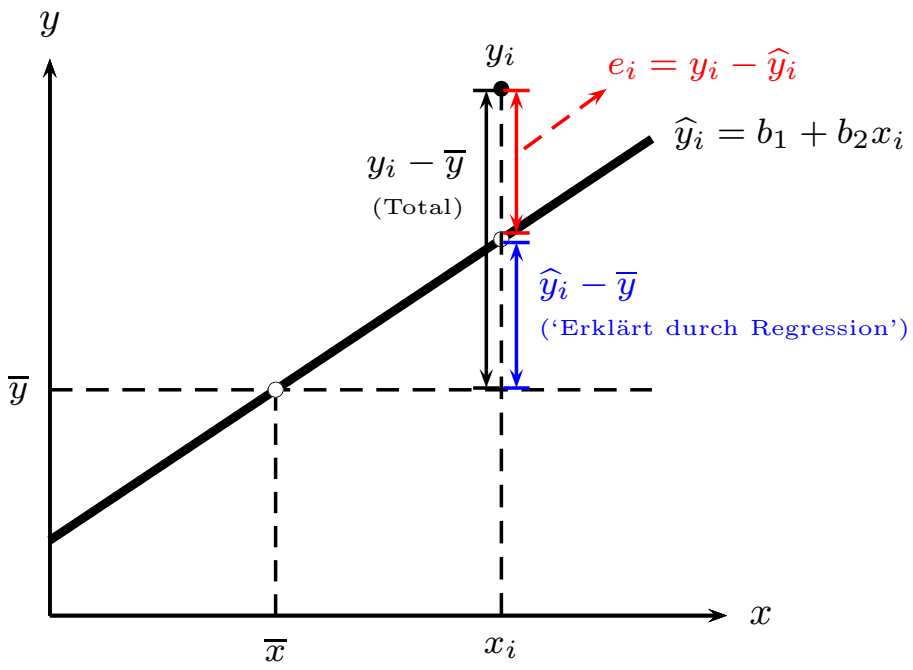


Abbildung 2.11: Zerlegung der gesamten Streuung von y in einen 'erklärten' und einen 'unerklärten' Teil.

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 $y_i - \bar{y}$.

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 $y_i - \widehat{y}_i = e$.

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 12).

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 12) 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.

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

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.

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?).

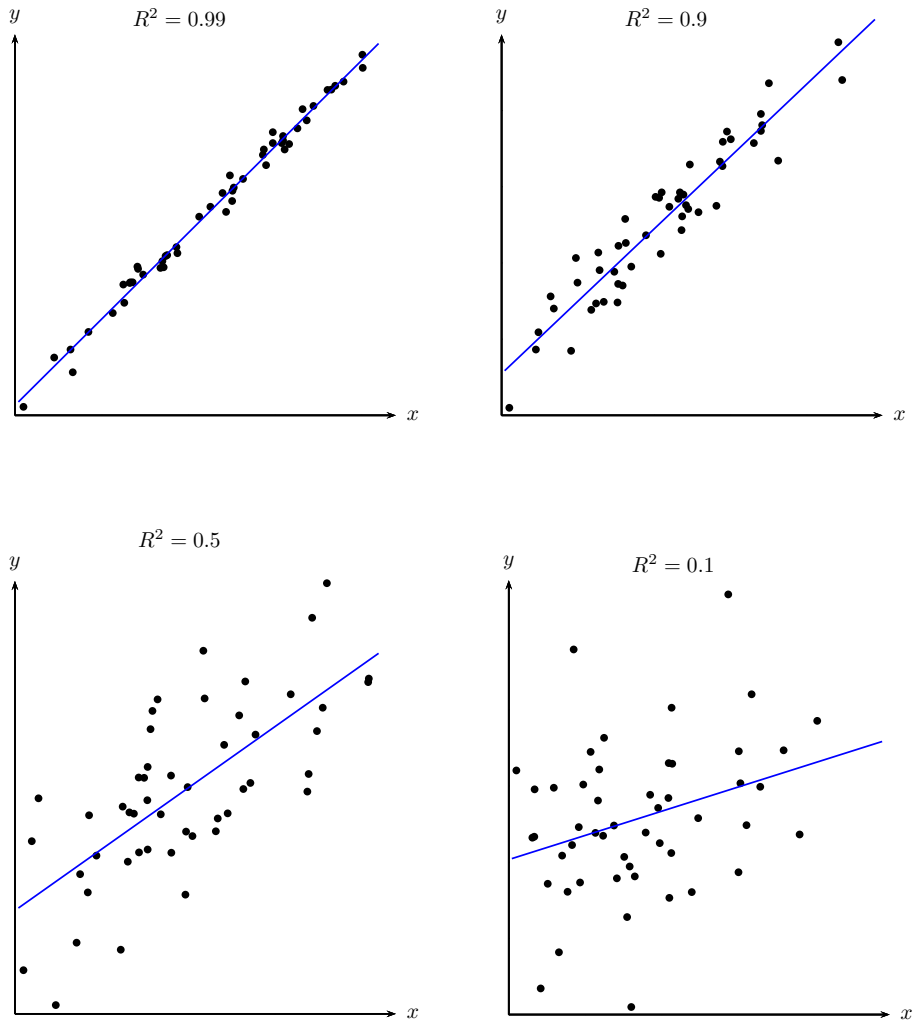


Abbildung 2.12: Das Bestimmtheitsmaß R^2 ist ein Indikator für die Streuung um die Regressionsgerade.

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).

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

Lösung: Zeigen Sie zuerst, dass

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

Einsetzen gibt

$$R^2 = \frac{[\mathrm{cov}(y,\widehat{y})]^2}{\mathrm{var}(y)\,\mathrm{var}(\widehat{y})} = \frac{b_2^2[\mathrm{cov}(y,x)]^2}{\mathrm{var}(y) b_2^2\,\mathrm{var}(x)} = \frac{[\mathrm{cov}(y,x)]^2}{\mathrm{var}(y)\,\mathrm{var}(x)} = r^2_{y,x}$$

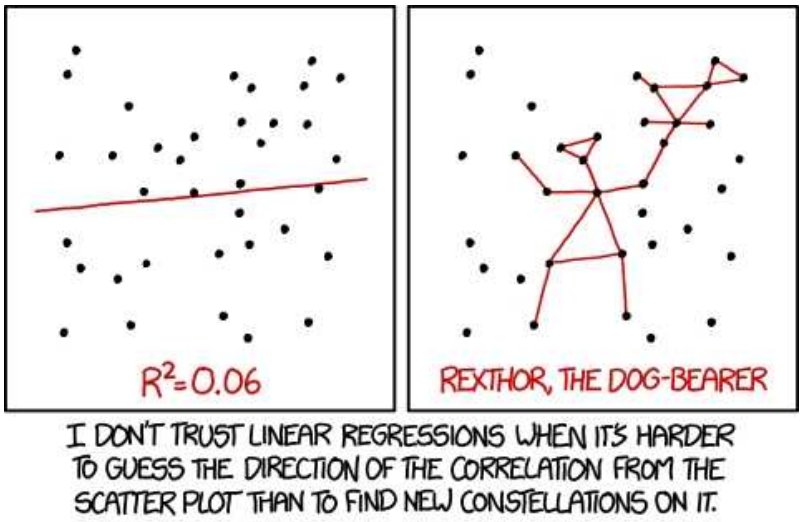


Abbildung 2.13: Quelle xkcd, `http://xkcd.com/1725/`

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 2.14 zeigt eine solche 3-dimensionale Abbildung mit der abhängigen y Variable auf der Vertikalachse und zwei erklärenden Variablen x_1 und x_2 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 2.14 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

