Contents

Chapter 2

Basics of descriptive regression analysis – the algebra of OLS

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

2.1 Introduction

Statistics is generally concerned with methods for collecting and evaluating quantitative information. Traditionally, a distinction is made between descriptive and inductive statistics. While the aim of descriptive statistics is often a condensation of information for given data, inductive statistics is mainly concerned with possible conclusions from an observed sample to a non-observable population.

Regression analysis can also be used for both purposes. Although it is used in econometrics almost exclusively in the sense of inductive statistics, we start here with descriptive regression analysis. The reason for this is mainly didactic, this allows us to separate the more technical aspects from the somewhat more abstract concepts of stochastic regression analysis; this is to allow for the easiest possible introduction to the subject.

We will argue that descriptive regression analysis can be seen more or less as a generalisation of the method for calculating simple means. Beyond that, however, regression analysis additionally allows us to present the relationship between two or more variables in a compact way.

This is what this chapter will be about. After a few general considerations, we will get to know the technique with the help of which we can calculate the coefficients of a linear regression. Building on this, we will look at interpreting the results before generalising the technique to more than two variables and examining a few important special cases.

All later chapters build directly on these simple concepts, so it is worth looking at these basics a little more closely.

[0]Translated with `www.DeepL.com/Translator` (free version)
German version: `https://www.uibk.ac.at/econometrics/einf/kap02_ols.pdf`

2.2 Linear relationships

> *"Nothing is so firmly believed as what we least know."* (Michel de Montaigne, 1533–1592)

One of the central tasks of econometrics is the *'measurement of relationships'*. To do this, the interrelationships of interest must first be formally represented. This is done with the help of mathematical functions.

A *function* $y = f(x)$ is essentially an 'input-output' relationship, it yields the value of an *dependent* variable y for given values of the explanatory variable x, or in the case of several explanatory variables $y = f(x_1, x_2, \ldots, x_k)$, where f denotes the functional form and k the number of explanatory variables.

For the time being, we will restrict ourselves to the very simplest case, linear functions with only one explanatory variable x.

$$y = b_1 + b_2 x$$

Here b_1 and b_2 stand for simple numbers describing the linear relationship between the variables x and y.

If we draw this function in a coordinate system we get a straight line. The *intercept* b_1 indicates the point of intersection with the vertical y axis (ordinate), i.e. it measures the value of y at the point $x = 0$. The coefficient b_2 of the explanatory x variable measures the slope of the straight line, and is therefore unsurprisingly called *'slope'*. For linear functions, the slope coefficient b_2 is equal to the derivative

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

and indicates by how many units y changes when x increases by one unit. Therefore, b_2 measures the *marginal effect* of a change from x to y.

Because of the linear functional form, this applies here not only to infinitesimal changes but also to discrete changes, i.e., if x increases by one unit ($\Delta x = 1$), then y *changes* by b_2 units ($\Delta y = b_2$, or, $\Delta y / \Delta x = b_2$). An important advantage of the linear functional form is that the *marginal effect* b_2 does not depend on x, i.e., it can be used for *all* values of x.

2.2.1 Exact and 'imprecise' relations

Although such simple linear relationships may at first seem like a caricature of a complex reality, they occur frequently in everyday life.

For example, when we fill up a car, we know that the amount to pay is the product of the price and the number of litres filled up. If we denote the amount to be paid by y and the number of litres filled up by x, the relationship between x and y is described exactly by the function $y = b_1 + b_2\, x$ (for $x \geq 0$).

Here, the slope coefficient b_2 denotes the price, i.e. if we fill up with an *additional* litre, the amount to be paid increases by b_2 euros. From the intercept b_1 we know

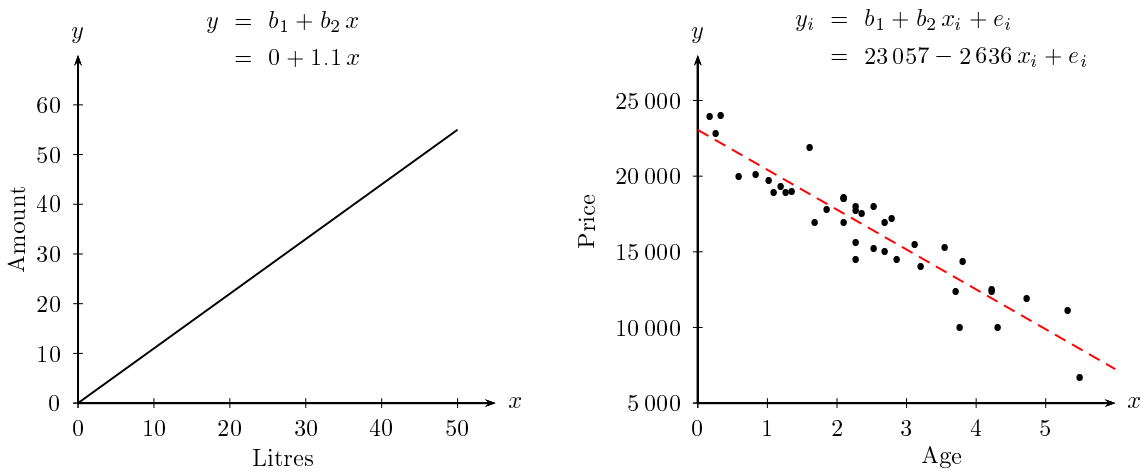


Figure 2.1: Left panel: an exact relation between liters refueled and amount to pay for a price $b_2 = 1.1$ Euro. Right panel: an 'inexact' correlation between the age of used cars and their price

Table 2.1: Prices (in euros) and age (in years) of 40 used cars (age_rd is age rounded to whole years);
`http://www.hsto.info/econometrics/data/auto40.csv`

Obs.	price	age	age_rd	km	Obs.	price	age	age_rd	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

that it must be equal to zero in this example, because if we fill up zero litres ($x = 0$) we don't have to pay anything either ($y = 0$), so the function starts at zero. This function is shown graphically in the left panel of Figure 2.1 for a price $b_2 = 1.1$.

The right panel of Figure 2.1 shows a different relationship, the relationship between the age of used cars of a specific type and their price. Each dot shows age and price for a specific used car, in total the 40 dots represent age and prices of 40 different cars (the underlying data are reflected in Table 2.1). Obviously, the 'average' price decreases with age, but the relationship no longer holds exactly.

This has various causes, on the one hand the cars differ in other characteristics not shown here (mileage, equipment, colour, ...), but also sellers and their motives, the location and much more, which can differ from observation to observation.

Nevertheless, it is clearly recognisable that older cars are *'on average'* cheaper, and that this relationship can be *approximated* relatively well by the dashed straight line shown in Figure 2.1.

How can we write such 'approximate' relations in general? We could, using the '$\approx$' sign (*'is approximate'*), write $y \approx b_1 + b_2 x$, but it is impossible to do arithmetic with '$\approx$'. Therefore we need a more suitable form of representation. The solution is simple, we introduce a 'rest', so called *'residuals'*, which should capture all other (unobserved) influencing factors. For these residuals we use the symbol e.

These residuals e will of course differ from observation to observation (i.e. here from car to car), so for each observation we need a separate equation

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

where n denotes the total number of observations.

Since this would be a bit cumbersome to write, it is usually written shorter in the following form

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

where i denotes the running index and n the number of observations. Sometimes one also writes $i \in \mathbb{N}$, i.e., the index i is an element of the natural numbers $\mathbb{N}$.

The residual e_i takes the value that is necessary for equation i to be exactly fulfilled. If we rewrite the above equation to $e_i = y_i - b_1 - b_2 x_i$, we see that there is a direct connection between the residuals e_i and the coefficients b_1 and b_2.

Two important notes are in order at this point:

1. only the expressions of the variables y_i and x_i are observable (i.e. in our example price and age of the used cars), the coefficients b_1 and b_2 as well as the residuals e_i are *not* directly observable.

2. only the expressions of the variables y_i, x_i as well as the residuals e_i differ between observations, the coefficients b_1 and b_2 are supposed to hold for all observations, so they are *not* observation-specific, or in other words, we assume them to be *constant*. We can imagine that the coefficients b_1 and b_2 of the linear function describe, in a sense, the relationship behind the data. Whether a value is observation-specific or not can often be seen from the subindex i, only observation-specific values have a subindex i.[1]

   Note that we have not claimed the coefficients b_1 and b_2 are *'in reality'* constant, this is an *assumption* that will allow us to calculate the two unknown coefficients from the data in the first place.

Note that the above equations would be rather meaningless if no restrictions were imposed on the residuals. Without these restrictions, any value of b_1 and b_2 would be compatible with the observed values of y_i and x_i! Therefore, the restrictions for the residuals, which are necessary to achieve the 'best possible' fit, play an essential role in all the following explanations and in econometrics in general. We will show that the main restrictions required are that the mean over all residuals is zero, and that the correlation between the residuals and the x variable is zero. More on this later!

In the following, we will look at how we can calculate the two coefficients b_1 and b_2 of the linear function $y_i = b_1 + b_2 x_i + e_i$ from the observed data y_i and x_i with $i = 1, \ldots, n$, because this allows us a very compact description of the data in the sense of descriptive statistics, similar to how the mean value provides a compact summary of a single data series.

In the car example, the straight line equation approximates the observations relatively well, but it is also clear that this approximation only gives satisfactory results for a certain range of x. For a 10-year-old car, for example, the regression line would yield a negative price. Price increases for vintage cars can of course not be represented at all by this straight line. This means that the relationship between age and price is not actually linear.

But as this example shows, even non-linear relationships can often *over a limited range* of variables be approximated relatively well by a linear function.

Intercept and regression constant We have so far referred to both b_1 and b_2 as coefficients, although b_1 is at least not 'visibly' multiplied by a variable. We can, however, imagine that b_1 is multiplied by a vector of ones, as is clear from the following vector notation

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

[1] Caution, the indices 1 and 2 of the coefficients b_1 and b_2 have a different meaning than the index i.

Table 2.2: Alternative labels for y and x of the function $y = b_1 + b_2x$

y	x
left-hand side variable	right-hand side variable
dependent variable	[independent variable]
explained variable	*explanatory variable*
regressand	*regressor*
response variable	covariate
effect variable	control variable

The unity vector is often called 'regression constant' in this context, and the intercept b_1 is simply the coefficient of the regression constant.[2]

Alternative terms for y and x When estimating a regression equation $y_i = b_1 + b_2x_i + e_i$ one also says that y is regressed on x. For the variables y and x a number of different names have been used in the literature, some of which are summarised in Table 2.2.

In the following we will mostly refer to y as *dependent variable* and x as *explanatory variable*. One should not take the term 'explanatory' too literally, because this does not necessarily mean that y is 'explained' by x; with this method we can at best show that there is a linear relationship between y and x, but the method alone in no way provides us with a substantive 'explanation' for this relationship, and of course certainly no indication of a possible causal relationship between y and x. Nevertheless, in the following we will stick to the designations *dependent* and *explanatory* variable, because they have become common in the literature.

The explanatory variables x are also often called regressors, while the term regressand for y is not quite as common.

Especially in statistics, the explanatory variables are often called *covariates*, in more technical contexts the term *control variables* is also common for the x variables.

In older textbooks, the x variable is also often called an *'independent variable'*. While the term 'dependent variable' for y is quite appropriate and common, the term 'independent variable' for x can be misleading, as this could be confused with 'statistical independence', which is a entirely different concept. Therefore, calling x an independent variable is generally discouraged.

In the next section we will learn a method that allows us to calculate the coefficients b_1 and b_2 from the observed values of the variables x and y in such a way that the relationship between x and y is described 'as well as possible'.

2.3 The OLS Method

"If you want to build tall towers, you have to spend a long time at the foundation." (Anton Bruckner, 1824–1896)

[2]The literature is unfortunately sometimes a bit confusing in this respect, in some older textbooks the terms 'intercept' and 'regression constant' are also used synonymously.

The term acronym OLS stands for *'Ordinary Least Squares'*. Our concrete concern in this section will be to find a formula into which we can insert the observed data y and x, and which gives us as a result *'best possible'* numerical values for the not directly observable coefficients b_1 and b_2 of a straight line equation $y_i = b_1 + b_2 x_i + e_i$. What exactly is meant by 'best possible' will be explained later, but we will see that the OLS method solves exactly this problem.

We begin our considerations with a mental decomposition each element y_i of the dependent variable y into two parts, into a *systematic component* $b_1 + b_2 x_i$, in which the relationship underlying the data is expressed in the form of a straight line equation, and into the remainder, i.e. the *residuals* e_i

$$y_i = \underbrace{b_1 + b_2 x_i}_{\substack{\text{systematic} \\ \text{component } \widehat{y}_i}} + \underbrace{e_i}_{\substack{\text{resi-} \\ \text{duals}}}$$

Let us illustrate this decomposition by looking at Figure 2.2. The top panel shows 5 data points and an imaginary straight line that fits these observation points 'as best as possible'. We will call this straight line *'regression line'* in the future. Assuming we already had this regression line, we could use it to decompose each observed y_i into two parts, into a value that lies exactly *on* the regression line, $\widehat{y}_i$ (pronounced y_i roof), and into the difference between this $\widehat{y}_i$ lying on the regression line and the actual observed value y_i. This difference is, of course, the residual e_i, i.e. $y_i = \widehat{y}_i + e_i$ (with $\widehat{y}_i = b_1 + b_2 x_i$) for $i = 1, \ldots, n$. The lower panel in Figure 2.2 shows this decomposition.

We call the exactly *on* the regression line 'fitted' values $\widehat{y}_i$ *systematic component*, that is, the component described by the variable x and the coefficients b_1 and b_2.

But for the calculation of these 'fitted' values $\widehat{y}_i$ we need, in addition to the x variable, the (for the time being) unknown coefficients b_1 and b_2

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

A 'good' regression line should satisfy two conditions:

1. the proportion of the 'systematic' component should be as large as possible, implying that the residuals should provide as small an explanatory contribution as possible;

2. this requires that the correlation between the 'systematic' component and the residuals should be as small as possible. We will show a little later that the OLS method gives us exactly such values for b_1 and b_2 that guarantee that the correlation between the 'systematic' component and the residuals is exactly zero.

To actually calculate the coefficients, one could come up with the idea of choosing the values b_1 and b_2 in such a way that the sum of all residuals $\sum_i e_i$ becomes as small as possible.

However, this would lead to positive and negative deviations cancelling each other out when summing. One can simply show that the sum of residuals is always zero

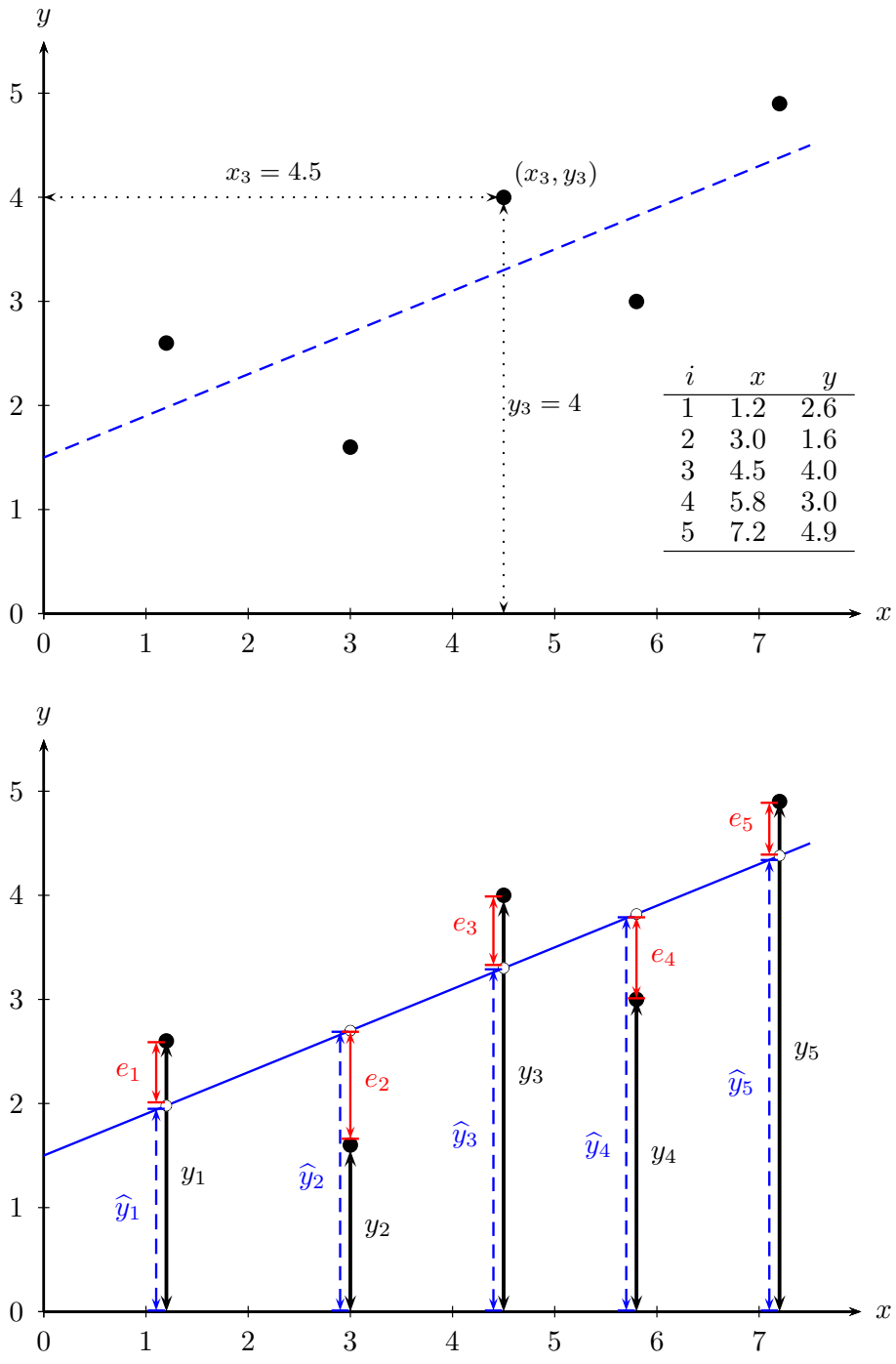


Figure 2.2: Decomposition of y_i into a systematic component $\widehat{y}_i$ and an unsystematic residual e_i (for $i = 1, \ldots, 5$).