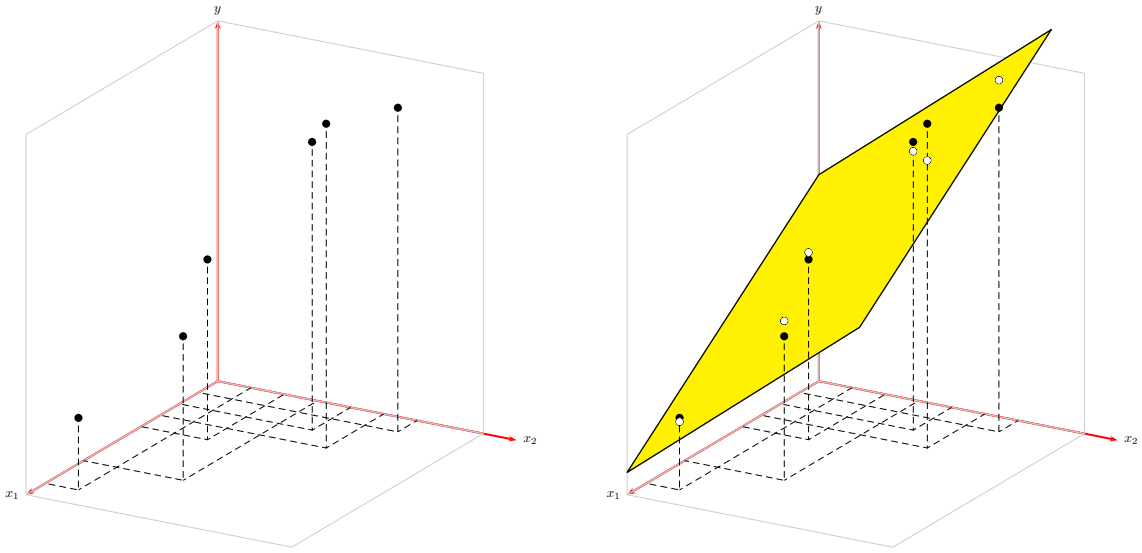


Abbildung 2.14: 3-dimensionale Abbildung der Daten und der Regressionsebene $\widehat{y}_i = 5.73 - 0.51x_{i1} + 0.76x_{i2}$ (gefittete Werte auf der Regressionsebene sind als hohle Kreise dargestellt)

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 2.8.1 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.

Zur Verdeutlichung noch einmal ausführlich in Vektornotation

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

Gleich wie früher im bivariaten Modell können die gefitteten Werte $\widehat{y}$ wieder als *lineare Approximation an die bedingten Mittelwerte* interpretiert werden. Im Autobeispiel könnten wir z.B. die lineare Approximation an den bedingten Mittelwert der Preise von Autos mit Alter = 1.5 *und* km = 20 000 berechnen. Wie wir bereits gesehen haben funktioniert dies selbst dann, wenn unsere Stichprobe kein einziges Auto mit diesen Charakteristika enthält, die angenommene lineare Funktionsform ermöglicht die Berechnung einer solchen Approximation.

Die unterstellte lineare Funktionsform ist auch der Grund dafür, dass die Koeffizienten einfach als **marginale Effekte** interpretiert werden können, denn die Regressionskoeffizienten sind einfach die partiellen Ableitungen und können als solche interpretiert werden.

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 $\widehat{y}$ verändert, wenn x_2 um eine Einheit zunimmt *und* x_3 unverändert bleibt, d.h. *ceteris paribus*. Analoges gilt für b_3

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

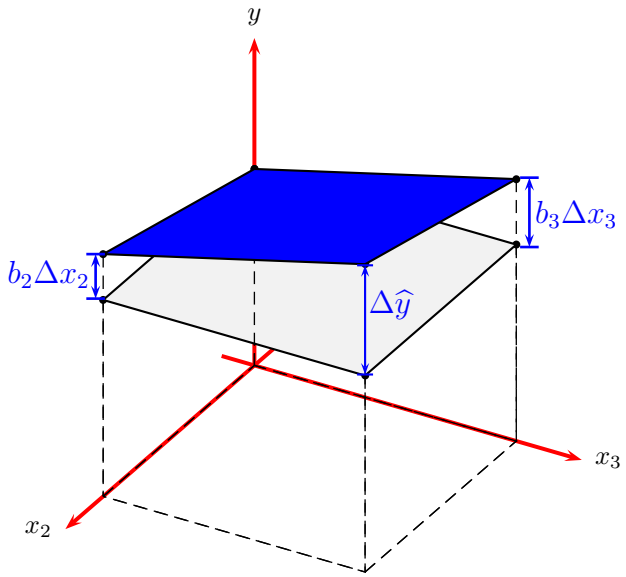


Abbildung 2.15: Ceteris-paribus Interpretation der Koeffizienten;
Für die Regressionsebene $\widehat{y}_i = b_1 + b_2 x_{i2} + b_3 x_{i3}$ folgt durch Bildung erster Differenzen $\Delta\widehat{y} = b_2\Delta x_2 + b_3\Delta x_3$, woraus für die Koeffizienten folgt

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

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.[8]

Eine Regression des Verkaufspreises auf Alter *und* Kilometerstand gibt

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

[8]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.

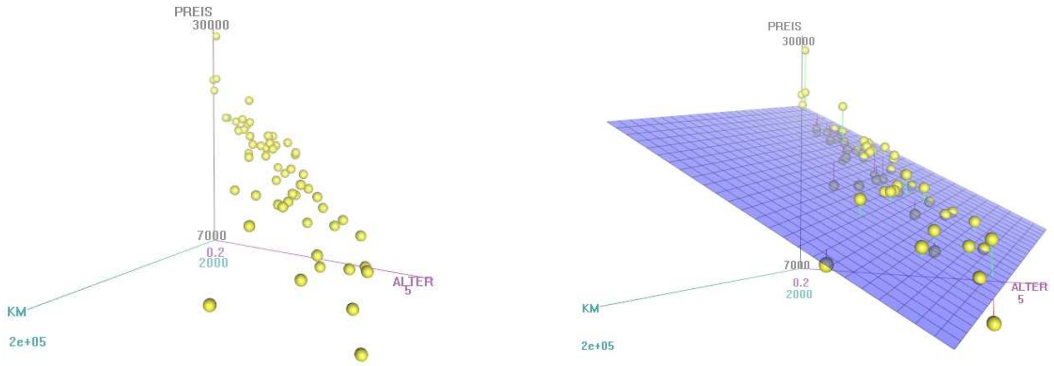


Abbildung 2.16: 3-dimensionale Abbildung des Autobeispiels mit Hilfe des R packages `Rcmdr` (Fox, 2005).

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

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

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

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 ein 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.

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.

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.[9] Diese Interpretation ist 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 linearen Funktionsform. Falls die Daten durch eine lineare Funktionsform nur sehr schlecht approximiert werden oder wesentliche erklärende Variablen fehlen kann diese Interpretation zu irreführenden Schlussfolgerungen führen.

Tatsächlich haben wir durch die Wahl der linearen Funktionsform die Daten gewissermaßen auf das Prokrustes-Bett[10] 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 $\bar{R}^2$ (*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 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

[9] Man beachte, dass sich diese ceteris-paribus Interpretation nur auf die *im Modell vorkommenden* x Variablen bezieht.

[10] Prokrustes – eine Figur aus der griechischen Mythologie – war bekannt dafür Reisenden ein Bett anzubieten, und sie 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.

Tabelle 2.4: 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

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 $\bar{R}^2$ wird versucht dieses Problem zumindest zu mildern, indem ein Korrekturfaktor eingeführt wird.

$$R^2 = 1 - \frac{\sum_i e_i^2}{\sum_i (y_i - \overline{y})^2} \quad \Rightarrow \quad \bar{R}^2 = 1 - \frac{\frac{\sum_i e_i^2}{(n-k)}}{\frac{\sum_i (y_i - \overline{y})^2}{(n-1)}} = 1 - \left(\frac{\sum_i e_i^2}{\sum_i (y_i - \overline{y})^2} \right) \left(\frac{n-1}{n-k} \right)$$

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.

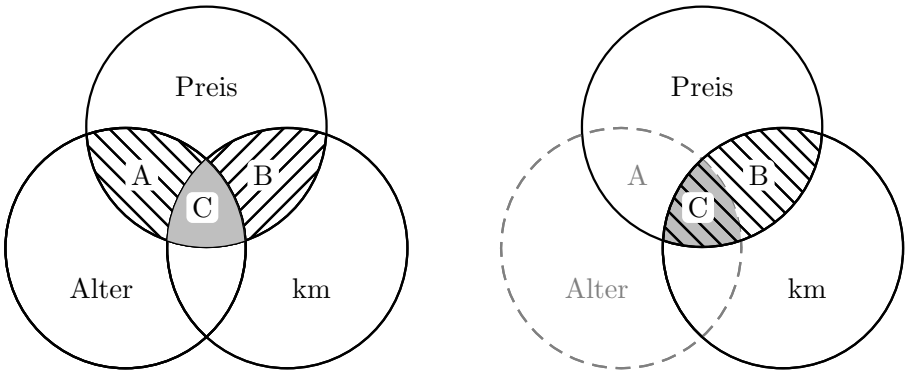
2.6.1 Nichtberücksichtigung relevanter Variablen

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 2.4 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

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.



$\widehat{\text{Preis}} = b_1^l + b_2^l \text{Alter} + b_3^l \text{km}$ $\qquad$ $\widehat{\text{Preis}} = b_1^k + b_2^k \text{km}$

Abbildung 2.17: '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).

Einen intuitiven Einblick gibt das Venn Diagramm in Abbildung 2.17. 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. 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 der nicht berücksichtigte Regressor sowohl mit der abhängigen Variable y als auch mit dem berücksichtigten Regressor x korreliert ist.

Das linke Panel des Venn Diagramms in Abbildung 2.17 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 2.8.1). Durch die Mittelwerttransformation fällt das Interzept weg, was die folgende Darstellung vereinfacht (um die Lesbarkeit zu erhöhen verzichten wir zudem auf den Beobachtungsindex i)

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

Wir vergleichen nun den Steigungskoeffizienten b_2 dieses 'langen' Modells mit dem Steigungskoeffizienten eines 'kurzen' Modells, in dem wir $\ddot{y}$ nur auf $\ddot{x}_2$ regressieren

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

Der OLS Steigungskoeffizient des 'kurzen' Modells ist

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

Um zu erkennen, was bei der Nichtberücksichtigung von $\ddot{x}_3$ passiert, setzen wir in die obige OLS-Formel für den Steigungskoeffizienten des kurzen Modells b_2^k für $\ddot{y}$ das 'lange' Modell $\ddot{y} = b_2^l \ddot{x}_2 + b_3^l \ddot{x}_3 + e^l$ ein und vereinfachen

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

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

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

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

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

$$x_3 = a_1 + a_2 x_2 + u, \quad \Rightarrow \quad a_2 = \frac{\mathrm{cov}(x_2, x_3)}{\mathrm{var}(x_2)}$$

wobei u die Residuen dieser Regression bezeichnet.

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

$$\boxed{b_2^k = b_2^l + b_3^l \, a_2} \tag{2.13}$$

Wenn – und nur wenn – b_3 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 2.18 zeigt das Problem noch einmal: wenn x_3 nicht berücksichtigt wird, wird x_2 neben seiner direkten Wirkung b_2 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, 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 (2.4) 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 (2.4) war

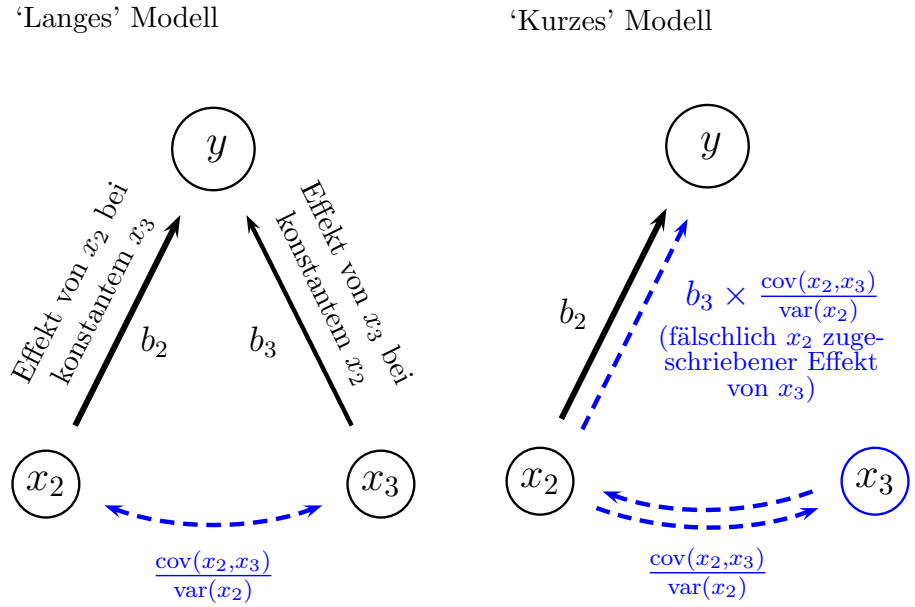


Abbildung 2.18: Nichtberücksichtigung einer relevanten Variable x_3 führt dazu, dass ein Teil der Auswirkungen von x_3 fälschlich x_2 zugeschrieben wird. Wenn das 'wahre' Modell $y = b_1 + b_2^l x_2 + b_3^l x_3 + e^l$ ist und irrtümlich ein kurzes Modell $y = b_1^k + b_2^k x_2 + u$ geschätzt wird ist $b_2^k = b_2^l + b_3^l \frac{\text{cov}(x_2,x_3)}{\text{var}(x_2)}$.

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

und die Hilfsregression $\text{km} = a_1 + a_2\,\text{Alter} + u$ 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 (2.4) 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).

Wozu war das nun alles gut? Die ganze Tragweite dieses Resultats wird erst später im Rahmen der stochastischen Regressionsanalyse deutlich werden, dort werden wir dieses Phänomen einen *"Omitted Variable Bias"* nennen.

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. $\text{cov}(\text{km}, \text{Alter}) > 0$). Da $b_2^k = b_2^l + b_3^l \times a_2$ und $b_3^l \times a_2 < 0$ folgt $b_2^k < b_2^l$.

Tabelle 2.5: 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 $y = b_1^l + b_2^l x_2 + b_3^l x_3 + e^l$.
Da $b_2^k = b_2^l + b_3^l \times \mathrm{cov}(x_2, x_3)/\,\mathrm{var}(x_2)$ gilt:

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

2.6.2 Partielle Regression und 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. Lovell (1963) 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 Regressionen berechnet werden kann.

Als Frisch and Waugh (1933) dieses Ergebnis bewiesen waren Computer noch kaum verfügbar, deshalb waren multiple Regressionen weit schwieriger zu berechnen als bivariate Regressionen, dieses Ergebnis hatte damals also 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 der drei folgenden bivariaten Regressionen berechnen

$$\begin{aligned} y &= c_1 + c_2 x_3 + e^y \\ x_2 &= a_1 + a_2 x_3 + e^{x_2} \\ e^y &= b_2 e^{x_2} + e \end{aligned}$$

wobei e^y die Residuen der ersten bivariaten Gleichung und e^{x_2} die Residuen der zweiten bivariaten Gleichung bezeichnet.

In Worten: wir schätzen zuerst zwei Hilfsregressionen, zuerst regressieren wir y auf die zu eliminierende Variable x_3, und anschließend den interessierenden Regressor x_2 auf die zu eliminierende Variable x_3, und speichern die Residuen dieser beiden Hilfsregressionen.

Wenn wir dann die beiden Residuen dieser Hilfsregressionen aufeinander regressieren erhalten wir exakt den gleichen Steigungskoeffizienten b_2 sowie die gleichen Residuen, die wir aus der ursprünglichen multiplen Regression erhalten hätten.

Durch die beiden 'kurzen' Regressionen auf x_3 wird gewissermaßen der (lineare) Einfluss von x_3 auf y und x_2 eliminiert. Im Englischen wird dies häufig *'partialling out'* genannt. Wie schon erwähnt wurde dieses Resultat von Lovell (1963) für mehrere Variablen verallgemeinert.

Der Beweis dieses Theorems erfolgt üblicherweise unter Zuhilfenahme von Matrixalgebra. Wir werden hier einen deutlich einfacheren Beweis skizzieren, der Lovell (2008) folgt.

Erinnern wir uns, die OLS Methode ist im wesentlichen eine Zerlegungsmethode, sie zerlegt eine abhängige Variable y in einen systematische Komponente $\widehat{y}$ und eine *damit unkorrelierte* Restkomponente, die Residuen e.