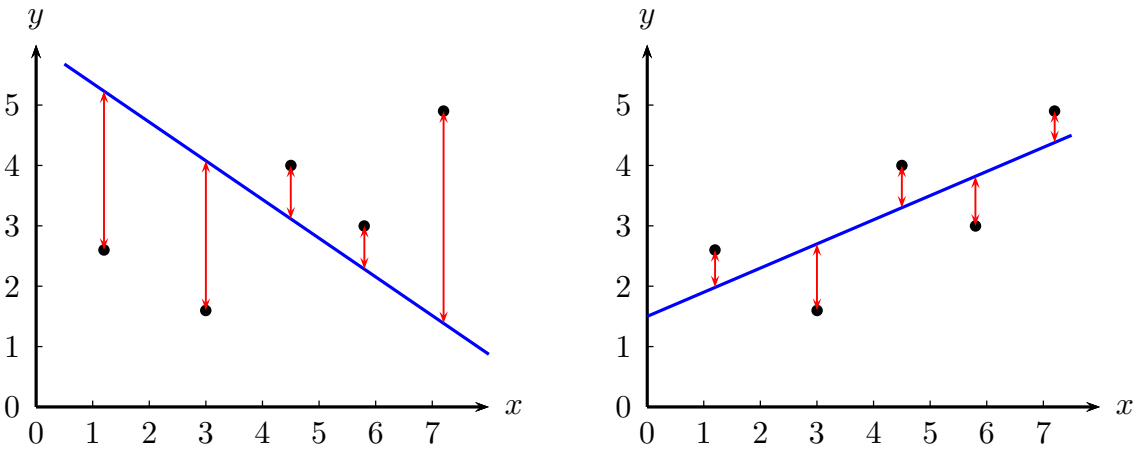


Figure 2.3: The sum of the deviations $\sum_i e_i = \sum_i (y_i - \widehat{y}_i)$ has the same value in both Figures, since positive and negative values cancel each other out.

for every straight line that is laid through the mean values of x and y. Therefore, this method is unsuitable to obtain an approximation.

Figure 2.3 illustrates the problem: the *sum of deviations* $\sum_i e_i$ has the same value in the left and right graphs, although the straight line in the right graph obviously approximates the points much better.

This problem could be avoided by minimising the absolute value of the deviations. However, this raises two problems: First, this problem is much more difficult to solve numerically, and second, it does not disproportionately weight large deviations more heavily than small deviations. In fact, people are often risk-averse and will prefer to see large errors 'punished' disproportionately more than small errors.

The simplest solution to these problems is to choose the coefficients b_1 and b_2 such that the *sum of squared deviations* (i.e. $\sum_i e_i^2$) is minimised. This is exactly the principle of the OLS method.

This also explains the name **Method of Ordinary Least Squares**.

This rather simple basic idea of the OLS method can be easily explained with the help of Figure 2.4. Note that the function $y_i = b_1 + b_2 x_i + e_i := \widehat{y}_i + e_i$ can be rewritten as $e_i = y_i - \widehat{y}_i$. In Figure 2.4, the squares of the residuals $e_i^2 = (y_i - \widehat{y}_i)^2 := (y_i - b_1 - b_2 x_i)^2$ are plotted. In a thought experiment, we can rotate and shift the straight line of this Figure, that is, change the values of b_1 and b_2, until the *sum* of the drawn square areas becomes as small as possible. The values of b_1 and b_2 that provide the *smallest possible sum of square areas* are the OLS coefficients we are looking for.

This thought experiment provides a good intuition, but this approach is hardly suitable for practical work. We need a general method that allows us to calculate the unobservable coefficients b_1 and b_2 from the observable data x and y, and we will now derive such a formula.

Before we begin, a quick note. You may wonder what the point of all this 'calculating' is, if the finished formulas themselves are already implemented even in Excel and are very easy to use. Well, we will see in the following chapters that the application

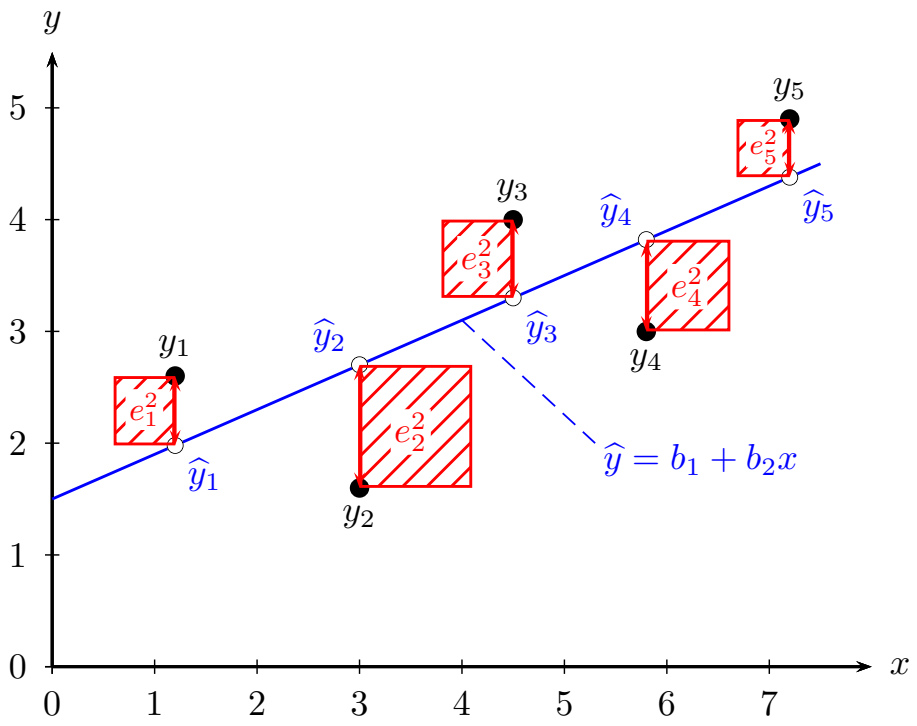


Figure 2.4: According to the OLS method, b_1 and b_2 are chosen in such a way that the *sum of squared deviations* becomes as small as possible, i.e. the total area of the shaded squares is minimised.

of this formula only leads to the desired results under very specific conditions. An understanding of the mechanics of the OLS method will also allow us to understand the limitations of this approach and to take appropriate action in a further step if the assumptions are violated. A naive application of these methods often leads to misleading or at least unnecessarily inaccurate results. To avoid such mistakes, a sound understanding of the basics is required, and for such an understanding, a bit of arithmetic is sometimes surprisingly useful.

We can easily represent the relation between the area of a square and the two coefficients b_1 and b_2 by rewriting $y_i = b_1 + b_2 x_i + e_i$

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

The area of a single shaded square in Figure 2.4 is $e_i^2 = (y_i - b_1 - b_2 x_i)^2$, and the area of *all* squares is simply the sum over $i = 1, \ldots, n$

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

What we are looking for are the values of b_1 and b_2 for which the *sum of all areas* – i.e. the sum of squared residuals $\sum_i e_i^2$ – is minimal, so the minimization problem is

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

where the b_1 and b_2 under the 'min' statement should indicate that these are the two quantities we are looking for.

The rest is simple arithmetic. We derive partially according to the unknown coefficients b_1 and b_2, set these two derivatives equal to zero. This gives the first-order conditions, or necessary conditions for a minimum.[3] The derivatives are[4]

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

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

As can be seen, these first-order conditions imply

(*'first order conditions'*, FOC)

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

These two conditions are of utmost importance, the solutions not only define the two coefficients we are looking for, they will be encountered again and again later, because the essential properties of the OLS method follow directly from these two conditions!

The first of these first-order conditions, $\sum_i e_i = 0$, follows from the derivative after the intercept b_1, i.e. it is only valid if the regression equation contains an intercept. The second condition follows from the derivative with respect to the slope coefficient b_2 and ensures – together with the first condition – that the covariance between x and e is always zero.[5]

The coefficients b_1 and b_2 we are looking for are the solutions of the minimisation problem and therefore guarantee that these two first-order conditions are satisfied! The simple structure – only the minimum of a quadratic function is determined – ensures that the solution is unique.

Now we finally want to calculate the two unknown coefficients b_1 and b_2 from the two first-order conditions. To do this, we reshape them a bit, noting that we can

[3] One can show that the second-order conditions, i.e. sufficient conditions for a minimum, are satisfied as well.

[4] For the derivatives we need the chain rule, i.e. if $y = f(z)$ and $z = g(x)$ it follows $y = f[g(x)]$ and the derivative is

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

[5] $\sum_i x_i e_i = \sum_i e_i(x_i - \bar{x} + \bar{x}) = \sum_i e_i(x_i - \bar{x}) + \bar{x}\sum_i e_i = \sum_i e_i(x_i - \bar{x}) = \sum_i (e_i - \bar{e})(x_i - \bar{x}) = n\,\mathrm{cov}(e, x) = 0$ since $\sum_i e_i = 0$ and $\bar{e} = 0$ if the regression contains an intercept (in the first step, only the constant $\bar{x}$ is subtracted and added).

draw 'anything without subindex i' in front of the sum sign, and that $\sum_i b_1 = nb_1$ because b_1 is a constant

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

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

These are the so-called normal equations, which we solve for the coefficients b_1 and b_2.

To do this, we multiply the first equation by $\sum x_i$ and the second equation by n (note that $\sum x_i$ is a simple number).

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

and subtract the first equation from the second one to eliminate b_1

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

from which follows

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

This is exactly the function we are looking for. On the right-hand side, only the observable x_i and y_i occur. If we substitute the observations into this formula, we get as a result the value of the slope coefficient b_2, which minimises the sum of squared residuals!

Once b_2 is calculated the intercept b_1 can be easily calculated, we divide both sides of the normal equation (2.4) by n and get

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

It is common to denote the mean of a variable with a bar above the variable name, e.g. $\bar{y}$ (pronounced y bar) for the mean of y. Of course, $\bar{y} := \frac{1}{n} \sum_{i=1}^{n} y_i$, where the symbol ':=' is read as 'is defined'. Note that the means are not observation-specific, and therefore have no subindex i (means are simple numbers).

Using this notation for the means, we obtain for the intercept

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

These two OLS formulas above already solve our problem, but the formula for the slope coefficient (2.6) looks a bit 'unappetising'. Fortunately, this formula can be represented much more simply using variances and covariances.

We recall that the *empirical variance* – a descriptive measure of dispersion for given observations – and the *empirical covariance* – a descriptive measure of the relationship between two variables – are defined as[6]

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

With the help of these definitions, the OLS coefficients can be written more simply as

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

where the equation for the intercept was taken from equation (2.7). Note that this only applies to regressions with intercept!

Proof:* To show that

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

we divide the numerator and denominator of the middle expression of equation (2.6) by n and get

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

and consider that the mean of x or y is defined as $\bar{x} := \frac{1}{n}\sum_i x_i$, respectively $\bar{y} := \frac{1}{n}\sum_i y_i$.

Thus the above expression can be written as

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

Then we add and subtract from the numerator $n\bar{x}\bar{y}$ and from the denominator $n\bar{x}^2$. This gives

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

[6]Note that this is the population variance var^p. In contrast, the sampling variance is defined as $\text{var}(x) := \frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2$. The following relationship holds for both definitions.

Next, we rewrite the definition of the means a little, from $\bar{x} = \frac{1}{n}\sum_i x_i$ follows $n\bar{x} = \sum_i x_i$ or $n\bar{y} = \sum_i y_i$, and put this in

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

pull out the sum sign

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

and factorise

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

This already looks much simpler. The formula is even easier to remember if we divide the numerator and denominator by n (or $n-1$), because then we see that equation (2.6) can be written more simply as the ratio of empirical covariance to empirical variance

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

□

Attn.: If all x_i have the same numerical value b_2 is not defined, since for constant x $\mathrm{var}(x) = 0$! We will see later that this is a special case of *perfect multicollinearity* (i.e. x is a multiple of the regression constant).

Example 1: Figures 2.2 and 2.4 are based on the following data:

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

Using the OLS formulae we found earlier, we can now calculate the coefficients b_1 and b_2 that minimise the sum of squared residuals.

To do this, we extend the table by the columns xy and x^2 and form the respective sums:

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

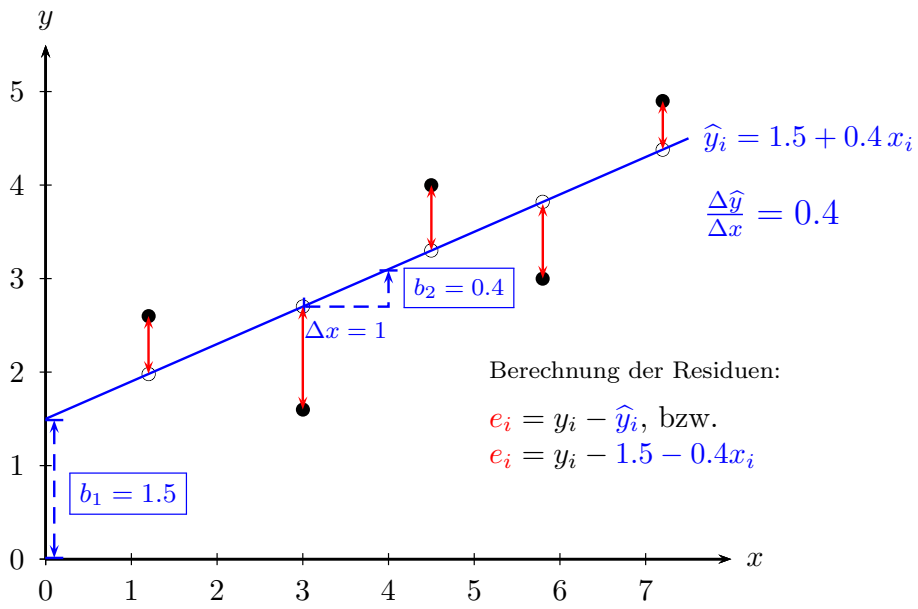


Figure 2.5: Example

If we substitute in equations (2.6) and (2.7) we get

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

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

The regression equation plotted in Figure 2.5 is thus

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

or using the alternative formula (2.8) for mean transformed data

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

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

Example 2: In this example we show that the usual mean can also be derived by the OLS method, namely by a regression on the regression constant only.

Let

$$y_i = b_1 + e_i$$

The residuals in this case are $e_i = y_i - b_1$. The OLS method is based on minimising the sum of squared residuals, i.e.

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

. Deriving by the unknown coefficient b_1 and setting this derivative to zero gives the value of b_1 that minimises the sum of squared residuals

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

from which follows

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

So an OLS regression on the regression constant actually yields the arithmetic mean, so you can consider the mean as a special case of an OLS estimator!

Example 3: We have suggested on several occasions that the OLS method is 'optimal' in a certain sense, without specifying more precisely what this optimality refers to. In this exercise we will show that the fitted values $\widehat{y}_i$ calculated by the OLS method have a very special property, namely that the dispersion around these OLS fitted $\widehat{y}_i$ is smaller than the dispersion around all other values $\widetilde{y}_i$ calculated by any other *linear* function.

This is analogous to the mean of a variable, because we know from the mean $\bar{x}$ that it minimises the sum of squared deviations (or the empirical variance), i.e. for any number z it holds

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

Why?

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

since $\sum_i (x_i - \bar{x}) = \sum_i x_i - n\bar{x} = n\bar{x} - n\bar{x} = 0$ (note $\bar{x} := \frac{1}{n}\sum_i x_i \Rightarrow \sum_i x_i = n\bar{x}$).

Because $\sum_i(\bar{x}-z)^2 > 0$ for $\bar{x} \neq z$, $\sum_i(x_i-\bar{x})^2 < \sum_i(x_i-z)^2$ must hold.

Show that the fitted values $\widehat{y}_i$ calculated by the OLS method also have this property.

To do this, compare the $\widehat{y}_i = b_1 + b_2 x_i$ calculated with the OLS coefficients b_1 and b_2 with the fitted values of any other linear function $\widetilde{y}_i = c_1 + c_2 x_i$ and prove that

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

Solution: To show this we proceed analogously as above.

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

The first two terms on the right-hand side are quadratic and can therefore never become negative. Let us therefore first look at the third term $2\sum_i(y_i-\widehat{y}_i)(\widehat{y}_i-\widetilde{y}_i)$, taking into account that $y_i-\widehat{y}_i := e_i$ are the OLS residuals.

So

$$\begin{aligned}\sum_i(y_i-\widehat{y}_i)(\widehat{y}_i-\widetilde{y}_i) &= \sum_i e_i(\widehat{y}_i-\widetilde{y}_i) \\ &= \sum_i e_i\left[(b_1+b_2x_i)-(c_1+c_2x_i)\right] \\ &= \sum_i\left[(b_1-c_1)+(b_2-c_2)x_i\right]e_i \\ &= (b_1-c_1)\underbrace{\sum_i e_i}_{=0} + (b_2-c_2)\underbrace{\sum_i x_i e_i}_{=0} \\ &= 0\end{aligned}$$

since the two first order conditions $\sum_i e_i = 0$ and $\sum_i x_i e_i = 0$ hold for the OLS residuals (see equations (2.2) and (2.3), page 11).

So it follows

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

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

This should not be too surprising, since we derived the OLS coefficients by minimising the sum of squared residuals ;-) ■

More examples:

1. Calculate the OLS formula for regression without intercept, i.e. for the model $y_i = bx_i + e_i$.
2. Show that $\sum_i (x_i - \bar{x}) = 0$.
3. Show that $\sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_i (x_i - \bar{x})y_i$.

2.4 Conditional means

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

We have now learned a method to calculate the two coefficients b_1 and b_2 that are not directly observable from observed data, without really justifying why we need them. In this section we will make up for this and provide a more intuitive insight into how we can interpret the fitted values $\widehat{y}$ and the coefficients. We will expand on these insights in the next section on the multiple regression model and go deeper in the dummy variables section, and they provide us with the foundations for understanding the stochastic regression model in the next chapter.

Recall that the OLS method is primarily a *decomposition method*, a variable of interest y is decomposed into a systematic component $\widehat{y}$ and a non-systematic component – the residuals e.

For the interpretation, we are only interested in the *systematic* component, since we can learn little from the residuals in a proper specification.

The *systematic* component is

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

or for the earlier used car example $\widehat{\text{price}}_i = 23\,057 - 2\,636\,\text{age}_i$ (see Figure 2.1, page 3), with the price here measured in euros and the age in years.

The *systematic component* is simply the fitted price, and this is described (or 'explained' within a model) by a *linear function* as a function of age.

For a deeper understanding, we will now go into two questions in a little more detail, namely.

1. what is a intuitive interpretation of the systematic component $\widehat{y}_i$, and
2. what is the significance of the (linear) functional form?

We will argue below that we can interpret linear regression simply as *linear approximation to the conditional means*.

To do this, we will return to the example with the used cars, but for the time being we apply a little trick: we round the explanatory variable 'age' to whole years in

order to obtain several observations for each year. This turns the continuous variable 'age' into a discrete variable, which we call 'age_rd'; in this example, the variable 'age_rd' takes an integer value between 0 and 5, i.e. age_rd $\in \{0, 1, 2, \ldots, 5\}$ (see Table 2.1, page 3).[7]

Table 2.3 shows the same observations as Table 2.1, but arranged differently, grouped by rounded age (age_rd). For age_rd = 0 (i.e. $0 <$ age ≤ 0.5), for example, there are three observations. Since we now have several observations for each rounded age, we can *calculate the mean values for each age level*; the average price for the three cars with age_rd = 0 is, for example, 23 567 euros.

Table 2.3: Car prices by rounded age. $\bar{y}$ denotes the arithmetic mean by age group and $\widehat{y}$ the fitted values of the regression $\widehat{y}_i = 22\,709 - 2\,517x_i$. (The numbers are taken from Table 2.1 (page 3), they are just arranged differently here).

	age_rd = 0	age_rd = 1	age_rd = 2	age_rd = 3	age_rd = 4	age_rd = 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	
i		19700	18600	15000	12350	
c		19300	18500	16900	12500	
e		19000	18500	15500	12350	
s			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

In the following we call the mean value for an age group an *conditional* mean, we write

$$(\overline{\text{price}}|\text{age\_rd} = 0) = 23\,567$$

and read this as: Mean price, *given* age is zero years.

If we do this for all ages we get the *conditional mean function*, each age group 'age_rd' is assigned a conditional mean value

[7] In this case we use the variable 'age_rd' to form *age categories*, such variables are therefore called *categorical* variables; we will discuss these in more detail in the section on dummy variables.

$$(\overline{\text{price}}|\text{age\_rd}) = \begin{cases} 23567 & \text{for age\_rd} = 0 \\ 19409 & \text{for age\_rd} = 1 \\ 17696 & \text{for age\_rd} = 2 \\ 15781 & \text{for age\_rd} = 3 \\ 12403 & \text{for age\_rd} = 4 \\ 9900 & \text{for age\_rd} = 5 \end{cases}$$

compare Table 2.3 row $\overline{y}$.

This allows – in the sense of descriptive statistics – a 'condensation' of the information from Table 2.3, instead of the 40 observations we have only 6 mean values, one for each age category.

With the help of this conditional mean function we can easily see that average prices fall with age, e.g. by 4158 euros in the first year, by 1713 euros in the second year, etc., see line $\Delta\overline{y}$ $(:= \overline{y}_j - \overline{y}_{j-1}$, with $j = 1, \ldots, 5)$ in Table 2.3.

We achieve even greater 'information compression' if we apply the OLS method to the 40 observations from Table 2.1.

For the rounded explanatory variable 'age_rd' we obtain

$$\widehat{\text{price}}_i = 22\,709 - 2\,517\text{age\_rd}_i$$

For example, for cars with age_rd = 4 we get the fitted value $\widehat{\text{price}}|(\text{age} = 4) = 22\,709 - 2\,517 * 4 \approx 12641$, and analogously the fitted values for the other age classes (rounded), see also line $\widehat{y}$ in Table2.3

$$(\widehat{\text{price}}|\text{age\_rd}) = \begin{cases} 22709 & \text{for age\_rd} = 0 \\ 20192 & \text{for age\_rd} = 1 \\ 17675 & \text{for age\_rd} = 2 \\ 15158 & \text{for age\_rd} = 3 \\ 12641 & \text{for age\_rd} = 4 \\ 10124 & \text{for age\_rd} = 5 \end{cases}$$

compare Table 2.3 row $\widehat{y}$.

Note that due to the linear functional form, the *change* $\Delta(\widehat{\text{price}}|\text{age\_rd}) = b_2 = 2\,517$ is constant.

To calculate these values, we only need the two OLS coefficients b_1 and b_2, so we achieve even greater 'information compression', albeit at the expense of accuracy. This is the usual *'trade off'* between information compression and accuracy that we will encounter more often.

Figure 2.6 shows the underlying data, the conditional averages and the values fitted using the OLS method.

Obviously, the conditional means (i.e. means by age category) and the OLS-fitted values are very close, in some cases so close that they partially overlap in the figure.[8]

[8]This need not always be the case, but in this example the observations are relatively well approximated by a linear function.

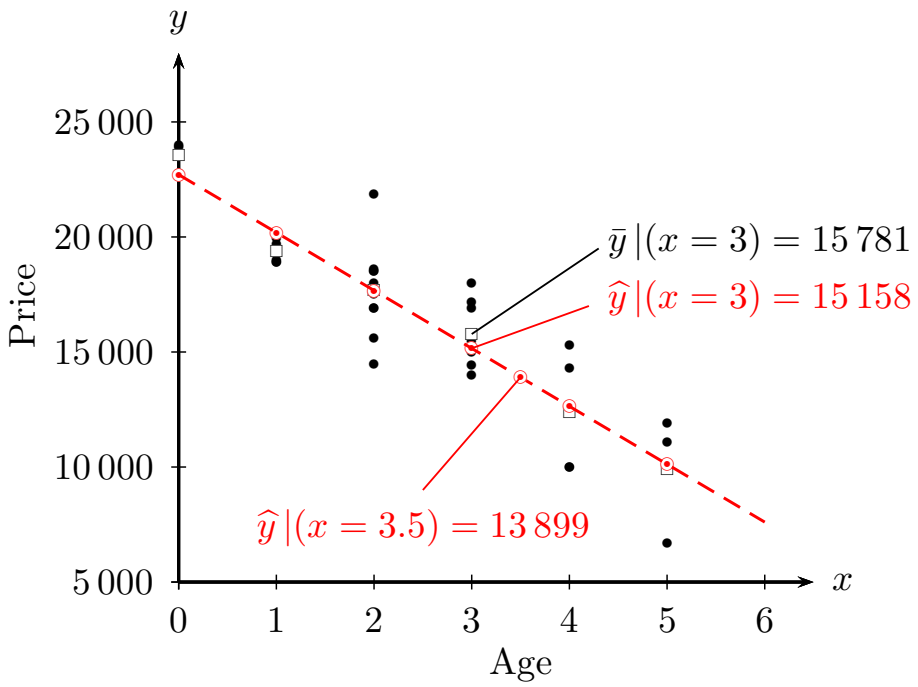


Figure 2.6: Descriptive regression as a linear approximation to the 'conditional mean function'. (• observations; □ conditional means; ⊙ linear approximation).

Intuitively, we can think of the fitted values $\widehat{y}$ lying on the regression line as *linear approximation to the conditional means*. We will elaborate on this interpretation later when we discuss dummy variables. Here it serves mainly as a preparation for the stochastic regression analysis, where we will interpret the $\widehat{y}$ similarly as a linear approximation to the *conditional expected values*.

Next we turn to the linear functional form. Given the assumed linear functional form $\widehat{y} = b_1 + b_2 x$ we can calculate $\widehat{y}$ for any x, for example in our example we can calculate the fitted price $\widehat{y}_i$ for a car with age 3.5 years: $(\widehat{y}|x = 3.5) = 22\,709 - 2\,517 \times 3.5 \approx 13\,899$, although there is not a single car with an age of 3.5 years in this data set. Nevertheless, we can think of $(\widehat{y}|x = 3.5) = 13\,899$ as a linear approximation to the (hypothetical) average price of cars with an age of 3.5 years. Note, however, that this interpretation is based on the *assumed* linear functional form that made this interpolation possible.

This intuition remains valid even if we do not have repeated y observations for manifestations of the x variable, such as in the original example from Figure 2.1 (page 3).

In this sense, in the descriptive regression analysis we will present the fitted values $(\widehat{y}|x = \underline{x})$ generally as a linear approximation to the conditional means for x, where $\underline{x}$ denotes a concrete possible expression of x (e.g. age_rd $= 3.5 = \underline{x}$)

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

where here $\overset{\text{lin}}{\approx}$ stands for 'linear approximation'.

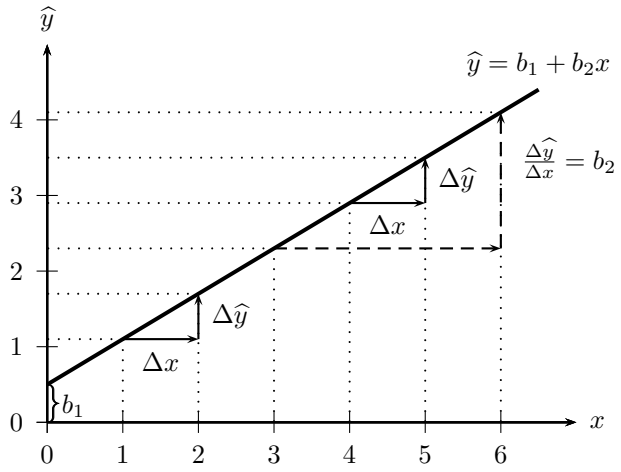


Figure 2.7: Linear function $\widehat{y} = b_1 + b_2 x = 0.5 + 0.6x$. An increase of x by one unit is accompanied by a change of $\widehat{y}$ by $+0.6$ units.

Since it would be extremely awkward to speak of a 'linear approximation to the conditional mean' each time, we will simply speak of a 'mean' price or average price in the future, but it is important to note for later discussion that we interpret the fitted values $\widehat{y}_i$ as a linear approximation to the conditional means.

In most cases we are interested in how a change in x 'on average' affects y, for example, by how many euros the 'average' price of used cars falls if the age increases by one year.

Using the OLS method, we can answer this question at least for a linear approximation to the conditional mean of y, because the first derivative (i.e. the differential quotient $d\widehat{y}/dx$) of the regression function[9] gives us the answer we want, the slope coefficient b_2.

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

This first derivative is often referred to as *'marginal effect'*, where the term 'marginal' refers to an infinitesimally small change in x.

For linear functions, however, it does not matter whether we consider infinitesimally small or discrete changes, the *marginal effect* in this case is equal to the slope coefficient b_2, and thus constant over the entire course of the function

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

but of course this only applies to linear functional forms (cf. Figure 2.7).

The slope coefficient b_2 thus tells us that an increase in x by one unit is accompanied by a change in $\widehat{y}$ by b_2 units, whereby we can interpret $\widehat{y}$ in the descriptive regression analysis as a linear approximation to the conditional mean.

[9] We omit the subindex i here, because the linear approximation does not only apply to the observed x_i, but because we can, at least in principle, calculate an associated $\widehat{y}$ for each x; of course, this is usually done only for $x_{\min} \leq x \leq x_{\max}$ makes sense

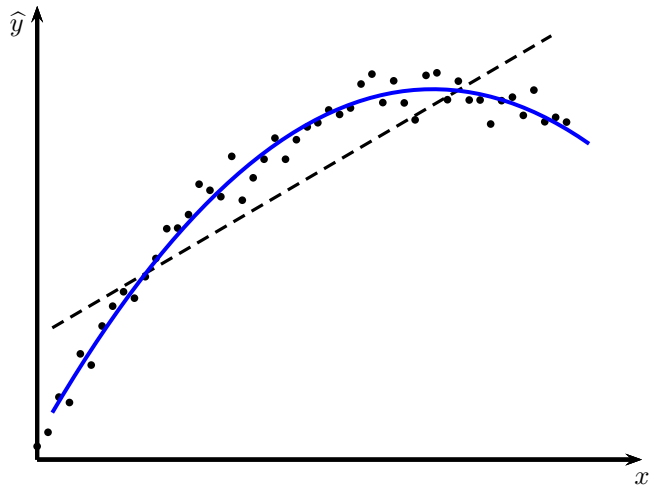


Figure 2.8: A linear function $\widehat{y} = b_1 + b_2 x$ can give a very poor fit if the actual relationship is non-linear. Obviously, in this case, a non-linear function like the blue line would give a much better fit, but for non-linear functions the marginal effect (slope of the tangent) is not constant, i.e. is different for each x.

$$\widehat{\text{price}} = 22\,709 - 2\,517\,\text{age\_rd} \qquad \rightarrow \qquad \frac{d\,\widehat{\text{price}}}{d\,\text{age\_rd}} = 2\,517$$

For a correct interpretation of these regression coefficients it is essential to have the dimensions of y and x: in this example $\widehat{y}$ is measured in euros and x in years, i.e. if age increases *by one year* the average price $\widehat{y}$ decreases by 2 517 *Euro*.

It is also important to stress that such a regression per se does not tell us anything about a possible causal relationship, it merely describes an association. The mere data can never tell us anything about a possible cause-effect relationship, this would be an interpretation far beyond mere description. In a later chapter on *endogeneity* we will discuss the possibility of causal statements in more detail, and we will see that causal statements always require special justification.

Note also that with the OLS method we have assumed *a priori* a linear functional form, and that the interpretation of the coefficients follows directly from this assumed functional form.

In the example with the used cars, the conditional means were very well approximated by a linear function, but of course this need not always be the case.

Figure 2.8 shows data points that are obviously described much better by a non-linear function than by the dashed simple regression line.

In this very special case, the points can be well described by a quadratic function $\widehat{y} = b_1 + b_2 x + b_3 x^2$, and we will see later that such functions can also be easily calculated with the OLS method. However, even in this simple case, the marginal effect is no longer constant, but changes with x; if we derive the quadratic function

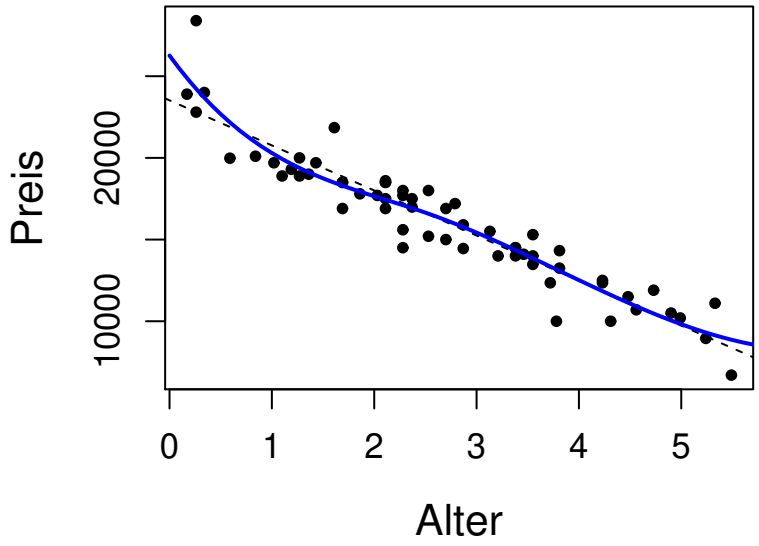


Figure 2.9: Spline function for the prices of used cars

with respect to x we obtain

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

i.e., the marginal effect (the slope of the tangent) is different for each x in this example.

In addition, there are estimation procedures for more complex forms of non-linearities, e.g. spline functions. Figure 2.9 shows such a non-linear estimation for the car example.

Obviously, this function can depict the data 'more accurately', one can see, for example, that the 'conditional mean price' falls more in the first year than in the later years. However, this 'more exact' description also has costs, the 'information compression' is clearly smaller, also the marginal effects differ for all manifestations of x, and can therefore no longer be simply stated.

Here again a more general principle becomes visible, there is a *'trade-off'* between the accuracy of the description and the 'information compression', or 'simplicity'.

The greater simplicity is often achieved by more restrictive assumptions (e.g. linearity of functional form). In most cases, this simplicity has the advantage of making the results easier to interpret, but this advantage usually comes at a cost in terms of accuracy.

The optimal degree of data compression depends on the purpose of use; if you want to drive from Rome to Hamburg by car, you need a road map with a different scale than if you are looking for the nearest hotel in Rome.

In general we can say

$$\text{data} + \text{theory(assumptions)} \rightarrow \text{conclusions}$$

There is no data analysis that is completely devoid of theory and the assumptions underlying the theory. Even for the calculation of a simple mean, it must be clarified beforehand 'what' is to be counted, or in other words, a classification must be made. As a rule, stronger assumptions permit more far-reaching conclusions, but the extent to which these are then correct depends largely on the extent to which the assumptions were correct. Therefore, we should always be very aware of the assumptions on which our analysis is based and what the consequences are if the assumptions are violated.

In the used car example, the linear functional form assumption fits obviously fairly well for ages 0 – 5, but the same assumption would obviously yield fairly nonsensical fitted prices for 10 year old used cars.

Representation of regression equations and standard errors

In textbooks, regression equations are often given in the following form

$$\begin{array}{rcccc} \text{Price} & = & 22\,709.303 & - & 2\,517.267\ \text{age\_rd} \\ & & (532{,}689)^{***} & & (190{,}125)^{***} \\ & & & & \\ & & R^2 = 0.822, & & n = 40 \end{array}$$

Below the coefficients in brackets are the *standard errors of the coefficients*. Their significance will only become fully apparent in the stochastic regression analysis, and in the chapter on the properties and of OLS estimators we will discuss their calculation and exact interpretation in detail.

The standard errors of the coefficients can be interpreted as an indicator of the precision of the estimate of the coefficient. In essence, they allow us to assess whether there is a signal in the noise of the data that stands out strongly enough from the null hypothesis to be taken seriously.

To do this, you can remember as a rough rule of thumb that for $n > 50$ the null hypothesis can be rejected with an error probability of 5% if the coefficient is at least twice as large as his standard error.

This is the case for both coefficients above, and is also symbolically expressed by the asterisks next to the standard errors. As we will show in detail in the chapter on hypothesis testing, one star symbolises that the null hypothesis that the 'true' coefficient (of an unobserved population) is equal to zero can be rejected at a significance level of 10%, two stars imply a significance level of 5%, and three stars imply a significance level of 1%. All this will be discussed in detail later.

The above notation in rows is ill-suited when many regressors are used, or when several regression equations are to be compared. Therefore, a tabular representation is almost exclusively chosen in the literature. Table 2.4 shows in column (1) the same regression output as above in tabular form.

	Dependent variable:	
	price	
	(1)	(2)
Constant	22,709.300***	23,056.710***
	(532.689)	(468.871)
age_rd	−2,517.267***	
	(190.125)	
age		−2,635.669***
		(166.935)
Observations	40	40
R^2	0.822	0.868
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 2.4: Representation of regressions in tabular form, here in column (1) with 'age_rd' rounded to whole years (cf. above regression in row form) , and in column (2) with age not rounded. This plot was generated using the R-package `stargazer`

2.5 Coefficient of determination: R^2

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

The regression line can describe the data – more or less well depending on the nature of the data.