Unser Ausgangspunkt ist eine einfache 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 $\sum_i x_{ih} e_i = 0$ für $h = 2, 3$.
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) $\text{cov}(y, x_3) = 0$ und $\text{cov}(x_2, x_3) = 0$ ist, dann folgt $b_3 = 0$.

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 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

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

Aus Gleichung (2.15) wissen wir aber, dass $\text{cov}(x_3, e^y) = 0$, und aus Gleichung (2.16), dass $\text{cov}(x_3, e^{x_2}) = 0$, 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 folgendem Code bewerkstelligt werden:

```
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)

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
```

Das (mehr oder weniger) gleiche in Stata:

```
clear all
insheet using http://www.hsto.info/econometrics/data/auto40.csv, delimiter(";")
destring alter, dpcomma replace
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 die partiellen Zusammenhänge korrekt darzustellen, indem wir zuerst mittels Hilfsregressionen den linearen Einfluss aller anderen Variablen eliminieren, und anschließend die Residuen dieser Hilfsregressionen in einem Streudiagramm darstellen.[11] 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 2.17).

Abbildung 2.19 zeigt den Zusammenhang zwischen Preis und Kilometerstand von Gebrauchtautos links ohne Berücksichtigung des Alters, und rechts nachdem für das Alter kontrolliert wurde (die drei übereinanderliegenden Grafiken wurden mit EViews, R und Stata erzeugt, sind aber ansonsten identisch. Tabelle 2.6 zeigt den Programmcode, mit dessen Hilfe diese Grafiken erstellt wurden).

Aufgrund der Abbildungen im linken Panel würden wir einen sehr engen Zusammenhang zwischen km und Preis erwarten. Im rechten Panel wurde der lineare Einfluss des Alters auf beide Variablen eliminiert; dadurch wird deutlich, dass uns ein einfaches bivariates Streudiagramm einen falschen Eindruck vom Zusammenhang zwischen Kilometerzahl und Preis vermitteln würde.

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 (sehr) irreführenden Schlussfolgerungen führen!

Tatsächlich haben wir die Daten gewissermaßen auf das Prokrustes-Bett[12] unserer linearen Spezifikation gespannt!

Die Annahme der *Linearität* ist allerdings nicht ganz so restriktiv wie es auf den ersten Blick scheinen mag, denn sie bezieht sich nur auf Linearität in den Parametern, aber *nicht* auf Linearität in den Variablen. Modelle, die nicht-linear in den Variablen sind, können ganz normal mit OLS geschätzt werden.

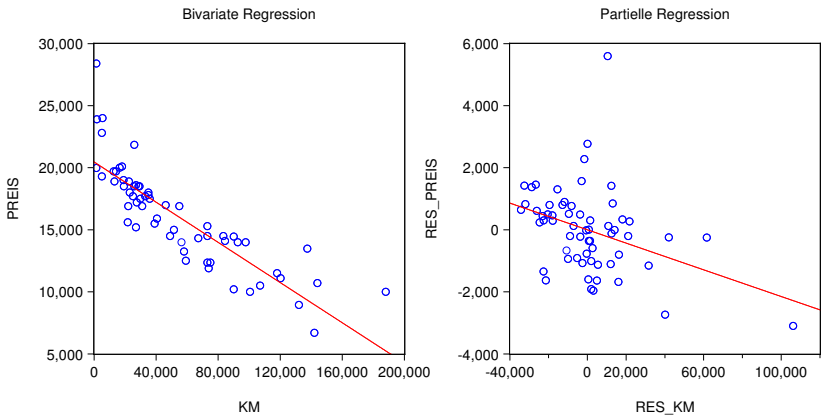
So können wir z.B. für das Modell

$$y = b_1 + b_2 x_2^2 + b_3 \log(x_3) + b_4 x_2 x_3 + e$$

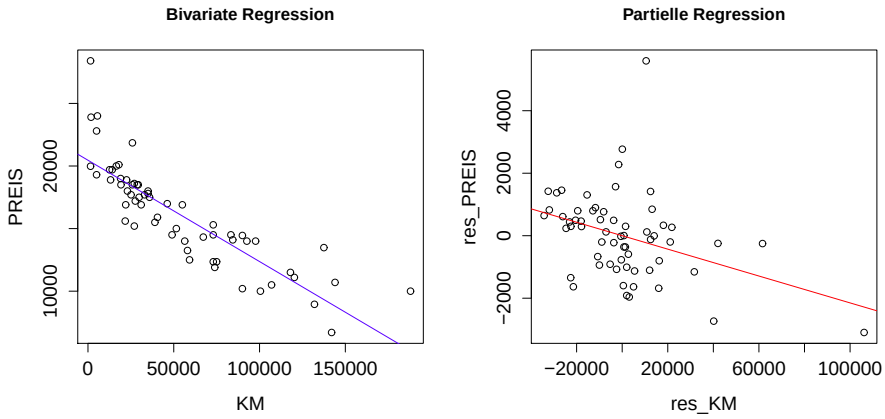
[11] Allerdings ist dabei zu beachten, dass dadurch die Skalierung geändert wird.

[12] 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.

EViews:



R:



Stata:

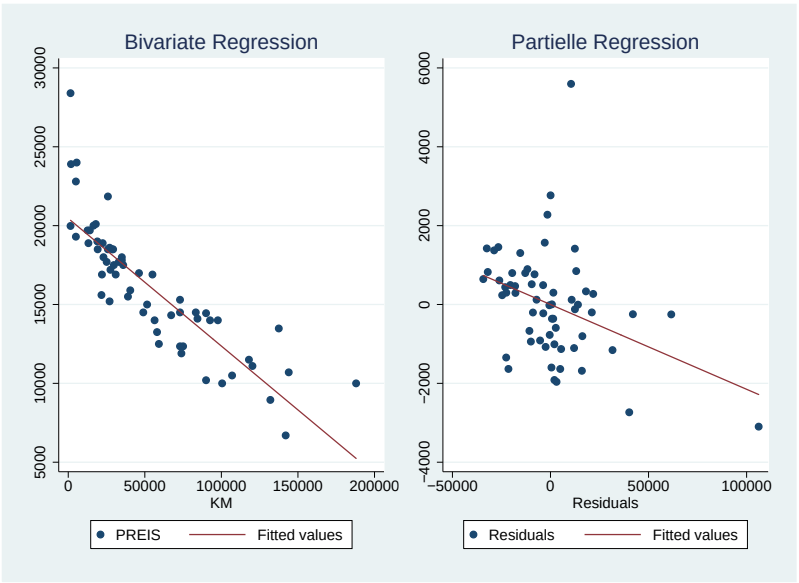


Abbildung 2.19: Bivariate und partielle Regression: bei der partiellen Regression werden Residuen nach Eliminierung des Alters geplottet (Standardoutput von EViews, R und Stata).

Tabelle 2.6: EViews-, R- und Stata-Programmcode, der Abbildung 2.19 erzeugt.

EViews:

```
wfopen(type=text) "http://www.uibk.ac.at/econometrics/data/auto.csv" delim=";"
group Gr_P KM PREIS
freeze(Graph1) Gr_P.scat linefit
Graph1.addtext(t) Bivariate Regression
equation eq_PREIS.ls PREIS c ALTER
eq_PREIS.makeresid res_PREIS
equation eq_KM.ls KM c ALTER
eq_KM.makeresid res_KM
group Gr_res res_KM res_PREIS
freeze(Graph2) Gr_res.scat linefit
Graph2.addtext(t) Partielle Regression
graph Graph3.merge Graph1 Graph2
GRAPH3.align(2,1,0)
```

R:

```
rm(list=ls())
Auto <- read.csv2("http://www.hsto.info/econometrics/data/auto40.csv", dec=".")
attach(Auto)
res.Preis <- resid(lm(Preis ~ Alter))
res.km <- resid(lm(km ~ Alter))
par(mfrow=c(1,2),cex.main=0.85)
plot(km,Preis, main="Bivariate Regression")
abline(lm(Preis ~ km),col="blue")
plot(res.km,res.Preis, main="Partielle Regression")
abline(lm(res.Preis ~ res.km),col="red")
```

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)
```

neue Variablen definieren $z_2 = x_2^2,\ z_3 = \log(x_3)$ und $z_4 = x_2 x_3$ und die Koeffizienten des Modells

$$y = b_1 + b_2 z_2 + b_3 z_3 + b_4 z_4 + e$$

wie üblich mit OLS schätzen.