Figure 2.10 shows two extreme cases, in the left panel the points are very close to the regression line, i.e. the *'fit'* is very good and the data is well described by the regression line – the loss of information is rather small when the data is described by the regression line. In contrast, the data in the right panel are less well described by the regression line, i.e. the 'fit' is poor. In the second case, if one *exclusively* knows the regression line, one gets only a poor idea of the underlying data – the loss of information when describing the data by a regression line is large.

.

It would be practical if we had a simple measure of how 'good' the fit of the regression line is to the observation points. Such a measure for the goodness of fit does exist, namely the coefficient of determination R^2.

We will show in a moment that the coefficient of determination can be interpreted as the share of the dispersion of y explained by x in the total dispersion of y.

Since it is a proportion, the coefficient of determination R^2 for ordinary regressions with intercept can only assume values between zero and one. The better the 'fit',

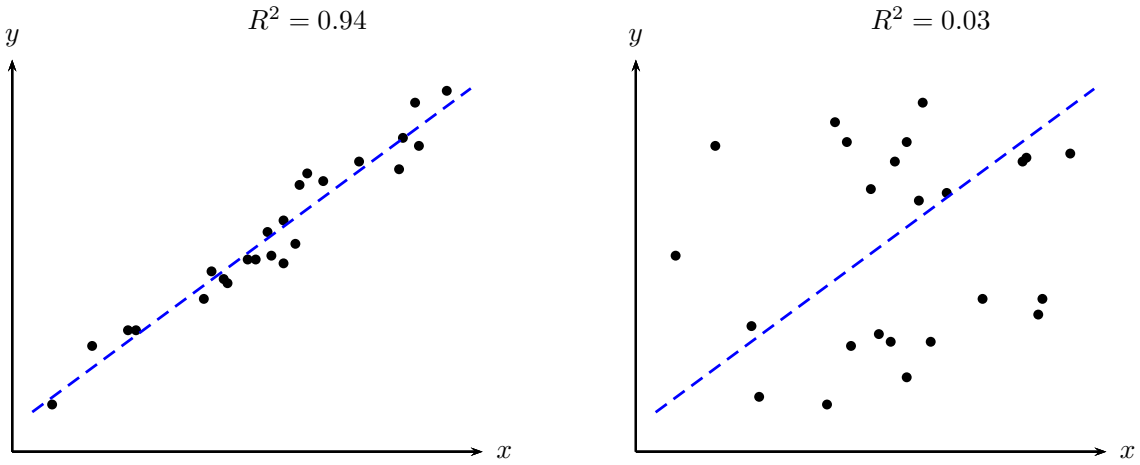


Figure 2.10: The relationship between two variables can be described more or less well by a regression line.

the closer the coefficient of determination is to one. The left panel of Figure 2.10 shows a relatively good 'fit' with a coefficient of determination of $R^2 = 0.94$. If the coefficient of determination takes the value one ($R^2 = 1$), the observation point lie exactly on the regression line. Conversely, the worse the 'fit', the closer the coefficient of determination is to zero. The right panel in Figure 2.10 shows a very poor 'fit' with a coefficient of determination of $R^2 = 0.03$.

The coefficient of determination is most simply interpreted as a descriptive measure to assess the *'goodness of fit'* of the regression lines to the observation points.

Essentially it is based on a scatter decomposition, we decompose the total scatter of y into an 'explained' and an 'unexplained' part; Figure 2.11 shows the idea.

First, note that a regression line with intercept always passes through the mean of x and y.

This follows directly from the first order conditions and can be easily shown by substituting the mean $\bar{x}$ into the equation for the fitted values $\widehat{y}_i = b_1 + b_2 x_i$, i.e.

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

where $\widehat{y}_{\bar{x}}$ denotes the value of $\widehat{y}$ for $\bar{x}$.

If the regression line passes through the point $(\bar{x}, \bar{y})$ then $\widehat{y}_{\bar{x}} = \bar{y}$ must be. This is indeed so, to see this we put the OLS formula for the intercept $b_1 = \bar{y} - b_2\bar{x}$ into the equation above and get

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

Note that this only holds if the regression contains an intercept, because we used $b_1 = \bar{y} - b_2\bar{x}$ here to show this.

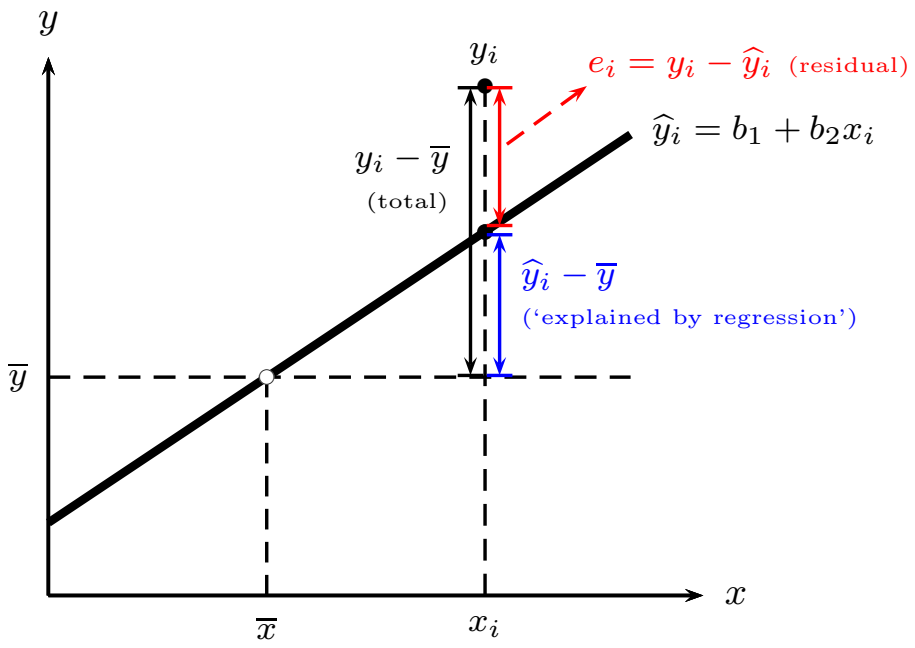


Figure 2.11: Decomposition of the total dispersion of y into an 'explained' and an 'unexplained' part.

Let us come back and remember that the OLS method is primarily a decomposition method, it helps us to decompose a variable y_i into a systematic component $\widehat{y}_i$ and the unsystematic 'rest' e_i.

For example, suppose there is a positive correlation between height x and weight y. This correlation is of course not exact, you know the story of the span-long Hansel and the noodle-thick Dirn, but at least on average we expect taller people to weigh more.

What is the best estimate for a person's weight if we don't know that person's height? Exactly, the average weight of all persons $\bar{y}$, or in other words, the weight of a person with average height $\bar{x}$, because we have just shown that the regression line always passes through the point $(\bar{x}, \bar{y})$. If the person actually has weight y_i we make the error of $y_i - \bar{y}$.

Suppose we now learn that this person is 190 cm tall. In this case we will use this information to revise our estimate, $\widehat{y}_i = b_1 + b_2\,190$. If we don't know the actual weight y_i this information allows us to improve the estimate, but it is still only an estimate, we still have to expect an error $y_i - \widehat{y}_i = e_i$.

This consideration allows us to split the error we would make without knowing x_i, i.e. $y_i - \bar{y}$, into two parts, one part we can 'explain' by knowing x $\widehat{y}_i - \bar{y}$, and the rest $y_i - \widehat{y}_i = e_i$.

Figure 2.11 summarises these considerations. We have picked out a single observation (x_i, y_i) and start by decomposing for this observation the total deviation of y_i from the mean $\bar{y}$, i.e. the distance $y_i - \bar{y}$, into a 'explained by regression' distance $\widehat{y}_i - \bar{y}$ and an 'unexplained' distance $e_i = y_i - \widehat{y}_i$.

For a single observation as in Figure 2.11

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

By dispersion here we mean the sum of squared deviations. Therefore we square the above expression and sum over all observations

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

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

We will now show that, due to the properties of the OLS method, the third term on the right-hand side is always zero when the regression contains an intercept. This property follows from the first order conditions $\sum_i e_i = 0$ and $\sum_i x_i e_i = 0$ (equations (2.2) and (2.3), page 11).

This can be easily shown, the third term of equation (2.10) is

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

Since $\sum_i e_i = 0$ always holds for regressions with intercept (equation (2.2), page 11), it only remains to show that $\sum_i \widehat{y}_i e_i = 0$.

To do this, we set $\widehat{y}_i = b_1 + b_2 x_i$.

$$\begin{aligned}\sum_i\widehat{y}_i e_i &= \sum_i(b_1+b_2x_i)e_i \\ &= \sum_i(b_1e_i+b_2x_ie_i) \\ &= b_1\sum_i e_i+b_2\sum_i x_ie_i=0\end{aligned}$$

This expression is also zero because the first order conditions for the OLS residuals guarantee that $\sum_i e_i = 0$ and $\sum_i x_i e_i = 0$. This showed that for regressions with intercept, the cross term of equation (2.10) is always zero (i.e. $\sum_i(\widehat{y}_i-\bar{y})(y_i-\widehat{y}_i)=0$).

Therefore, the total dispersion of y around the mean breaks down into just two terms, the dispersion 'explained' by x and the 'unexplained' dispersion

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

resp.

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

where TSS stands for 'Total Sum Squared', i.e. the total dispersion of y_i around the mean $\bar{y}$. ESS is the 'Explained Sum Squared', the dispersion of the fitted values $\widehat{y}_i$ around the mean $\bar{y}$, and SSR stands for 'Sum of Squared Residuals', the dispersion of the y_i around the regression line, which is the sum of squared residuals.

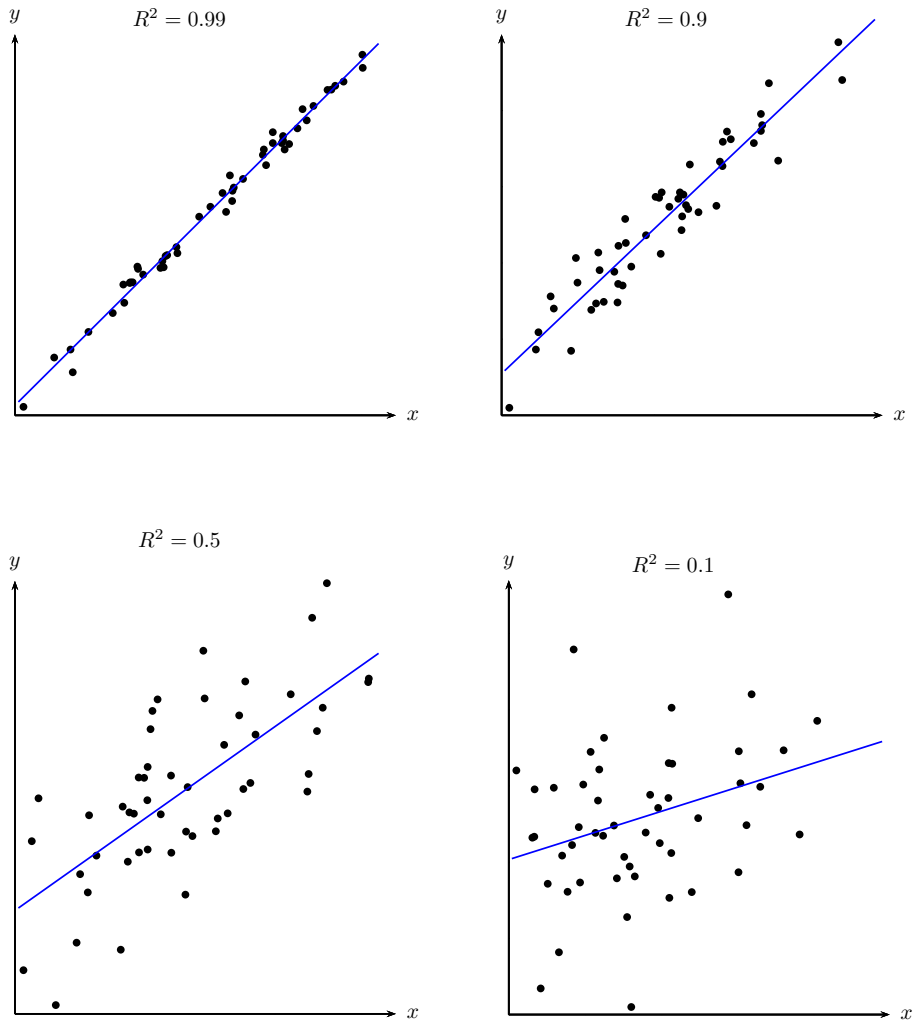


Figure 2.12: The coefficient of determination R^2 is an indicator for the dispersion around the regression line.

Finally, the coefficient of determination is defined as the proportion of the dispersion explained by the regression line ESS in the total dispersion of y, i.e. 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 other words, the coefficient of determination R^2 indicates what proportion of the total dispersion of y is explained by the regression line (or more precisely, by the explanatory variable x).

Since it is a proportion, the coefficient of determination for regression equations with intercept is always between zero and one (this must be true for regression equations without intercept *not*! Why?).

To give an impression of the fit with different R^2, Figure 2.12 shows some regression lines with different R^2.

On the other hand, completely different data can lead to the same R^2, the classic example being the Anscombe data, see Figure 2.13.

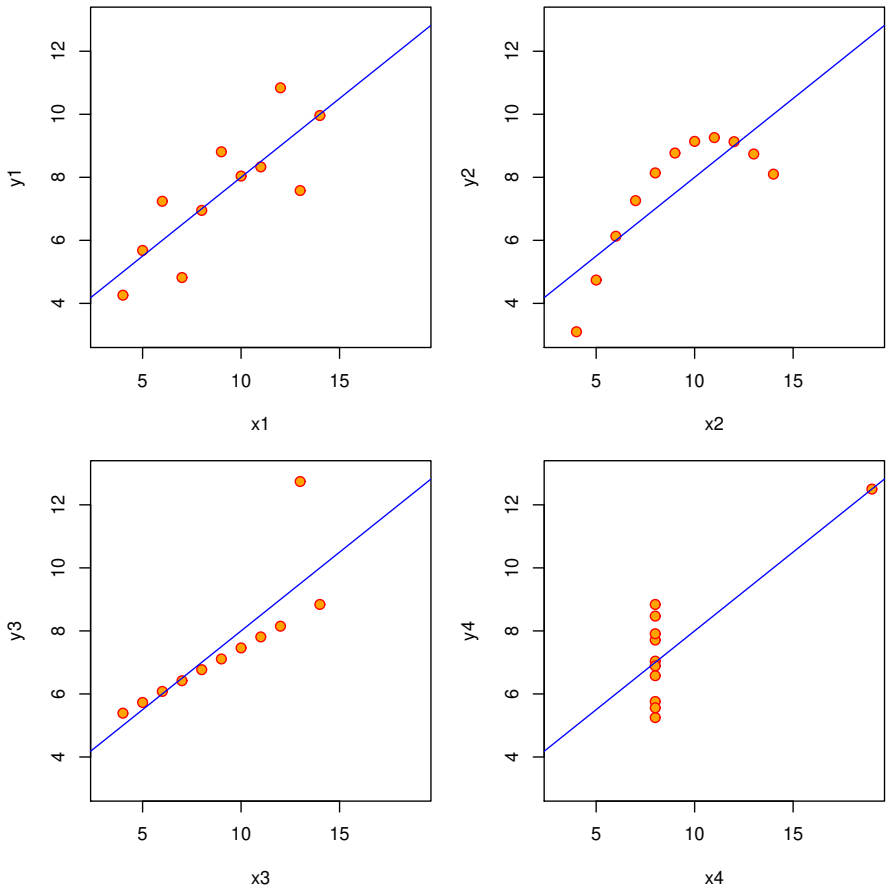


Figure 2.13: Anscombe data set, all four regressions have almost the same $R^2 = 0.666$! Source: Anscombe, Francis J. (1973) Graphs in statistical analysis. The American Statistician, 27, 17-21; taken from R package `datasets`, anscombe, Examples.

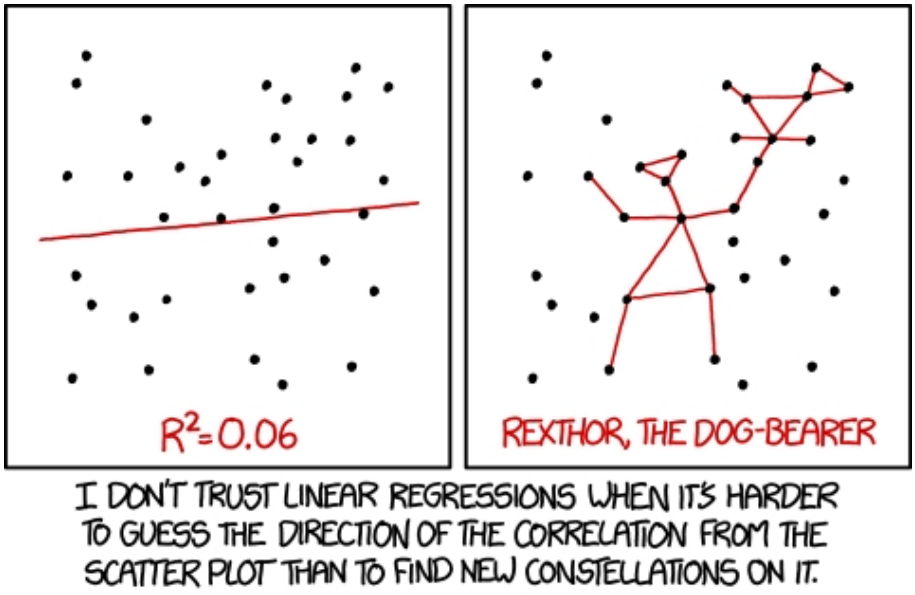


Figure 2.14: Quelle xkcd, `http://xkcd.com/1725/`

Since the R^2 is almost always given with the regression output and is easy to understand, beginners often tend to attach too much importance to the R^2. In particular, there is a widespread misconception that a high R^2 is associated with a more accurate measurement of the regression coefficients, and therefore a high R^2 is 'good' for the interpretation of the results. This is false, for example, if a regression equation is misspecified, it may have a very high R^2 even though the regression equation is more or less useless. On the other hand, a regression equation with a low R^2 may allow a rather accurate measurement of the regression coefficients (in which we are finally interested) if enough observations are available.

Exercises

1. Show that the coefficient of determination R^2 is the square of the (Pearson's) correlation coefficient between the observed values y and the fitted values $\widehat{y}$, i.e. $R^2 = [\text{corr}(y, \widehat{y})]^2 := r^2_{y,\widehat{y}}$.

   Notes. Pearson's correlation coefficient is defined as.

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

   Consider that $y = \widehat{y} + e$ and and the variance calculus rules $\text{cov}[x, (y + z)] = \text{cov}(x, y) + \text{cov}(x, z)$. Furthermore, we recall that

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

   and that in regressions with intercept $\text{cov}(\widehat{y}, e) = 0$ (why actually?).

2. Show that in a bivariate regression the coefficient of determination is also equal to the square of a correlation coefficient between y and x (this is only true for bivariate regressions).

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

Solution: First show that

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

Insertion gives

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

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)

So far we have only looked at measuring the relationship between two variables x and y. Most correlations in the real world are of course much more complex, almost always several explanatory variables act on a dependent y variable. For example, the price of used cars is not exclusively explained by age, but also by mileage, equipment, previous accidents, colour and much more.

Fortunately, the OLS method can be generalised very easily for the case with several explanatory variables.

The case with two explanatory variables can still be represented graphically in a 3-dimensional space; Figure 2.15 shows such a 3-dimensional mapping with the dependent y variable on the vertical axis and two explanatory variables x_2 and x_3 on the horizontal axes. While in the bivariate model we were looking for a regression *straight line* that would represent the data as well as possible, in the case with two explanatory variables we are looking for a regression *plane* that minimises the sum of squared residuals. The left panel in Figure 2.15 shows the observation points in space, the right panel shows the corresponding regression plane with the fitted values $\widehat{y}_i$ lying on this plane. Higher-dimensional cases, i.e. cases with more than two explanatory variables, can no longer be represented graphically, but the mathematical calculation is just as simple.

For two explanatory variables the regression function can be written as

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

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

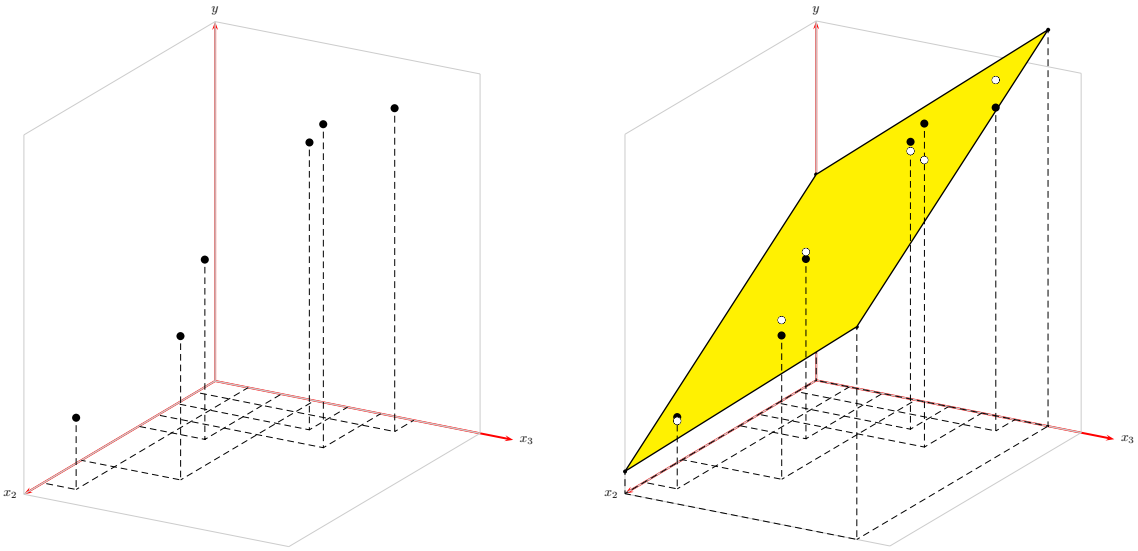


Figure 2.15: 3-dimensional mapping of the data and the regression plane $\widehat{y}_i = 5.73 - 0.51x_{i2} + 0.76x_{i3}$ (fitted values on the regression plane are shown as hollow circles).

where n again denotes the number of observations. Note that we now need two sub-indices for the explanatory x, the first sub-index $i = 1, \ldots, n$ still denotes the observation (i.e. the row of the data matrix), the second sub-index denotes the explanatory variable (i.e. the column of the data matrix).

We can calculate the three unknown coefficients b_1, b_2 and b_3 the same as before by minimising the sum of squared residuals:

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

We are looking for the values b_1, b_2 and b_3 that satisfy the following 1st order conditions:

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

Note that these equations again imply $\sum e_i = 0$, $\sum e_i x_{i2} = 0$ and $\sum e_i x_{i3} = 0$, since $(y_i - b_1 - b_2 x_{i2} - b_3 x_{i3}) = e_i$.

As solutions of these three first-order conditions one obtains after some arithmetic

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

where we introduce a new notation here for simplicity, two dots above a variable mean that from each observation i of a variable the mean of that variable was subtracted, i.e. $\ddot{y}_i := (y_i - \bar{y})$, $\ddot{x}_{i2} := (x_{i2} - \bar{x}_2)$ and $\ddot{x}_{i3} := (x_{i3} - \bar{x}_3)$ (see also section 2.12.1 Mean transformations). The running index $i = 1, \ldots, n$ of course again identifies the individual observation.

It should be noted that the OLS method also works with more than two explanatory variables, but the expressions in sum notation become rather confusing. We will show later that the multiple regression model can be written much more clearly with the help of matrices and can also be solved more easily.

Fortunately, these formulas for the OLS estimators are implemented in almost all statistical program packages (even in Excel), here it is only a matter of recognising that the calculation of the OLS estimators in the multivariate case follows the same basic principle as in the bivariate case.

With more than two explanatory variables, the multiple regression model is often written as

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

where h is the running index and k the total number of regressors including the regression constant, and the intercept b_1 is the coefficient of the regression constant $x_{i1} = 1$ as usual. For this model we need two running indices, i as a running index over the individual observations (i.e. rows) with $i = 1, \ldots, n$, and a running index h over the explanatory variables (i.e. columns) with $h = 1, \ldots, k$.

For a solution to exist, the number of explanatory variables k must be less than (or equal to) the number of observations n, i.e. $k \leq n$, and the explanatory variables must be linearly independent of each other (i.e. no perfect multicollinearity).

For clarification we write this once again in vector notation

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

An essential part of the charm of linear regression models lies in the simple interpretation of the coefficients as **marginal effects**, because due to the linear functional form, the regression coefficients are simply the partial derivatives and can be interpreted as such. The only new thing that is added is the *ceteris paribus* interpretation.

For the regression model

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

the regression coefficient b_2 indicates by how many units $\widehat{y}$ changes when x_2 increases by one unit *and* x_3 *remains constant*, i.e. *ceteris paribus*! Analogous is true for b_3

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

This ceteris-paribus interpretation is expressed by using the *partial derivative* ∂.

Attn.: This ceteris-paribus interpretation of the coefficients applies only with respect to the variables included in the regression, not to variables outside the regression model!

If in the car example km and age are regressed on price, the ceteris-paribus interpretation only applies to km and age, how does the fitted price change with an increase in km at constant age, and vice versa, but not with respect to other variables such as equipment features.

Example In an earlier section we examined the relationship between the price of used cars and their age. Of course, the price will depend not only on age, but also on many other factors, such as mileage.[10]

A regression of the sales price on age *and* mileage gives

[10] This is a very simple example of a hedonic pricing model (*'hedonic pricing model'*). Essentially, the price of a good is explained by its characteristics. Such pricing models are widely used, for example, in real estate markets.

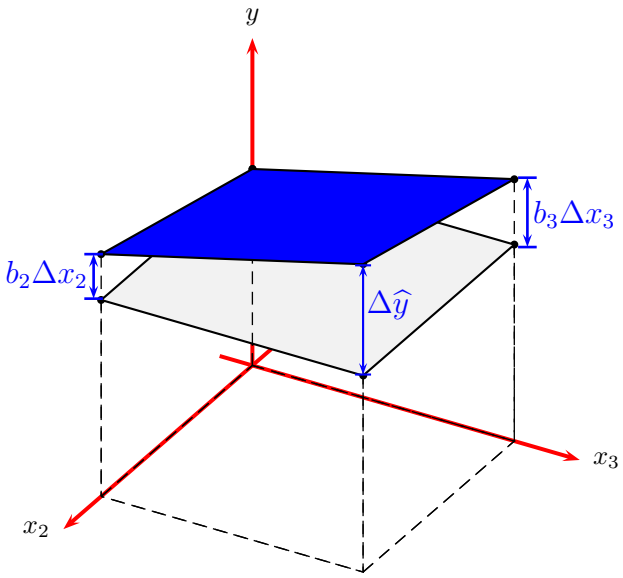


Figure 2.16: Ceteris-paribus interpretation of the coefficients; For the regression plane $\widehat{y}_i = b_1 + b_2 x_{i2} + b_3 x_{i3}$ it follows by forming first differences $\Delta\widehat{y} = b_2\Delta x_2 + b_3\Delta x_3$, from which follows for the coefficients

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

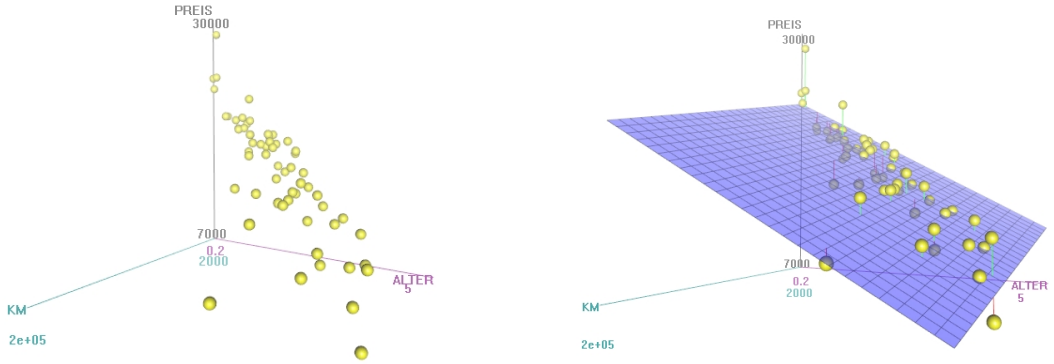


Figure 2.17: 3-dimensional illustration of the car example with the help of the R package `Rcmdr` (Fox, 2005).

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

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

This regression describes the relationship between price and age and mileage for 40 observations.

As before, we can interpret the fitted price for a car with given age and mileage as a linear approximation to the mean of this subcategory, e.g. the linear approximation for an average price of cars with an age of four years and a mileage of 100 000 km is equal to

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

where $\widehat{y}$ denotes the fitted price, x_2 the age and x_3 the mileage.

Mostly, however, we are interested in the individual coefficients. The intercept has a simple interpretation in this case, it gives the average price of a 'used new car', i.e. a used car with age = 0 and km = 0, however, the intercept is rarely of interest, so it is rarely mentioned when interpreting the results.

More interesting are usually the slope coefficients. Based on this regression, we would expect the price of a used car of this brand to fall by 1896 euros on average if the age increases by one year *and the mileage remains constant* (i.e. ceteris paribus)

$$\frac{\partial\, \widehat{\text{price}}}{\partial\, \text{age}} = 1896.264$$

Similarly, we must expect the price to fall by approximately 0.031 euros for each kilometre driven (i.e. by approximately 3 cents/km or by 31 euros per thousand kilometres), *if age remains unchanged* (ceteris paribus).

$$\frac{\partial\, \widehat{\text{price}}}{\partial\, \text{km}} = 0.031$$

Because of the linear functional form, this interpretation applies not only infinitesimally, but also to discrete changes in the explanatory variables. If, for example, an 'average' car is driven 30000 kilometres over a period of two years, an average loss in value of $1896.264 \times 2 + 0.031 \times 30000 = 4722.838$ must be expected on the basis of this regression. But of course this does not refer to the actual average prices, but to the fitted prices $\widehat{y}$ lying on the regression plane (i.e. to the linear approximation).

To emphasise the ceteris paribus interpretation, it is sometimes said that in the multiple regression model we *control* for the influence of the other explanatory variables, i.e. the coefficient of age measures the average depreciation per year when controlling for mileage. This language usage goes back to the experimental origins of regression analysis.

In this ceteris paribus interpretation of the coefficients as marginal effects lies a major advantage of the multiple regression model; it allows for the control of multiple influencing factors acting simultaneously on the dependent variable y. This ceteris

paribus interpretation of the coefficients is of course valid even if the data were not collected in a ceteris paribus way. For example, to determine the isolated influences of age on price *at constant mileage* we do not need data from cars with different ages and *same mileage*, because of the linear functional form the marginal ceteris paribus effects can be calculated even if each age – mileage combination is observed only once.

Linear regression therefore allows (or more correctly, enforces) a ceteris paribus interpretation of the coefficients even for non-experimental data. However,this interpretation is made possible only by the assumption of a linear functional form.

There are two important points to note here:

1. This ceteris paribus interpretation refers exclusively to the x *variables explicitly considered in the regression model*! For example, if you estimate a wage equation and take into account education, work experience and gender, the ceteris paribus interpretation *only* refers to these considered variables education, experience and gender, but not to *not considered* (and often unobservable) variables such as quality of education, social skills, assertiveness, intelligence, etc.

   This is a key difference to randomised controlled experiments (RCTs), which at least in principle allow indirect 'control' of unobserved factors as well. More on this in the chapter on *endogeneity*.

2. The marginal effect refers to the systematic part of the regression $\widehat{y}$, i.e. to the linear approximation to the conditional mean, not to the y_i of an individual. It is tempting to think of the marginal effect as affecting y, but this is wrong. $\widehat{y}_i$ is a conditional mean and tells us no more about an individual observation than the average income of a country tells us about the income of an individual of that country!

   As we will see in a moment, misspecification (e.g. not taking a relevant variable into account) will usually result in a coefficient not correctly measuring the impact on y. This will always be the case if a change in x also affects the error process e. More on this later in the section on 'nonconsideration of relevant variables' and especially in the later chapter on endogeneity.

But of course the ceteris paribus interpretation is also permissible if the explanatory variables are correlated with each other, as is to be expected in our example with the mileage and age of the cars.

However, as mentioned above, this ceteris paribus interpretation is only made possible by the assumption of the linear functional form.

In fact, by choosing the linear functional form, we have in a sense fitted the data to the Procrustean bed[11] of our specification; we will have more to say about this later.

Note also that so far we have only described the 'average' relationship for the given 40 observations, so this has been a purely descriptive analysis so far.

[11] Procrustes – a figure from Greek mythology – was known to offer travellers a bed, and then 'fit' the hapless wanderers to the size of the bed by brute force. If the wayfarer was tall he chopped off his feet, if the wayfarer was short he stretched him.

A stochastic interpretation: In the vast majority of cases, regressions are used to make inferences about an unobserved population, i.e. in the sense of *inductive statistics*. We will explain the details of this in detail in the following three chapters, here only a brief preview.
The numbers in brackets under the coefficients are the standard errors of the coefficients (if nothing else is mentioned), i.e. a measure of the accuracy of the coefficients. As a *rough rule of thumb* you can remember that the absolute value of the coefficients should be at least twice as large as the standard error (for approx. $n > 30$). This is often indicated by two or three asterisks next to the coefficients or standard errors (usually one asterisk indicates that the coefficient is different from zero at a significance level of 10%, two asterisks 5% and three asterisks indicate a significance level of 1%).

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

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

For the interpretation one can proceed in the following steps:

1. What relationship is shown and what do we expect to see? Do the signs agree with our expectations?
2. Are the coefficients statistically significantly different from zero? If no, do not interpret these coefficients further.
3. Interpret the quantitative significance of the statistically significant coefficients using the ceteris paribus assumption. The intercept is interpreted only if there are special reasons for doing so.
4. Briefly mention the goodness of fit (R^2) and to what extent we can assume that the assumptions underlying the estimation are satisfied (the details of which follow in the next chapters).

The regression equation above shows the dependence of the price of a given type of used cars on its age and mileage. We would a priori expect the average price to fall with increasing age and mileage. The equation shows that this is indeed the case, both signs are negative.
If we want to generalise the conclusions from this *sample* (with $n = 40$), we have to take into account possible sampling errors. Since the ratio of coefficient to standard error is so small, we can almost rule out the possibility that this is just a random result in this case (details follow later). The probability that there is no relationship between price and age or mileage in the population is less than one percent in this case.
From this we conclude that the average price falls by approx. 1900 euros with each year of age at constant mileage (i.e. ceteris paribus), and that the price decreases by approx. 3 cents with each additional kilometre at constant age. These two variables explain approx. 90% of the variance of the prices, so the fit of the regression line is pretty good.

The adjusted coefficient of determination $\overline{R}^2$ (*adjusted* R^2)

Everything said earlier about the coefficient of determination R^2 also applies to the multiple regression model, if the regression contains an intercept the R^2 is the proportion of the variance explained jointly by all x variables to the total variance of y (i.e. ESS/TSS).

However, there is a small problem in the multiple regression model: because the dispersion (variance) can never become negative, by including another regressor the R^2 cannot decrease and will usually increase. This is obvious, by including an additional regressor the fit can never get worse. Therefore, the usual coefficient of determination is not suitable for comparing regressions with a different number of explanatory x variables.

The adjusted coefficient of determination $\overline{R}^2$ attempts to at least mitigate this problem by introducing a correction factor.

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

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

With an increasing number of explanatory variables k, the factor $(n-1)/(n-k)$ becomes larger, compensating for the fact that $\sum_i e_i^2$ becomes smaller with increasing k. Therefore, the corrected coefficient of determination $\bar{R}^2$ is more suitable for comparing two regressions with a different number of explanatory variables.

In the context of stochastic regression analysis, we will see later that the sum of squared residuals $(n-k)$ has *degrees of freedom*, while the total dispersion of y in the denominator $(n-1)$ has *degrees of freedom*, so as a mnemonic, one can imagine that for the corrected coefficient of determination $\bar{R}^2$, the corresponding dispersions are simply adjusted for the degrees of freedom.

Alternative – and theoretically better founded – measures for model selection are the *Akaike Information Criterion* (AIC) and the *Bayes Information Criterion* (BIC), which are frequently used in time series econometrics.