Man beachte, dass dieses Modell zwar nicht-linear in den Variablen x_2 und x_3 ist, *aber linear in den Parametern* b_1, b_2 und b_3. Um Modelle mit OLS schätzen zu können müssen diese nur linear in den Parametern sein, Linearität in den Variablen ist nicht erforderlich.

Hingegen benötigt man für Modelle, die *nicht-linear in den Parametern* sind, wie z.B.

$$y = b_1 + b_2^2 x_1 + \log(b_3) x_2 + b_2 b_3 x_2 + e$$

andere Methoden, auf die wir hier nicht eingehen werden.

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."*
> (Machlup, 1974, 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*[13] 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

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

[13] 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 $u = 1 - v$. (`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:

$$\begin{aligned}
(\overline{\text{StdL}}|m = 1) &= (16 + 16 + 14 + 12 + 18 + 14)/6 = 15 \\
(\overline{\text{StdL}}|w = 1) &= (12 + 12 + 14 + 14 + 10 + 13)/6 = 12.5 \\
(\overline{\text{StdL}}|v = 1) &= (16 + 14 + 12 + 18 + 14 + 14 + 10)/7 = 14 \\
(\overline{\text{StdL}}|v = 0) &= (16 + 12 + 12 + 14 + 13)/5 = 13.4
\end{aligned}$$

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 17). 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

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

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

Für Männer ist $m_i = 1$, also $\widehat{\mathrm{StdL}}_i = b_1 + b_2 \times 1 = b_1 + b_2$, 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.

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

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 $m_i = 0$).

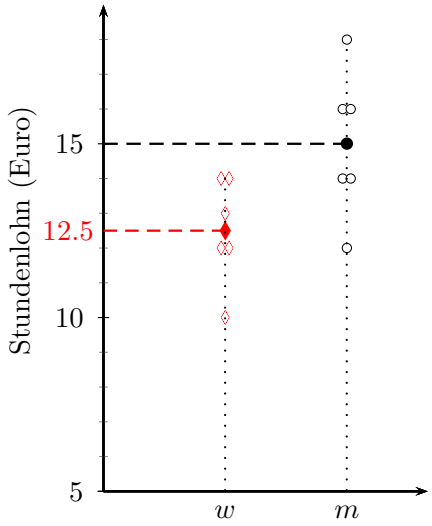
$$\overline{\mathrm{StdL}}|(m = 1) - \overline{\mathrm{StdL}}|(m = 0) = 15 - 12.5 = 2.5 = b_2$$

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

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

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 ($b_1 = 15$), und der durchschnittliche Stundenlohn für Frauen ist um 2.5 Euro *niedriger* als der von Männern ($b_2 = -2.5$). Abbildung 2.20 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



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

bzw.

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

Abbildung 2.20: Stundenlöhne von Männern und Frauen, siehe Tabelle 2.7.

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.[14]

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:* Sei y_i eine von insgesamt n Beobachtungen einer intervallskalierten Variable, und d eine Dummy Variable; n_1 Elemente dieser Dummy Variable haben die Ausprägungen Eins und n_0 Elemente den Wert Null ($n_1 + n_0 = n$).

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

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

[14] wir werden diesen Fall später unter der Bezeichnung *perfekte Multikollinearität* ausführlich diskutieren.

1. Sei $\bar{y}_1$ der Mittelwert der y_i für die gilt $d_i = 1$, und $\bar{y}_0$ der Mittelwert aller y_i für die $d_i = 0$. 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 $d_i = 1$

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

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

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

Lösungsskizze:

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

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

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

Hinweis: $\text{cov}(y,d) = \frac{1}{n}\sum_i y_i d_i - \bar{y}\bar{d}$ und

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

6. Zeigen Sie, dass in einer Regression auf eine Dummy Variable $y_i = b_1 + b_2 d_i + \varepsilon_i$ 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 (mit $d_i = 0$) und im Steigungskoeffizienten die Differenz der Mittelwerte beider 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 diesen Daten 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, dass die ceteris paribus Interpretation nur in Bezug auf die im Modell berücksichtigten Regressoren gilt! In der einfachen Regression $\widehat{\text{StdL}}_i = 12.5 + 2.5m_i$ greift deshalb keine ceteris paribus Annahme, wir erhalten als Resultat die einfachen unbedingten Mittelwerte wie in Abbildung 2.20 dargestellt.

Wenn wir aber die Bildungsjahre zusätzlich berücksichtigen erhalten wir das in Abbildung 2.21 dargestellte Resultat, *bei gleicher Bildung* verdienen Männer um 2.95 Euro *mehr* als Frauen. Aufgrund der unterstellten linearen Funktionsform gilt dies für jedes Bildungsniveau.

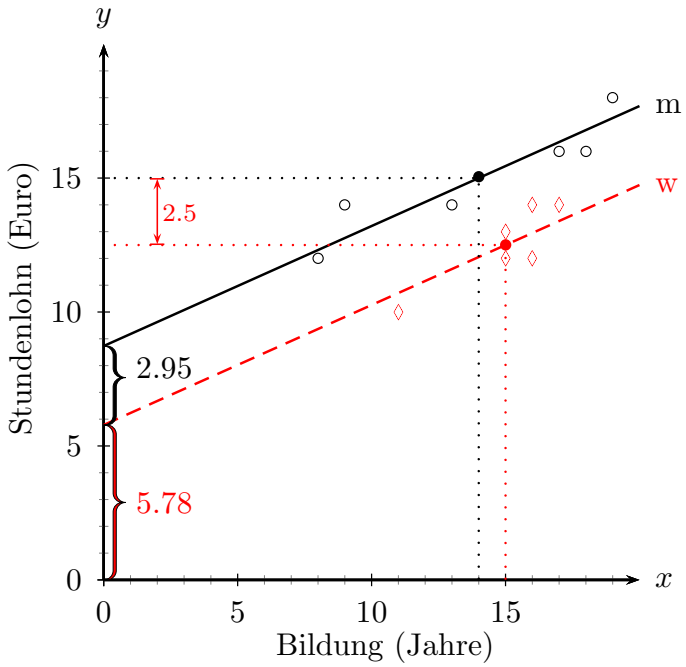


Abbildung 2.21: Unterschiede im Interzept; $\widehat{\text{StdL}} = 5.78 + 2.95\,m + 0.45\,\text{Bildg}$
Der einfache Lohnunterschied zwischen Männern und Frauen beträgt 2.5 Euro, aber dieser berücksichtigt nicht, dass hier die durchschnittliche Bildungsdauer von Frauen um ein Jahr länger ist. Der ceteris paribus Unterschied (d.h. bei gleicher Bildung) beträgt 2.95 Euro.

In diesem Beispiel beträgt die durchschnittliche Bildungsdauer für Frauen 15 Jahre und für Männer nur 14 Jahre. Wenn wir diese Werte einsetzen erhalten wir die einfachen Unterschiede. Wenn wir aber für Männer die Bildungsdauer von Frauen einsetzen sehen wir, dass der *ceteris paribus* Unterschied 2.95 Euro beträgt!

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

$$\widehat{\text{StdL}}|(m = 0 \;\&\; \text{Bildg} = 15) = 5.78 + 2.95 \cdot 0 + 0.45 \cdot 15 = 12.5$$

$$\widehat{\text{StdL}}|(m = 1 \;\&\; \text{Bildg} = 14) = 5.78 + 2.95 \cdot 1 + 0.45 \cdot 14 = 15$$

$$\widehat{\text{StdL}}|(m = 1 \;\&\; \text{Bildg} = 15) = 5.78 + 2.95 \cdot 1 + 0.45 \cdot 15 = 15.45$$

Beim einfachen Mittelwertvergleich $\widehat{\text{StdL}}_i = 12.5 + 2.5m_i$ haben wir also die mittleren Stundenlöhne einer Personengruppe mit einer mittleren Bildungsdauer von 15 Jahren mit denen einer Personengruppe mit einer mittleren Bildungsdauer von 14 Jahren verglichen, also unter sonst *nicht* gleichen Bedingungen.