2.6.1 Omitted variables

Let us return to our example with the used cars. The multiple regression explaining the price is $\text{price} = b_1 + b_2\text{age} + b_3\text{km} + e$; column (1) of Table 2.5 shows the result of this estimation again for comparison purposes. Column (2) shows the result of a regression *only* on age, and column (3) the result of a regression *only* on mileage. Since these two regressions have fewer explanatory variables we will call these 'short' models.

In the two 'short' models (2) and (3) we get much larger slope coefficients in absolute terms than those in the multiple ('long') regression model (1). What happened?

If we regress *only* on age, the slope coefficient measures not only the influence of age, but also indirectly the influence of mileage not taken into account. Since age and

Table 2.5: Price of used cars.

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

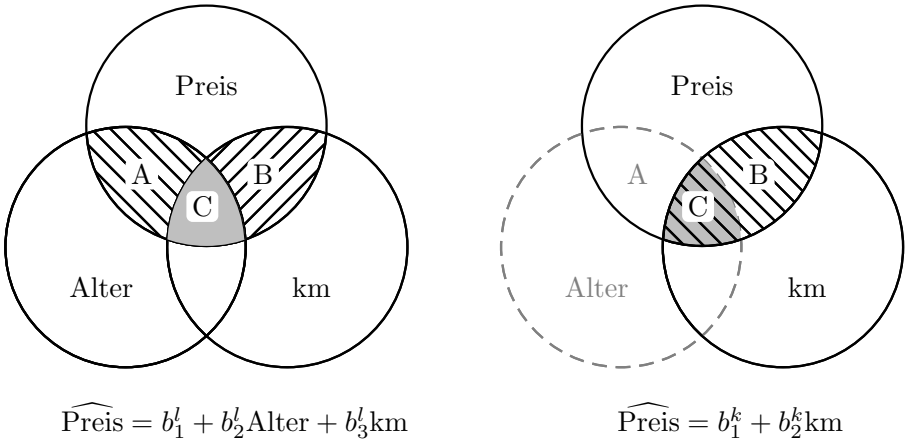


Figure 2.18: 'Long' and 'short' model; In the 'long' model (left panel), the overlap area C is not included in the estimation of the slope coefficients. If age is incorrectly omitted, the area C enters the estimation of the coefficient for mileage (*'Omitted Variables Bias'*, right panel).

mileage of used cars are usually positively correlated, we overestimate the influence of age, part of the price loss is due to the higher average mileage of older cars.

An intuitive insight is provided by the Venn diagram in Figure 2.18. The dispersion of the variables price, age and mileage is symbolised by circles, and the correlation between the variables by the intersections of the circles.

In the correctly specified model (left panel), area A enters into the estimate of the coefficient for age and area B enters into the estimate of the coefficient for mileage. The overlap area C, resulting from the correlation between age and mileage, cannot be clearly assigned to one of the variables and therefore does not enter into the estimate of the slope coefficients (but very much into the R^2).

This is different in the case of the misspecified model in the right panel. When age is not considered as an explanatory variable, the areas B *and* C enter into the estimation of the coefficient on mileage, the area C at least partially wrongly, as it is also attributable to the age not taken into account.

This erroneously gives the mileage a greater significance than it actually deserves, since it also partly captures the effect of the age not taken into account! The consequences are severe, the coefficient of mileage no longer measures the correct marginal effect, but is in a sense 'polluted' by the erroneously *not* considered variable age (note that the ceteris paribus interpretation only applies to the variables

considered in the model). Therefore, we get a far inflated price loss of 8 cents per kilometre instead of the 3 cents of the 'long' model that results when mileage *and* age are taken into account.

The same applies if we regress only on age and do not take mileage into account. In this case, we would wrongly attribute part of the price loss, which is actually attributable to mileage, to age.

This problem is known in the literature as *omitted variables bias* and will concern us in detail later in the context of stochastic regression analysis. It should be noted that an *omitted variables bias* can only occur when the omitted variable is correlated with both the dependent variable y and the considered regressor x.

The left panel of the Venn diagram in Figure 2.18 can give us another insight. If the regressors age and mileage are very highly correlated this leads to the overlap area C becoming very large, and the areas A and B becoming correspondingly small. However, since only the areas A and B are included in the estimation of the coefficients, the estimation becomes correspondingly imprecise, which essentially leads to the same problem as a (too) small sample. This problem of a high correlation between the explanatory variables is called *multicollinearity* in econometrics.

In the extreme case, when the regressors age and mileage are perfectly correlated (i.e. are linearly dependent), the circles for age and mileage lie on top of each other and the coefficients can no longer be estimated individually, and are therefore no longer defined. This extreme case is called *perfect multicollinearity*. We will also discuss these cases of multicollinearity in detail in a later chapter.

First, however, let us take a closer look at the problem of missing relevant variables and show what happens when relevant variables are not taken into account.

The algebra of not taking relevant variables into account

We start with the simplest multiple regression model, where we mean-transform all variables, i.e. $\ddot{x}_i := x_i - \bar{x}$ (see section 2.12.1). The mean transformation eliminates the intercept, which simplifies the following presentation (to increase readability, we also omit the observation index i).

We now compare the coefficients of a *long* model (denoted by a superscript l)

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

with the slope coefficient of a 'short' model in which $\ddot{x}_3$ is not taken into account

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

We assume that the *long* model correctly represents the data generating process, and that the *short* model is *mis-specified*.

The OLS slope coefficient of the misspecified 'short' model (with superscript s) is

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

To recognise the consequences of this misspecification (i.e. the nonconsideration of $\ddot{x}_3$) we substitute into the above OLS formula for the slope coefficient of the short model b_2^s for $\ddot{y}$ the correctly specified 'long' model $\ddot{y} = b_2^l \ddot{x}_2 + b_3^l \ddot{x}_3 + e^l$ and simplify

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

Due to the first order conditions $\sum_i \ddot{x}_{i2} e_i^l = 0$, therefore holds

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

So there is a simple relationship between the slope coefficients of the 'short' and 'long' model.

Does the expression $\operatorname{cov}(x_2, x_3)/\operatorname{var}(x_2)$ sound familiar? Exactly, this is the OLS formula for the slope coefficient of a regression from x_3 to x_2

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

where e^* denotes the residuals of this regression, as usual.

Therefore, we can write the relationship between the slope coefficients of the 'short' and 'long' models more simply as

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

If – and only if – b_3^l and a_2 *equation* are different from zero, not taking x_3 into account will cause the coefficients of the 'short' and 'long' models to differ.

Figure 2.19 shows the problem once again: if x_3 is not taken into account, x_2 will be in addition to its direct effect b_2^l is also mistakenly a part of the effect of x_3, since x_2 is a proxy for x_3. The size of this 'proxy effect' depends on two factors: first, on the effect of x_3 on y, i.e. on b_3^l, and second, on the correlation between x_2 and x_3.

For the case with several variables not taken into account, the formulae are somewhat more complex, but the essence remains.

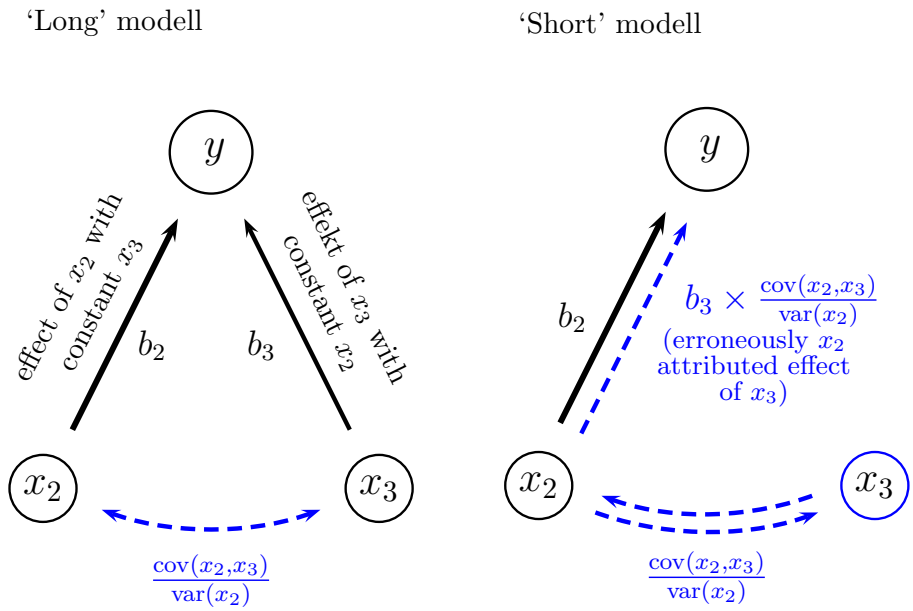


Figure 2.19: Failure to consider a relevant variable x_3 leads to part of the impact of x_3 being falsely attributed to x_2. If the 'true' model is $y = b_1 + b_2^l x_2 + b_3^l x_3 + e^l$ and a short model $y = b_1^s + b_2^s x_2 + e^s$ is erroneously estimated is $b_2^s = b_2^l + b_3^l \frac{\text{cov}(x_2,x_3)}{\text{var}(x_2)}$.

Example So what does this mean for our used car example? In Table(2.5) we have the estimate for one 'long' and for two 'short' models. To demonstrate the relationship, we restrict ourselves to the 'short' model with age.

As a reminder, the 'long' model from Table (2.5) was

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

and the auxiliary regression $\text{km} = a_1 + a_2\,\text{age} + \upsilon$ ist

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

We alternatively obtain the slope coefficient of the 'short' model from column (2) of Table (2.5) from $b_2^l + b_3^l \times a_2 = -1896.264 - 0.031 \times 23843.819 = -2635.669 = b_2^s$ (small deviations are due to rounding errors).

Now, what was all that for? The full implications of this result will only become clear later in the context of stochastic regression analysis; there we will see that not taking relevant variables into account leads to *endogenous regressors* and causes an *omitted variable bias*.

But already now this result allows us to estimate a possible 'error'. Indeed, whether the slope coefficient of the 'long' model is larger or smaller than the slope coefficient of the 'short' model depends only on the sign of the expression $b_3^l \times a_2$.

Suppose we had collected no data on the mileage of the cars and only prices and ages of the cars. We assume that price falls as mileage increases (i.e. $b_3^l < 0$), and

Table 2.6: Equation (2.12) allows an estimate of the direction of the error in estimating a 'short' model $y = b_1^s + b_2^s x_2 + e^s$ instead of a 'long' model $y = b_1 + b_2^l x_2 + b_3^l x_3 + e^l$.
Since $b_2^s = b_2^l + b_3^l \times \text{cov}(x_2, x_3)/\text{var}(x_2)$ holds:

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

that mileage and age are positively correlated (i.e. $a_2 > 0$, or $\text{cov}(\text{km}, \text{age}) > 0$). Since $b_2^s = b_2^l + b_3^l \times a_2$ and $b_3^l \times a_2 < 0$, it follows that $b_2^s < b_2^l$, so the effect of age on price is probably overestimated in the 'short' regression (note that prices are negative, so $b_2^s = -2635 < -1896 = b_2^l$).

2.6.2 The Frisch-Waugh-Lovell (FWL) theorem

Already in the very first issue of Econometrica (1933) Ragnar Frisch and Frederick V. Waugh pointed out an interesting property of the multiple regression model that can also give us a deeper understanding of how to interpret the regression coefficients.

This result was later confirmed by Michael C. Lovell (1963); he showed that this also holds for groups of variables. Since then, this result has been known as the *Frisch-Waugh-Lovell* (FWL) theorem.

Essentially, the FWL theorem shows that a coefficient of interest in a multiple regression can alternatively be calculated using multiple bivariate (short) regressions.

When Frisch and Waugh (1933) proved this result, computers were still hardly available, and because multiple regressions were far more time-consuming to calculate than bivariate regressions, this result did have practical significance at the time. Today, computing time is cheap, but this result is still important. It allows us deeper insights into the 'OLS mechanics', contributes to the understanding of regression coefficients in multiple regressions, and has numerous applications in advanced areas of econometrics, e.g. panel econometrics.

Specifically, the FWL theorem states the following: if, for example, we are interested in the coefficient b_2 of the multiple regression.

$$y = b_1 + b_2 x_2 + b_3 x_3 + e \tag{2.14}$$

we can alternatively calculate it using three bivariate regressions.

First we regress the two variables of interest y and x_2 on the variable x_3 to be eliminated

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

where e^y denotes the residuals of the first bivariate equation and e^{x_2} the residuals of the second bivariate equation.

Note that in each case the residuals include the effect 'adjusted' for the linear influence of x_3:[12] $e^y = y - (y|x_3)$ and $e^{x_2} = x_2 - (x_2|x_3)$

$$e^y = b_2\, e^{x_2} + e$$

The FWL theorem guarantees that a *bivariate* regression of these two residuals yields exactly the same coefficient b_2 and the same residuals e as the long regression 2.14!

We take advantage of the fact that OLS is a decomposition method; in the residuals of the two 'short' regressions on x_3, the (linear) influence of x_3 on y or x_2 was eliminated. This is often called *'partialling out'*. As already mentioned, this result was generalised by Lovell (1963) for more than 2 variables.

Proof: The proof of this theorem is usually done with the help of matrix algebra. Here we will sketch a much simpler proof that follows Lovell (2008).

Our starting point is again multiple regression

$$y_i = b_1 + b_2 x_{i2} + b_3 x_{i3} + e_i \tag{2.15}$$

The following is based on two properties of the OLS method:

1. The explanatory variables x_2 and x_3 are uncorrelated by construction with the residuals e. This follows directly from the first order conditions $\sum_{i=1}^{n} x_{ih} e_i = 0$ for all $h = 2, \ldots, k$.

2. If an explanatory x variable is not correlated with either the dependent variable y or the rest of the explanatory x variables, then the coefficient of that variable is zero. For example, in equation (2.15), if $\text{cov}(y, x_3) = 0$ and $\text{cov}(x_2, x_3) = 0$, then it follows that $b_3 = 0$.

We start by decomposing the dependent variable y and the explanatory variable x_2 into the systematic component explained by x_3 and the residuals using two OLS auxiliary regressions

$$y_i = c_1 + c_2 x_{i3} + e_i^y \tag{2.16}$$
$$x_{i2} = a_1 + a_2 x_{i3} + e_i^{x_2} \tag{2.17}$$

Note that due to the first order conditions $\text{cov}(x_3, e^y) = 0$ and $\text{cov}(x_3, e^{x_2}) = 0$.

We substitute these two equations into the long equation (2.15) and get.

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

[12] recall that the systematic part $\widehat{y}$ is the conditional means, i.e. $\widehat{y} = c_1 + c_2 x_3 = (y|x_3)$.

from this follows after rearrangement

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

If an explanatory variable *neither* is correlated with the dependent variable (y) *nor* with another explanatory variable (x_2), the coefficient of this variable must be zero.

However, from the first-order conditions of equations (2.16) and (2.17) we know that $\text{cov}(x_3, e^y) = 0$ (equation 2.16) and that $\text{cov}(x_3, e^{x_2}) = 0$ (equation 2.17), therefore the coefficient of x_3 must be zero, i.e. i.e. $b_2 a_2 - c_2 + b_3 = 0$. Therefore

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

Moreover, we already know that in a regression of mean-transformed variables, the intercept is zero. In our case, both the dependent variable e_i^y and the explanatory variable $e_i^{x_2}$ are residuals from regressions with an intercept, so their mean must be zero (first order condition!), so the residuals are already mean transformed. For this reason, the intercept is also zero ($b_1 - c_1 + b_2 a_1 = 0$) and we get as a result

$$e_i^y = b_2\, e_i^{x_2} + e_i$$

Note that b_2 from this equation is exactly equal to b_2 from 'long' regression (2.15), that is, we get b_2 from regressing the residuals of the two auxiliary regressions (2.16) and (2.17) exactly the same coefficient b_2 and also the same residuals e_i as from the 'long' regression (2.15). ■

We can therefore say that the coefficient b_2 of the 'long' regression (2.15) describes the impact of x_2 on y, after the linear influence of x_3 has been eliminated, or in other words, *after controlling for* x_3.

We have already mentioned that this theorem holds more generally, it is also possible to eliminate the linear influence of several variables by regressing on this group of variables in the auxiliary regressions.

Example: We can demonstrate this result again using the example of used cars. We use two auxiliary regressions to eliminate the linear influence of kilometres on price and age.

To do this we calculate the residuals of the two equations

$$\begin{aligned} \text{price} &= a_1 + a_2\,\text{km} + e^p \quad \rightarrow \quad e^p \\ \text{age} &= c_1 + c_2\,\text{km} + e^a \quad \rightarrow \quad e^p \end{aligned}$$

and then regress (without intercept!)

$$e^p = b_2\, e^a + e$$

In R, for example, this can be done with the code in script 2.1 (or in STATA see script 2.2).

Script 2.1: Example of Frisch-Waugh-Lovell Theorem, `R-Code`

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

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

eq_res
# Coefficients:
# res_Alter
#      -1896

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

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

Script 2.2: Example of Frisch-Waugh-Lovell Theorem, `STATA-Code`

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

We have used one consequence of the FWL theorem before without explicitly pointing it out, namely in the mean transformation $\ddot{x} := x_i - \bar{x}$. We have claimed that we can calculate the same coefficients from mean-transformed data as from the original data. Recall that a regression on the regression constant yields the mean $\bar{y}$; the residuals of this regression on the regression constant are therefore simply the mean-transformed data. The FWL theorem tells us that from a regression of these residuals on each other we get the same slope coefficient as from the original data.

Attn: the FWL theorem applies as well to the coefficients of stochastic regression model, but it does not apply to the *standard errors* of the coefficients! The reason is that in the residual regression it is not taken into account that degrees of freedom are lost through the two preceding auxiliary regressions.

Partial regression plots

Among other things, we can also use the FWL theorem to graphically represent the relationships between dependent and explanatory variables in *multiple* regression.

Remember, in a two-dimensional scatter plot we can only plot the result of a bivariate regression. If, however, other variables act on y and x, this leads to the fact that these variables not taken into account distort the relationship between y and x, i.e. cause a *omitted variables bias* (see section 2.6.1).

Therefore, graphical representations of bivariate correlations in scatter plots can be very misleading, an apparent correlation could also be due to variables not taken into account (spurious correlation).

The FWL theorem offers a simple way to correctly represent partial correlations by first eliminating the linear influence of all other (available) variables by means of auxiliary regressions, and then representing the residuals of these auxiliary regressions in a scatter plot.[13]. Such scatter plots are called *'partial regression plots'*, sometimes also *'added variable plots'*, *'adjusted variable plots'* or *'individual coefficient plots'*.

Example Let us return to our earlier example with used cars. Because the explanatory variables age and mileage are correlated and both influence price, a bivariate scatter plot of price vs. mileage draws an overly optimistic picture of the correlation; the omitted age influences the picture indirectly (see Figure 2.18).

Figure 2.20 shows the relationship between price and mileage of used cars on the left without taking age into account, and on the right after controlling for age. The corresponding R code is in script 2.3, and the Stata program code in script 2.4.

We see that the bivariate regression (left) overestimates the influence of kilometres, and that the partial regression (right) is strongly influenced by a single observation (the same scale was chosen for both graphs for better comparability).

[13]However, it should be noted that this changes the scaling. By adding the means to the residuals, it can be achieved that the partial regressions also run through the mean of the variable

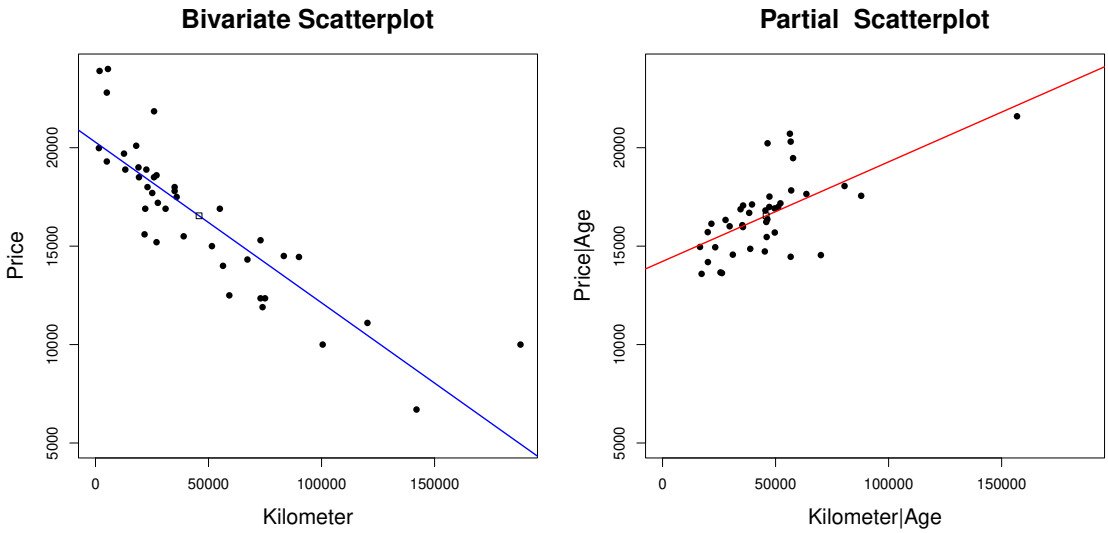


Figure 2.20: Bivariate and partial regression: in partial regression, the age of the cars is controlled for, i.e. residuals of regressions are plotted on age (so that the regression line still runs through the mean values, the respective mean values are added to the residuals).
For better comparability, the same axis scaling was chosen for both graphs.

Script 2.3: Bivariate scatter plot and partial regression plot, `R-Code`

```
rm(list=ls())
car <- read.csv2("http://www.hsto.info/econometrics/data/auto40.csv")
names(car) <- c("age", "price", "km", "liftback", "PS90")

car$res_price <- resid(lm(price ~ age, data = car)) + mean(car$price)
car$res_km    <- resid(lm(km ~ age, data = car))    + mean(car$km)

x11(width = 400, height = 300)  ## only for Windows
par(mfrow=c(1,2),cex.main=0.85)
plot(car$km, car$price, main = "Bivariate␣scatter␣plot",
     xlab = "km", ylab = "price",
     xlim = c(0, max(car$km)), ylim = c(5000, max(car$price)))
abline(lm(price ~ km, data = car), lwd = 1.6, col = "blue")
points(mean(car$km), mean(car$price), pch = 22)
#
plot(car$res_km, car$res_price, main = "Partial␣regression␣plot",
     xlab = "km␣|␣age", ylab = "price␣|␣age",
     xlim = c(0, max(car$km)), ylim = c(5000, max(car$price)))
abline(lm(res_price ~ res_km, data = car), lwd = 1.6, col = "red")
points(mean(car$km), mean(car$price), pch = 22)
```

Script 2.4: Bivariate scatter plot and partial regression plot, STATA-Code

```
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)
```

2.7 Dummy Variables

> *"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 variables are among the most useful things introductory econometrics has to offer. Indeed, very often we are interested in comparisons between groups, e.g. between countries, industries, or in the consequences of belonging to certain groups (e.g. gender). So far, we have studied only variables that could take any value within a range, i.e. *interval-* or *ratio-scaled*[14] variables. For example, to model the assignment of a person to a group, variables that can take only two values are sufficient, e.g., one (1) for *'true'* and zero (0) for *'false'*. This is why such variables are often called 0-1 variables, binary variables, or qualitative variables. In econometrics, the term *dummy variables* has come to be used for them.

Such dummy variables can be used in a regression model to examine the effects of qualitative differences, such as wage differences between men and women. Dummy variables are an extremely useful and flexible tool that can be used to examine a variety of questions, such as wage differences between men and women, whether countries in the tropics grow more slowly than countries in temperate climates, whether and how marginal propensity to consume changes after tax reform, or the extent to which the spending behavior of married versus single people differs.

Dummy variables can only take two values, zero and one, and are used for the coding of groups. If a (binary) characteristic is present, the number one is assigned to the dummy variable, and if this characteristic is *not* present, the number zero is assigned. For example, a dummy variable OECD is assigned the number one if a country is an OECD member, and zero if it is not an OECD member. Or, a dummy variable w (for female) is assigned the number one if the person is a woman, and zero otherwise. Of course, one could just as well create a dummy variable m for male

$$w_i = \begin{cases} 1 & \text{if person } i \text{ is a woman,} \\ 0 & \text{otherwise (i.e. man)} \end{cases} \qquad m_i = \begin{cases} 1 & \text{if man, and} \\ 0 & \text{other} \end{cases}$$

Tip: In logic, it is common to assign the number one to true statements and the number zero to false statements. When choosing the name of dummy variables, it is therefore recommended to choose the name in such a way that it can be concluded from the name to which characteristic the value '1' was assigned. For example, if

[14]For interval-scaled data, the order is fixed and the differences between two values can be interpreted in terms of content. For ratio-scaled variables, an absolute zero point also exists. In this section, we will look at cases where at least one explanatory variable is nominally or ordinally scaled. In the case of a *nominal scale*, the expressions cannot be placed in any *natural order*. Examples of nominal scaled characteristics are gender, religion, skin color, etc. With a *nominal scale*, there is a natural ranking, but the distances between the characteristic expressions cannot be quantified in a meaningful way. Examples include school grades, food quality grades, etc.

we were to give a dummy variable the name 'gender', we would not be able to infer from this variable name which gender was assigned the value one. If, on the other hand, we name the dummy variable 'female' it is clear that this variable has been assigned the value 1 for females and 0 for males. This can greatly facilitate the interpretation of dummy variables, as we will see in a moment.

We start with a simple example, Table2.7 shows hourly wages (wage) of 12 people ($n = 12$), as well as their gender, marital status and years of education.

Table 2.7: hourly wages (wage) of men (m) and women (w), marital status ($v_i = 1$ for married and zero otherwise; $u_i = 1$ for not married and zero otherwise), and education (in years).
Note that $w = 1 - m$ (resp. $m = 1 - w$ or $m + w = 1$) and $u = 1 - v$.
(`https://www.uibk.ac.at/econometrics/data/stdl_bsp1.csv`)

i	wage	m	w	v	u	educ
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

We can easily calculate the mean hourly wage $\overline{\text{wage}}$ and the *conditional* mean hourly wages for men, women, married, and unmarried from the data in Table2.7:

Mean wage:

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

Conditional mean values of wage:

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

In an earlier example, we showed that a regression *only* on the regression constant (i.e. a ones vector) yields the mean of the dependent variable (see page 16). We will now see in a moment that we can also compute the conditional means simply using OLS regression, namely by regressing on a dummy variable.

For the data from Table2.7, a regression on the dummy variable m yields

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

When $m_i = 0$, i.e., for women, we get $\widehat{\text{wage}}_i = b_1 + b_2 \times 0 = b_1$; therefore, we conjecture that the intercept b_1 provides the average hourly wage of women.

For men, $m_i = 1$, so $\widehat{\text{wage}}_i = b_1 + b_2 \times 1 = b_1 + b_2$, so we conjecture that $b_1 + b_2$ gives the average hourly wage of men, and the slope coefficient b_2 measures the difference between average hourly wages of men and women.

This is indeed correct, the conditional mean hourly wage for women is 12.5 euros, and men in this example earn on average 2.5 euros more than women, i.e. 15 euros.

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

Since the intercept in each case gives the mean of the 'zero category' (i.e. the mean of the category to which the value zero was assigned in the dummy variable), this 'zero category' is often called *reference category*.

The slope coefficient measures the difference between the mean of this reference category and the mean of the 'one category' (i.e. the category which was assigned the value one in the dummy variable), in this example by how many euros the average hourly wage of men (with $m_i = 1$) is higher than the average hourly wage of the reference category (i.e. women with $m_i = 0$).

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

Alternatively, we could have calculated a regression on the dummy variable w (for female); this yields

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

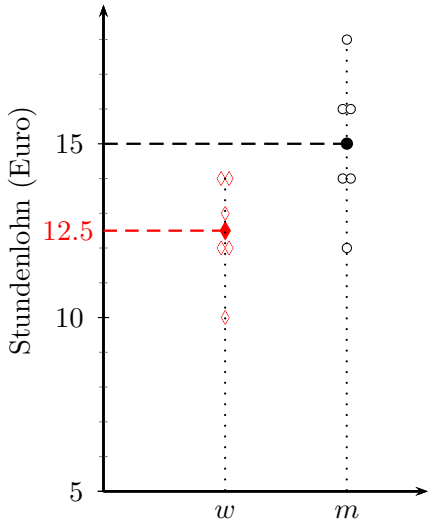
Since $w_i = 1$ for females and $w_i = 0$ for males, in this case males form the reference category whose mean hourly wage is measured in the intercept ($b_1 = 15$), and the mean hourly wage for females is 2.5 euros *lower* than that of males ($b_2 = -2.5$). Figure 2.21 shows this for the data from Table 2.7.

One might get the idea to compute a regression on a regression constant *and* the two dummy variables w and m, i.e. $y_i = b_1 + b_2 w_i + b_3 m_i + e_i$. However, this does not work because in this case there is a linear relationship between the regressors (the sum of the two dummies gives the regression constant, i.e. $w_i + m_i = 1$). Whenever there is an exact linear dependence between regressors the OLS function is not defined, infinite solutions exist.[15]

This is easy to see in the simplest case; if all manifestations of the regressor have the same manifestations (e.g. $x_i = 5$) x would be a multiple of the regression constant, and the variance of a constant is of course zero. Since $b_2 = \text{cov}(x, y)/\,\text{var}(x)$ and $\text{var}(x) = 0$, no solution to b_2 exists in this case. We will show later that this is true for all linear dependencies between regressors.

But we can compute a regression on both dummy variables m and w *without* regression constant. In this case, the estimated coefficients simply provide the means of both categories

[15] We will discuss this case in detail later under the name *perfect multicollinearity*.



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

or

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

Figure 2.21: hourly wages of men and women, see Table 2.7.

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

Exercise with solution hints: Let y_i be one of a total of n observations of an interval-scaled variable, and d be a dummy variable; n_1 is the number of elements of this dummy variable which equal one and n_0 is the number of elements with value zero ($n_1 + n_0 = n$).

The mean of y is $\bar{y} := \frac{1}{n}\sum_{i=1}^{n} y_i$, and the mean of d is $\bar{d} := \frac{1}{n}\sum_{i=1}^{n} d_i = \frac{n_1}{n}$ (why?).

We call the mean of all y_i for which $d_i = 0$ holds $\bar{y}_0$, and the average of all y_i with $d_i = 1$ we name $\bar{y}_1$.

We will now show in several steps that the following holds in general

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

i.e., the intercept b_1 of this regression is the mean of the group with $d_i = 0$ (reference group), and the slope coefficient b_1 is the difference of the group with $d_i = 1$ from the reference group.

1. Let $\bar{y}_1$ be the conditional mean of the y_i for which $d_i = 1$ holds, and $\bar{y}_0$ be the conditional mean of all y_i for which $d_i = 0$ (i.e. $\bar{y}_0 := \bar{y}|(d_i = 0)$ and $\bar{y}_1 := \bar{y}|(d_i = 1)$). Show in general that

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

   Solution hint: the data are sorted first, so that all observations with $d_i = 0$

come first, and then all observations with $d_i = 1$.

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

The sum of the *conditional* means weighted by the proportions is the overall mean!

2. Show that the empirical variance $\text{var}(y) := \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$ can also be written as.

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

Solution sketch:

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

since from $\bar{y} := \frac{1}{n}\sum_i y_i$ follows $\sum_i y_i = n\bar{y}$

3. Show that the empirical covariance $\text{cov}(y, x) := \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})$ can also be written as.

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

4. Show that for a dummy variable d it holds that

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

Note that $\sum_i d_i^2 = n_1$ (because $1^2 = 1$)

Solution sketch: note that for a dummy variable $\sum_i d_i = n_1$ and $\bar{d} = \frac{n_1}{n}$. Therefore

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

5. Show that for a dummy variable d holds.

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

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

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

6. Show that in a regression on a dummy variable $y_i = b_1 + b_2 d_i + e_i$ the slope coefficient

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

hint: $1-\frac{n_1}{n} = \frac{n_0}{n}$ (why?)

7. Show that the intercept b_1 can be calculated as

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

Solution sketch:

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

Therefore, the intercept b_1 of a regression on a dummy variable $y_i = b_1 + b_2 d_i + e_i$ provides the mean of the reference category $\bar{y}_0$, and the slope coefficient measures the difference between the means of the two categories ($b_2 = \bar{y}_1 - \bar{y}_0$).

□

Partial Effects: We have already discussed the simplest case in the previous example, a simple regression on a regression constant and a dummy variable d.

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

which in the intercept gives us the mean of the reference category (for which $d_i = 0$) as the slope coefficient the difference of the means of the two categories.

This difference corresponds to the *marginal effect* in metric scaled regressors, but since dummy variables by definition cannot change infinitesimally it is hardly appropriate to speak of an *marginal effect*; after all, we may be dealing with differences such as between men and women, a partial derivation makes little sense here.

As we have already seen, the coefficient of the dummy variable measures the difference from the 'reference category' $d_i = 0$

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

and the difference is the *"partial effect"*

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

This does not change significantly if further explanatory x variables are considered as regressors

2.7.1 Differences in intercept

We extend our dummy model by considering *additionally* a metrically scaled variable. To do this, we return to our example with hourly wages (see Table2.7) and additionally take into account the years of education ('educ').

A regression on the dummy variable m (for male) and 'educ' gives

$$\widehat{\text{wage}} = 5.78 + 2.95\,m + 0.45\,\text{educ}$$

(with $R^2 = 0.87$ and $n = 12$).

What happened? Suddenly the intercept is much smaller and the difference between men's and women's hourly wages is even larger (remember, a regression on the dummy variable only gave $\widehat{\text{wage}} = 12.5 + 2.5\,m$).

The now much smaller intercept is quickly explained, it gives the hypothetical mean hourly wage for women with zero years of education; no such person exists in this dataset.

But why do men now seem to earn on average 2.95 euros more than women? The answer follows from the ceteris paribus condition, *at equal education*!

Let us recall the chapter on not taking relevant variables into account. There we argued that the following relationship exists between the slope coefficient of a 'short' and 'long' model:

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

where b_2^k is the coefficient of the dummy variable of the short model $\widehat{\text{wage}} = 12.5 + 2.5\,m$, and $b_2 = 2.95$ $b_3 = 0.45$ are the above coefficients of the long model.

In the short model, the variable education is 'missing'. We regress this missing variable education in an auxiliary regression on the dummy variable m and obtain the slope coefficient a_2

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

This auxiliary regression tells us that women (the reference category) have an average of 15 years of education, and men have one year less (i.e. 14 years of education).

If we insert this into the formula for the non-accounted variable we get

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

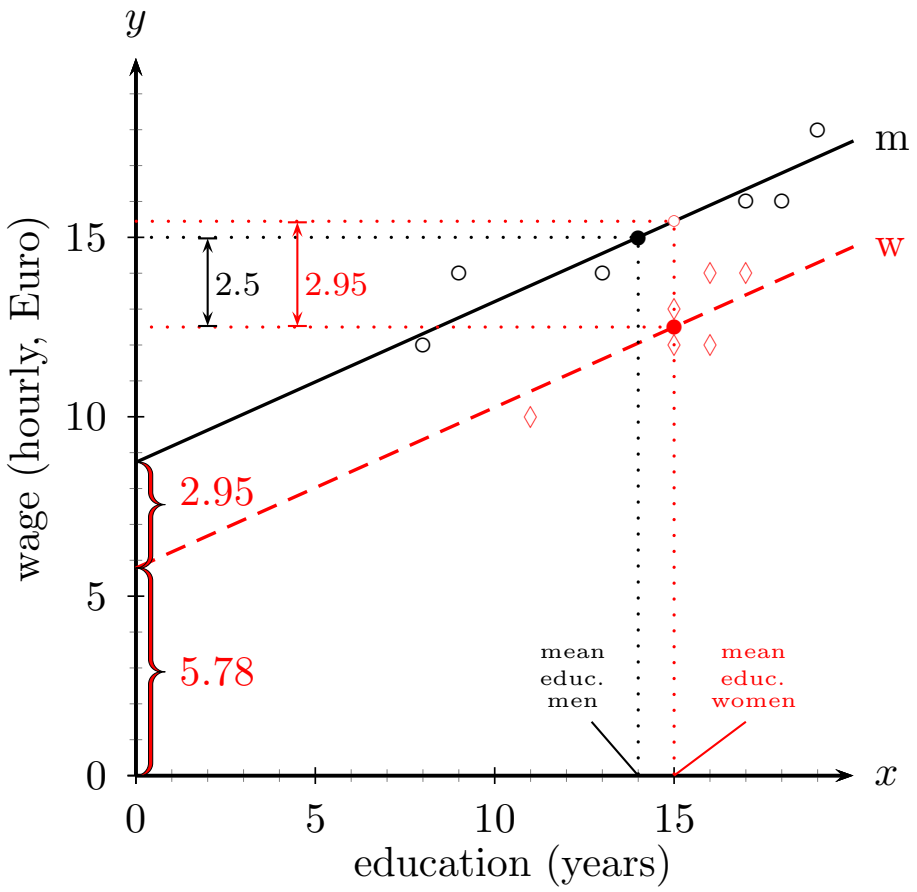


Figure 2.22: Differences in the intercept; $\widehat{\text{wage}} = 5.78 + 2.95\,m + 0.45\,\text{educ}$
The simple mean wage difference between men and women is 2.5 euros, but this does not take into account that here the average duration of women's education is 15 years, one year more than the average duration of men's education (14 years). The *ceteris paribus difference* (i.e. for the same duration of education) is 2.95 euros! Note also the significance of the assumed linear functional form

.