Da hier der durchschnittliche Lohn mit der Bildung steigt, und Frauen im Durchschnitt *mehr* Bildung haben als Männer, *unterschätzt* ein einfacher Mittelwertvergleich den ceteris paribus Lohnunterschied.

Die multiple Regression mit Berücksichtigung der Bildung erlaubt einen ceteris paribus Vergleich: wie groß *wäre* der Unterschied bei *gleicher Bildung*? Man beachte, dass die konkreten Zahlen auch auf der *angenommenen* linearen Funktionsform beruhen.

Alternativ kann man dies auch als Beispiel für die Nichtberücksichtigung relevanter Variablen interpretieren. Erinnern wir uns an das Kapitel über die Nichtberücksichtigung relevanter Variablen zurück; dort haben wir gezeigt, dass zwischen dem Steigungskoeffizienten eines 'kurzen' und 'langen' Modells folgende Beziehung besteht:

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

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

Im kurzen Modell 'fehlt' hier 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 wenig überraschend

$$b_2^k = b_2^l + b_3^l a_2 = 2.95 + 0.45 \times (-1) = 2.5$$

Durch einen einfachen Mitetlwertvergleich wird das durchschnittlich höhere Bildungsniveau der Frauen gewissermaßen verschleiert, vgl. Tabelle 2.5 (Seite 44). *Bei gleicher Bildung* (ceteris paribus) wäre die Lohndifferenz mit 2.95 Euro deutlich größer als der einfache Unterschied von 2.5 Euro!

Durch diese Spezifikation haben wir zwar zugelassen, dass sich das Interzept zwischen Männern und Frauen unterscheiden kann, aber wir haben a priori unterstellt, dass Bildung für Männer und Frauen die gleichen Auswirkungen auf den Stundenlohn hat; diese Spezifikation lässt keine unterschiedlichen marginalen Effekte der Bildung für Männer und Frauen zu! Der mit dieser Spezifikation gemessene marginale Effekt der Bildung beträgt 0.45 Euro, d.h. mit jedem zusätzlichen Bildungsjahr steigt der mittlere Stundenlohn für Männer *und* Frauen gleichermaßen um 0.45 Euro (vgl. Abbildung 2.21).

Dieser für beide Gruppen gleiche marginale Effekt 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.

Im Beisopiel mit den Stundenlöhnen

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

Ein solches Produkt zweier Regressoren nennt man einen *Interaktionseffekt*.

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 unterschiedliche 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 2.22), 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

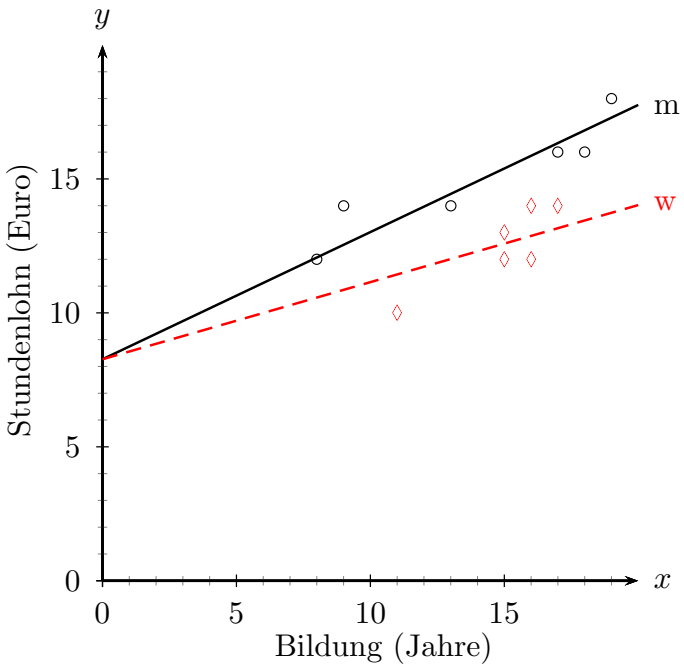


Abbildung 2.22: Unterschiede in der Steigung;
$\widehat{\text{StdL}} = 8.27 + 0.29\,\text{Bildg} + 0.19(m \times \text{Bildg})$

auch eine Interaktionsvariable zwischen Dummy Variable und metrisch skalierten x Variable verwenden

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

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

$$\begin{aligned} \text{für } m=0: \quad \widehat{y}^0 &= a_1 + a_2 x \\ \text{für } m=1: \quad \widehat{y}^1 &= c_1 + c_2 x \end{aligned}$$

mit $c_1 = b_1 + b_3$ und $c2 = b_2 + b_4$.[15]
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 2.23.

[15] 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.

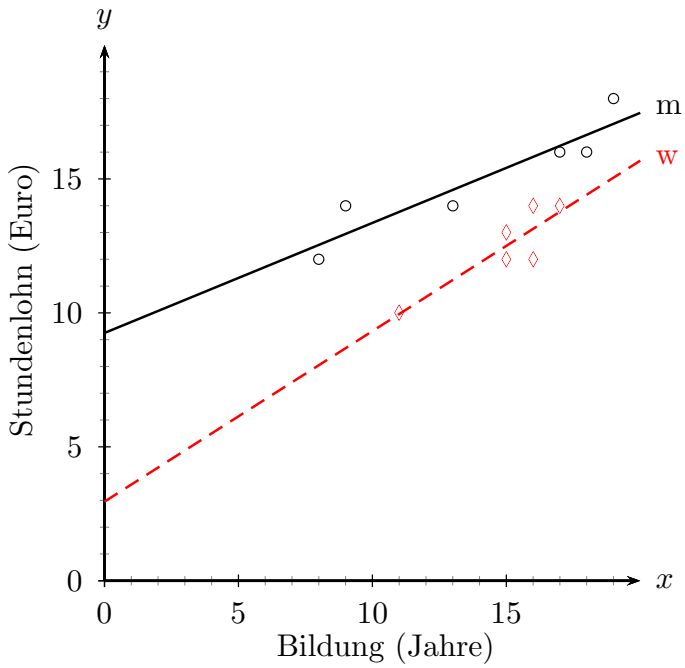


Abbildung 2.23: Unterschiede im Interzept und Steigung;
$\widehat{\text{StdL}} = 2.95 + 6.30m + 0.64\,\text{Bildg} - 0.23(m \times \text{Bildg})$

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.

2.7.4 Dummy Variablen für mehrere sich nicht überlappende Kategorien

Wir haben bereits gesehen, dass das Interzept einer Regression auf eine Dummy Variable d den Mittelwert der Referenzkategorie ($d_i = 0$) angibt, und der Steigungskoeffizient den mittleren Unterschied zwischen den Kategorien misst.

In manchen Fällen werden erklärende Variablen verwendet, die nur eine kleine Anzahl von Ausprägungen haben, z.B. die Nationalität von Touristen (nominalskaliert), Schulnoten (ordinalskaliert), oder die Anzahl der Kinder einer Frau.

Kehren wir noch einmal zurück zu dem Beispiel mit Gebrauchtautos, in dem wir den Preis mit dem Alter der Autos erklären. Um das Prinzip zu erklären runden wir das Alter wieder auf ganze Jahre, die Variable AlterJ hat deshalb nur die Ausprägungen $\{0, 1, 2, 3, 4, 5\}$.

In Spalte (1) von Tabelle 2.8 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 20) 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 ist es empfehlenswert, für jede Ausprägung der kategorialen Variable eine eigene Dummy Variable anzulegen, eine davon als Referenzkategorie zu wählen, und alle anderen Dummy Variablen als Regressoren zu verwenden.

Im Beispiel mit den Gebrauchtautos bilden wählen wir z.B. Autos mit einem 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 2.8 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.

Vergleichen Sie dies wieder mit Tabelle 2.3 (Seite 20); die Koeffizienten der Dummy Variablen messen den Unterschied im mittleren Preis zur Referenzkategorie AlterJ = 0.

Spalte (3) von Tabelle 2.8 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 das durchschnittliche Alter der Autos mit dem betreffenden Alter, wie wieder ein Vergleich mit Tabelle 2.3 (Seite 20) 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: Tabelle 2.8 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="")
```

Tabelle 2.8: Drei verschiedene Spezifikationen für die Preise von Gebrauchtautos, Alter gerundet auf ganze Jahre

	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]

In Stata erhält man einen vergleichbaren Output mit[16]

```
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
```

□

Eine Anwendung für Panel Daten: 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.

[16] Allerdings wird in Stata bei der Regression ohne Interzept (Option `nocons`) die erste Kategorie unterdrückt.

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 Durchschnitte über die Zeit zu bilden, und über diese Durchschnittswerte 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 erhalten wir das so genannte *'between'* Modell (deshalb der hochgestellte Index b)

$$\bar{y}_{i\bullet} = b_1^b + b_2^b \bar{x}_{i\bullet} + \bar{e}_{i\bullet}^b$$

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

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

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 *gepooltes Modell* genannt (deshalb der hochgestellte Index p).

Dieses gepoolte Modell kann mit OLS geschätzt werden, indem die Daten entsprechend angeordnet werden: man 'stapelt' einfach die Beobachtungen für die einzelnen Individuen übereinander (engl. *'stack'*).[17]

Das '*stacked model*' für 3 Individuen und 4 Zeitperioden würde in Vektorschreibweise

[17] Diese Anordnung der Daten muss natürlich nicht manuell erfolgen, aller Programme verfügen über spezielle Befehle für die Umorganisation der Daten. In Stata gibt es dafür die Befehle `xtset` und `reshape`; für R existieren zwei Packages `reshape` und `reshape2`, mit den sehr flexiblen Befehlen `melt` und `dcast`, für Paneldaten wird das Package `plm` verwendet. Die entsprechenden Befehle für EViews sind `pagestack`, `pageunstack` und `pagestruct`.

folgendermaßen aussehen ($i = 1, \ldots, 3$, $t = 1, \ldots, 4$)

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

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

Ein etwas allgemeineres und flexibleres Modell könnte individuenspezifische Interzepte zulassen, aber für alle Länder den gleichen Steigungskoeffizienten unterstellen. Dies kann einfach mit Hilfe entsprechender Dummy Variablen bewerkstelligt werden. Wir würden z.B. das folgende Modell schätzen

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{24} \\ y_{31} \\ y_{32} \\ y_{33} \\ y_{34} \end{pmatrix} = b_1 \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + b_2 \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + b_3 \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + b_4 \begin{pmatrix} x_{11} \\ x_{12} \\ x_{13} \\ x_{14} \\ x_{21} \\ x_{22} \\ x_{23} \\ x_{24} \\ x_{31} \\ x_{32} \\ x_{33} \\ x_{34} \end{pmatrix} + \begin{pmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{14} \\ e_{21} \\ e_{22} \\ e_{23} \\ e_{24} \\ e_{31} \\ e_{32} \\ e_{33} \\ e_{34} \end{pmatrix}$$

wobei das erste Individuum als Referenzkategorie dient.

Dieses Modell wird auch *'Least Squares Dummy Variable'* (LSDV) Modell genannt und wird üblicherweise kürzer als

$$y_{it} = b_1 + a_i + b_4 x_{it} + e_{it}$$

notiert, wobei die a_i symbolisch für die Individueneffekte stehen (z.B. Ländereffekte), also die Individuen-Dummies symbolisieren.

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 elegante Alternative. Erinnern wir uns an das 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 den interessierenden Steigungskoeffizienten b_4 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 nur auf Dummy Variablen 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 füe 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.

Anstatt eine Regression mit potentiell mehreren tausend Dummy Variablen zu berechnen genügt es also eine individuenspezifische Mittelwerttransformation durchzuführen, und eine einfache OLS Regression auf die derart transformierten Daten zu rechnen

$$(y_{it} - \bar{y}_{i\cdot}) = b_4(x_{it} - \bar{x}_{i\cdot}) + (e_{it} - \bar{e}_{i\cdot})$$

wobei $\bar{y}_{i\cdot}, \bar{x}_{i\cdot}$ und $\bar{e}_{i\cdot}$ Gruppen-Mittelwerte sind.

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 natürlich zu exakt den gleichen Schätzungen für die Steigungskoeffizienten wie das LSDV Modell, ist aber viel einfacher zu berechnen.[18]

Aus diesem Grund können wir die Koeffizienten des *'fixed effects'* Modells gleich interpretieren wie die Koeffizienten eines Dummy Variablen Modells (LSDV).

Man verliert bei dieser *'fixed effects'* Methode zwar die Individueneffekte a_i (d.h. die Koeffizienten der Individuendummies), aber diese sind ohnehin selten von Interesse.[19]

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, ...).

[18] 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 und geben die korrekten Standardfehler aus.

[19] Außerdem könnten die Individueneffekte einfach aus den Daten berechnet werden, $a_{i\cdot} = \bar{y}_{i\cdot} - b_4\bar{x}_{i\cdot}$.

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. Viele 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 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.

Dies ist natürlich ein extremes Beispiel, wenn wir uns 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.

2.7.5 Dummy Variablen für mehrere sich überlappende Kategorien

Im vorhergehenden Abschnitt Kategorien untersuchten wir sich gegenseitig ausschließende Kategorien, jede Beobachtung war in genau einer Kategorie; ein Auto kann z.B. nicht gleichzeitig zwei und vier Jahre alt sein. Aus diesem Grund waren die entsprechenden Dummy Variablen unkorreliert.

Nun sehen wir uns Fälle an, die sich gegenseitig nicht ausschließen, z.B. kann eine Person weiblich sein und als weiteres Merkmal verheiratet oder nicht verheiratet sein.

Kehren wir nochmals zurück zum Beispiel mit den Stundenlöhnen, siehe 2.7 (Seite 53). 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 Steigungskoeffizient zeigt, dass Verheiratete durchschnittlich um 0.6 Euro mehr verdienen,

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

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 schätzen wir ein zu kurzes Modell, und es tritt wieder das Problem der Nichtberücksichtigung relevanter Variablen 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 (Referenzkategorie) 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

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

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.

Im früheren Fall sich gegenseitig ausschließender Kategorien (d.h. nicht überlappender Kategorien) sind die Produkte der Dummy Variablen per Definition gleich Null (z.B. $m \times w$), deshalb ist in diesem Fall das Modell mit den einfachen Dummies bereits das gesättigte Modell. Aus diesem Grund waren die Koeffizienten der einfachen Dummies bereits die Mittelwerte der entsprechenden Kategorien.

2.7.6 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.

2.7.7 Stückweise lineare Funktionen*

Stückweise lineare Funktionen (*piecewise linear functions*) sind der einfachste Fall von *Spline Funktionen.*[20]

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.

Anstelle dreier einzelner Gleichungen kann alternativ auch eine Gleichung mit Dummy Variablen und Interaktionstermen geschätzt werden.

Dazu definieren wir zwei Dummy Variablen

$$\begin{array}{lll} D_1 & = 1 & \text{wenn } x \geq x^{*1} \quad \text{und 0 sonst;} \\ D_2 & = 1 & \text{wenn } x \geq x^{*2} \quad \text{und 0 sonst;} \end{array}$$

[20] 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."

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 2.24 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 2.24 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 $a_2 + b_2 + c_2$.

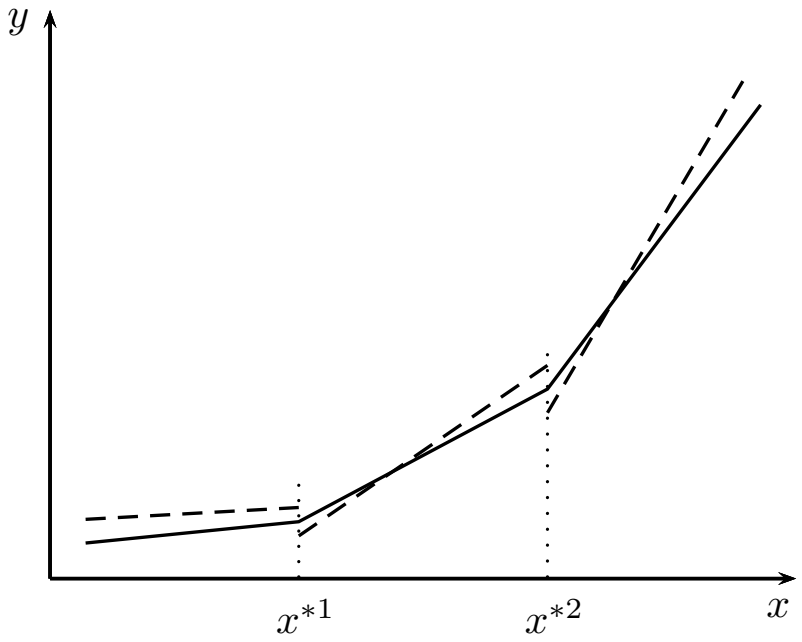


Abbildung 2.24: Einzelregressionen (strichliert) und stückweise lineare Regression (durchgezogene Linie).

2.8 Diverses

2.8.1 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.

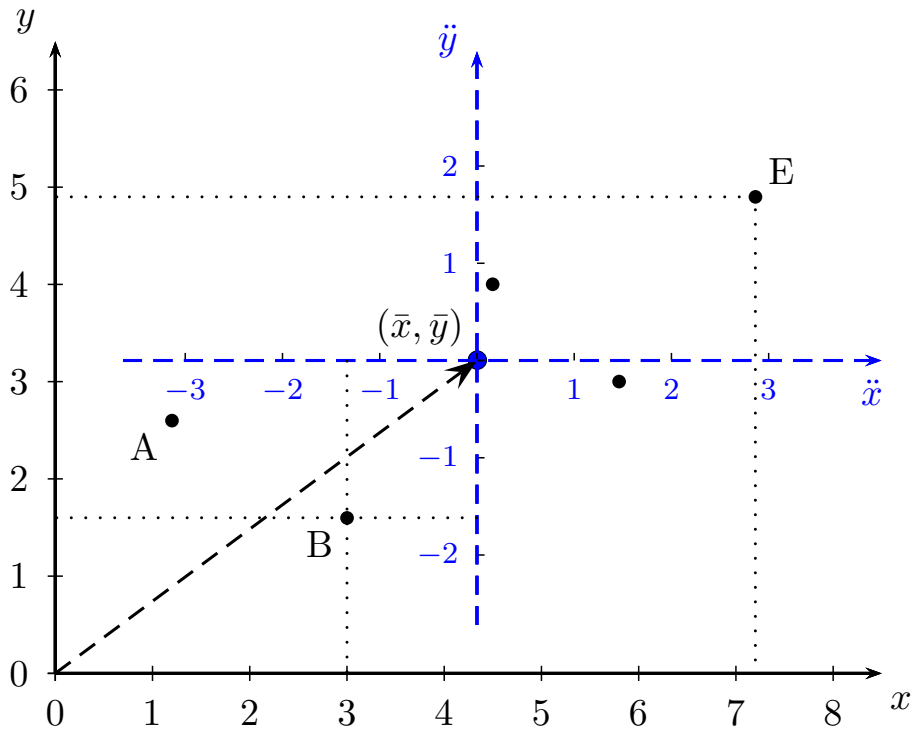
Die Mittelwerttransformation besteht einfach darin, dass von jeder einzelnen Beobachtung x_i einer Datenreihe der Mittelwert der selben Datenreihe $\bar{x}$ subtrahiert wird.

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 2.25 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 mittel-



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 2.25: 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]

wertransformierten 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.

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. Wir werden später zeigen, dass dieses Ergebnis als ein Spezialfall des Frisch-Waugh-Lovell Theorems interpretiert werden kann.

2.8.2 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 2.26 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.

2.8.3 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[21] publiziert wurde. Legendre suchte nach einer Methode, wie ein Gleichungssystem mit mehr Gleichungen als Unbekannten gelöst

[21] "Nouvelles méthodes pour la détermination des orbites des comètes." Paris 1805, Anhang: "Sur la Méthode des moindres quarrés", S. 72-80 .

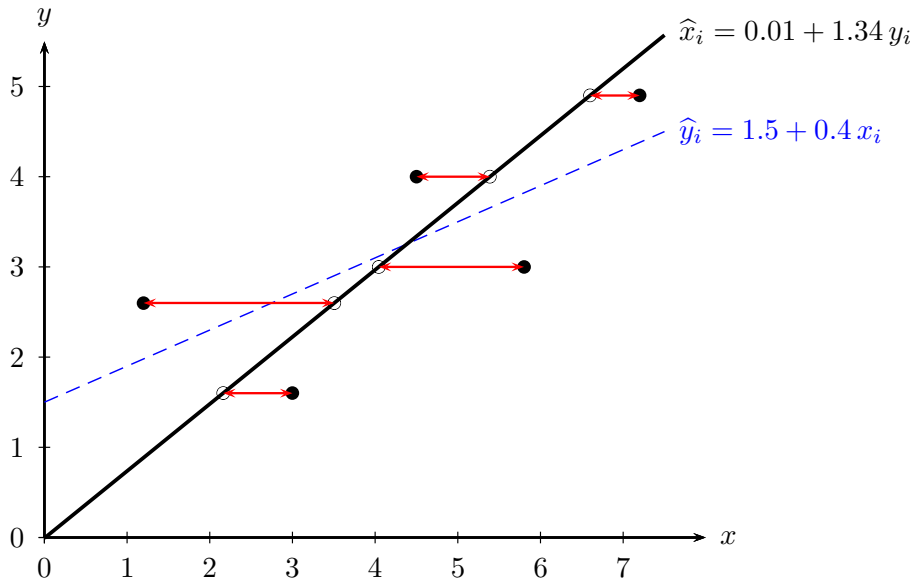


Abbildung 2.26: *'Reverse Regression'*: die Regression von x auf y ($\hat{x}_i = b_1^* + b_2{}^* y_i$), sowie strichliert die normale Regression $\widehat{y}_i = b_1 + b_2 x_i$.

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. Singh, 2010).

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. Vor dem Hintergrund des Burenkrieges (1899 – 1902), der einen Mangel an tauglichen Rekruten zutage brachte, untersuchte Galton den Zusammenhang zwischen zwischen der Körpergröße von Söhnen und Vätern; die Körpergröße der Soldaten war damals von größerer militärischer Bedeutung als heute.

Galton glaubte einen negativen Zusammenhang zwischen der Körpergröße von Vätern und Söhnen zu finden, eine *"regression towards the mean"*. Die der Analyse zugrunde liegende statistische Technik wurde in der Folge als 'Regression' bekannt.

Dieses Ergebnis beunruhigte die damaligen Eliten zutiefst, und führte zu Ängsten

vor Degeneration und langfristigem Niedergang der imperialen Größe. Es zeigte sich allerdings, dass Galtons Sorgen unbegründet waren, und dieses Phänomen ging als *"Galton's Fallacy"* in die Literatur ein. Später mehr davon ...

Literaturverzeichnis

Fox, J. (2005), 'The R Commander: A basic-statistics graphical user interface to R', *Journal of Statistical Software* **19(9)**, 1–42.

Frisch, R. and Waugh, F. V. (1933), 'Partial time regressions as compared with individual trends', *Econometrica* **1**(4), pp. 387–401.
URL: *http://www.jstor.org/stable/1907330*

Lovell, M. C. (1963), 'Seasonal adjustment of economic time series and multiple regression analysis', *Journal of the American Statistical Association* **58**(304), pp. 993–1010.
URL: *http://www.jstor.org/stable/2283327*

Lovell, M. C. (2008), 'A Simple Proof of the FWL Theorem', *The Journal of Economic Education* **39**(1), 88–91.
URL: *http://www.tandfonline.com/doi/abs/10.3200/JECE.39.1.88-91*

Machlup, F. (1974), 'Proxies and dummies', *The Journal of Political Economy* **82**(4), 892.

Singh, R. (2010), 'Development of Least Squares: A Survey', *The IUP Journal of Computational Mathematics* **3**(1), 54–84.