So from the 'long' regression with years of education we learn that a simple comparison of average numbers can be misleading.

In this example, women have a higher level of education on average, and the 'short' model does not take this into account. Since 'educ' is not a regressor in the 'short' model, it does not fall under the ceteris paribus assumption in the 'short' model, but it does in the 'long' model.

The level of education is positively correlated with the hourly wage ($b_3 > 0$), and men on average have a lower level of education ($a_2 < 0$, the 'short' model underestimates the ceteris paribus difference, see Table2.6 (page 46).

With the same education (ceteris paribus), the wage difference of 2.95 euros would be significantly larger than the difference in simple means of 2.5 euros! This is shown in Figure 2.22.

For Austria, the following estimate is obtained on the basis of EU-Silc (2019) data.

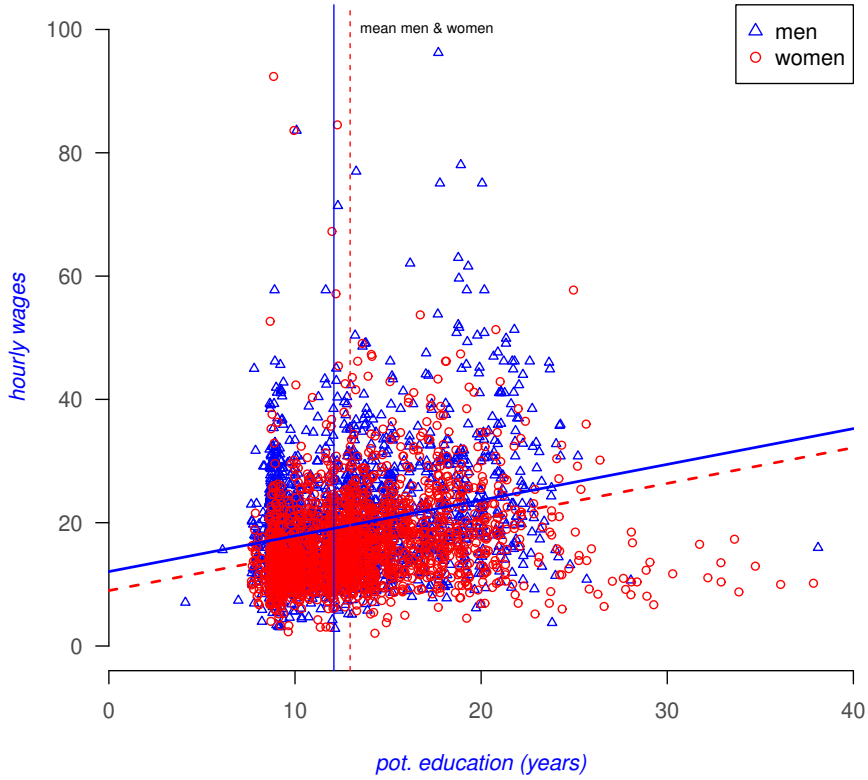


Figure 2.23: Hourly wages and (potential) years of education in Austria (hourly wages > 100 Euro truncated for clearer presentation, $n = 4151$). Source: EU-Silc 2019, Statistics Austria

$$\begin{array}{rcccccc} \text{wage} & = & 12.073 & + & 0.58\ \text{educ} & - & 3.085\ \text{female} \\ & & (0.506)^{***} & & (0.038)^{***} & & (0.305)^{***} \end{array}$$

$$R^2 = 0.061, \qquad n = 4720$$

which is shown in Figure 2.23. Note that the distribution of hourly wages is very right-skewed, and therefore conditional averages are not a very suitable measure; we will return to this in the section on logarithmic functions.

By specifying this with a simple dummy, we have allowed the intercept to differ between men and women, but we have assumed a priori that education has the same impact for men and women, with each additional year of education the mean hourly wage for men and women increases by 0.45 euros. This is of course a very restrictive assumption, but we can easily relax it.

2.7.2 Differences in slope

Introducing the product of a dummy variable with another metric-scaled variable as an additional regressor allows for different slopes of the regression lines for the two categories.

This is possible by introducing the *product* of dummy variables as an additional regressor. Such a product of two regressors is called *interaction effect* and will concern us in more detail later.

Suffice it to say here that the product of two dummy variables is always 1 if *both* dummy variables have the value 1, and 0 otherwise.

In the example with hourly wages

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

In this case, the *slopes* of the regression lines of the two categories may differ, for the category $m = 0$ the slope is b_2, and for the category $m = 1$ the slope is $b_2 + b_3$.

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

The slopes are

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

.

The coefficient of the interaction term b_3 measures the *difference in slopes* between the two categories, because

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

For our example with hourly wages we get

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

Accordingly, the mean hourly wage of women ($m = 0$) would increase by 0.29 euros with an additional year of education, and the mean hourly wage of men would increase by $0.29 + 0.19 = 0.48$ euros with each additional year of education. This specification thus allows the modelling of different effects of the metric-scaled variable on the two categories underlying the dummy variable.

However, this specification implies the same intercept for both categories (see Figure 2.24), which in most cases is a restriction that is difficult to justify theoretically. It is almost always wiser to allow different intercepts *and* different slopes.

2.7.3 Differences in intercept and slope

We can easily generalize the specification to allow for differences in intercept *and* slope. To do this, we just need to use both a dummy and an interaction variable between dummy variable and metric scaled x variable

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

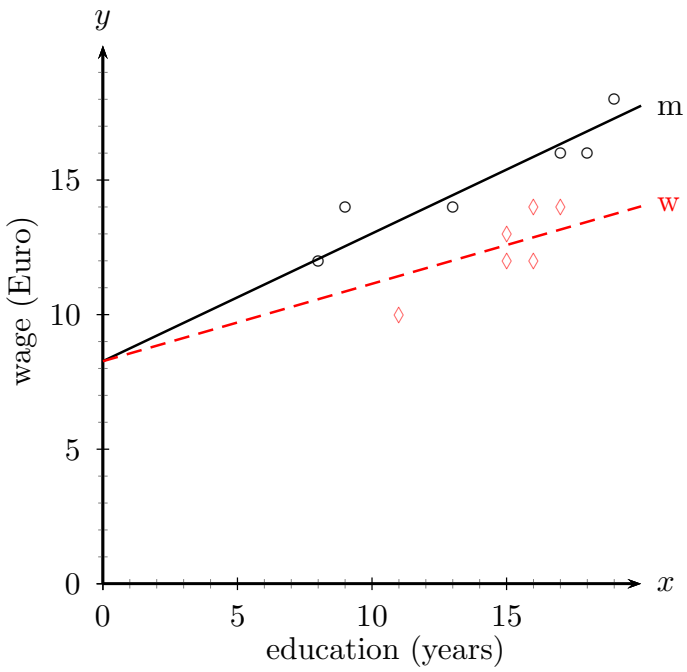


Figure 2.24: Differences in slope; $\widehat{\text{wage}} = 8.27 + 0.29\,\text{educ} + 0.19(m \times \text{educ})$

The difference between the two categories is again

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

Note that one obtains the same coefficients if one would calculate a separate regression for both groups

$$\begin{array}{lll} \text{for } m=0: & \widehat{y}^0 & = b_1 + b_2 x \\ \text{for } m=1: & \widehat{y}^1 & = c_1 + c_2 x \end{array}$$

where $c_1 = b_1 + b_3$ and $c2 = b_2 + b_4$.[16]

For our example with hourly wages we get

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

(with $R^2 = 0.89,\ n = 12$); see Figure2.25.

If we calculate separate regressions for men and women we get:

For women ($m = 0$):

$$\widehat{\text{wage}} = 2.95 + 0.64\,\text{educ}$$

For men ($m = 1$):

$$\widehat{\text{wage}} = 9.25 + 0.41\,\text{educ}$$

Please note that these are purely hypothetical numbers in this example.

[16]However, the standard errors will differ in these approaches because the dummy variable model implicitly assumes the same variance σ^2 (*homoscedasticity*) for both groups. Therefore, before applying the dummy variable model, it should be tested whether the variances are actually the same in all groups. You can find out how to do this in the chapter on heteroskedasticity.

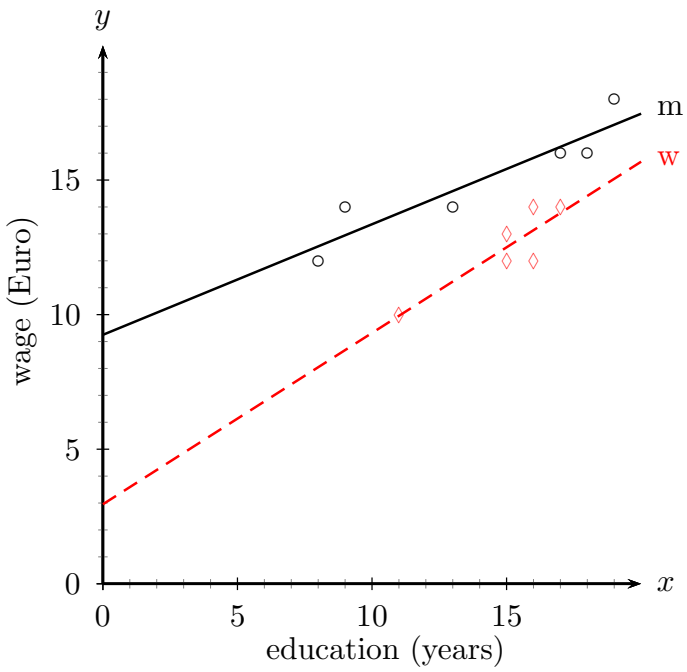


Figure 2.25: differences in intercept and slope; $\widehat{\text{wage}} = 2.95 + 6.30\, m + 0.64\, \text{educ} - 0.23(m \times \text{educ})$

2.7.4 Categorical variables with more than two values

Many categorical variables can have more than two possible values, e.g. nationality, highest level of education, school grades, etc. The different values are often coded in numbers, e.g. in the EU-SILC data of Statistics Austria the highest educational attainment are assigned numbers 1 to 6, see Table2.8, and unavailable values are coded with negative numbers (here -3 & -1).

Table 2.8: EU-SILC 2018 (Statistic Austria 2020)

```
P137000 Highest level of education
-3 Don't know
-1 Not specified
 1 Compulsory school
 2 Apprenticeship with vocational school
 3 Technical or commercial school
 4 Matura
 5 Graduation from a university, college
   or university of applied sciences
 6 Other qualification after Matura
```

Of course, one cannot calculate meaningfully with these numbers, they have no meaning in terms of content. The numbers were only assigned in order to be able to store the data in a space-saving way.

Nevertheless, it is very easy to work with data coded in this way, you only have

to choose one *reference category*, and for each further category a separate dummy variable.

In this concrete example with the educational qualifications, one would, for example, first identify the negative values with '`NA`' (for *'not available'*), choose as reference category e.g. '`compulsory school`', and for each further category create a dummy variable that takes the value 'one' if a person falls into this category, and 'zero' otherwise. So for this case we would need five dummy variables (most programs create this automatically, `R` uses `factor` objects for this).

The intercept then again measures the value of the reference category when all regressors (including dummies) are zero, and the coefficients of the dummies again measure the ceteris paribus difference from the reference category.

It is important that each observation falls exactly into one category, i.e. the categories form a partition.

Cases in which categories are not mutually exclusive (e.g., a variable 'nationality' if persons can have more than one citizenships) will be discussed in the next subsection (*Categories that are not mutually exclusive*).

For demonstration purposes, let us return to our example with the used cars, again using the age rounded to whole years '`age_rd`'; this variable has the expressions $\{0, 1, 2, 3, 4, 5\}$.

In column (1) of Table2.9, prices are regressed on this age with the 6 specifications. As shown earlier, the intercept (= 22 709.3) measures the average price of used cars with a rounded age of zero years, and the coefficient of age (= −2 517.27) measures the average decrease in price with each additional year. These are the same quantities we have already obtained in Table 2.3 (page 19) (except for different rounding).

Note that this specification only allows for a constant decrease in price, i.e. this specification forces an approximation according to which the decrease in price in year 1 must be exactly the same as in year 5. This may still be an acceptable approximation in this case, but at the latest in the case of nominal or ordinal scaled variables such an approximation no longer makes any sense at all.

In such cases, we (or the programme) create a separate dummy variable for each expression of the categorical variable with the exception of the reference category and use these as regressors.

In the example with the used cars, for example, we choose the age of zero years (i.e. age_rd = 0) as the reference category, and create a separate dummy variable for all other ages.

Column (2) of Table 2.9 shows the result of the regression. As expected, the intercept (= 23 566.67) gives the average price of cars with age_rd = 0, and the coefficients of the dummy variables measure the average difference in price to this reference category. Thus, cars with age four years are on average 11 163.81 cheaper than cars with age zero years. Note that this specification does not 'force' a constant decrease in price, but allows for different decreases in price from year to year.

Compare this again with Table2.3 (page 19); the coefficients of the dummy variables measure the difference in mean price to the reference category age_rd = 0.

Table 2.9: Three different specifications for the prices of used cars, age rounded to whole years. Compare results with Table 2.3 (page 19).

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

Finally, column (3) of Table2.9 shows the result of a regression in which no intercept is taken into account, but all six dummy variables for age are. As expected, the coefficients of the dummy variables in this case simply measure the average price of the cars with the age in question, as again shown by a comparison with Table2.3 (page 19). Note that the R^2 of the equation in column (3) must not be interpreted as usual because this regression does not contain an intercept!

2.7.5 Example: Heterogeneity and the Simpson paradox

Our objects of investigation usually differ in countless characteristics, and for practical reasons we are almost always forced to ignore a large part of this heterogeneity in our investigations.

Such *not considered heterogeneity* can have dramatic consequences, as the following hypothetical example demonstrates.

Suppose a university with two overcrowded degree programmes – psychology and mathematics – set up admissions tests.

After the test has been carried out, the following result will be announced

	women	men
Applicants	500	500
of which admitted	200	300
in per cent	40%	60%

Faced with a significantly higher rate of male admissions, the university management needed an explanation.

A brilliant statistician at the university comes up with the idea of looking at the results for the two courses separately, see table (2.10).

Psychology	women	men
Applicants	100	400
admitted	80	280
in per cent	80%	70%

Mathematics	women	men
Applicants	400	100
admitted	120	20
in per cent	30%	20%

Table 2.10: Admission quotas for women and men in two degree programmes. Although a total of 60% of men and only 40% of women were admitted, the admission rate for women is higher than that for men in both degree programmes.

The result is astonishing: women appear to have a significantly higher admission rate than men in both fields, although the aggregated figures would suggest the opposite. How can this be?

The result seems contradictory at first glance, but the figures speak for themselves.

This paradox can also be illustrated graphically using the example of a vector addition, see figure 2.26.

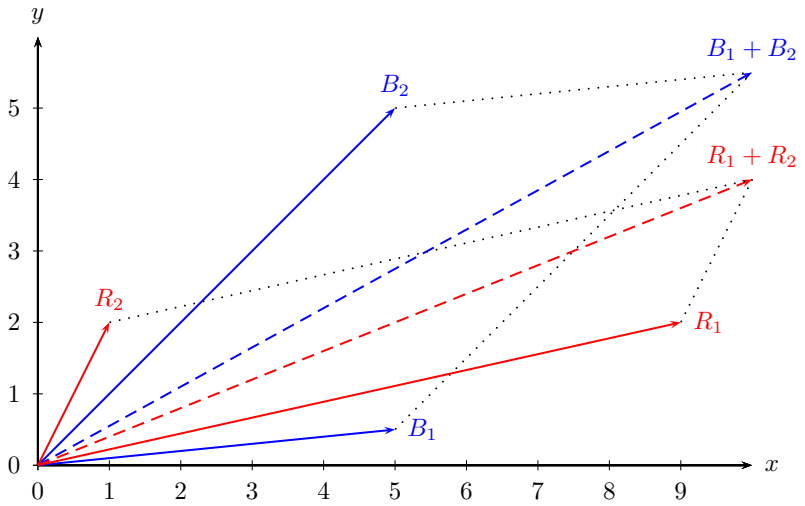


Figure 2.26: Simpson's paradox: although both red vectors (R) are steeper than the blue vectors (B), the vector sum of the blue vectors is steeper than the vector sum of the red vectors

.

This example is not far-fetched; it happened, for example, with admissions to graduate schools at the University of California, Berkeley, in the autumn of 1973. The

figures showed that more men than women were admitted overall, and the difference was so large that it could not be explained by chance. However, the breakdown by faculty showed that women were not discriminated against, but on the contrary were given a slight but statistically significant advantage (see `https://de.wikipedia.org/wiki/Simpson-Paradoxon`).

This paradox was named after Edward Simpson, who published the possibility of such a paradoxical result in 1951. However, the consequences of *'omitted variables'* were described much earlier by Karl Pearson (1899) and Udny Yule (1903).

We can easily reproduce this apparent paradox in a regression context using the techniques presented so far.

In essence, such a result can always occur if essential aspects - such as the field of study in this case - were not taken into account in the aggregated study, especially if the size and proportions of the subcategory groups differ greatly (e.g. in the above example, the proportion of women in psychology is only 20%, but 80% in mathematics).

So, as this example demonstrates, looking at the aggregated data can lead to completely misleading conclusions, and that taking group differences into account can sometimes be essential. The problem is that we can never really be sure that we have taken into account all the crucial group differences. It is therefore quite possible (although not very likely) that the results may be reversed if other characteristics are taken into account.

This hides a well-known phenomenon, namely the non-consideration of relevant variables (in this case categorical variables for the faculty assignment), which leads to a *'omitted variables bias'*.

Such problems are discussed in statistics under the keyword *'confounding'*, in econometrics these problems are known as *'unobserved heterogeneity'* or *'omitted variables bias'*. Since we can never know whether we have really taken all relevant influencing factors into account, we should always remain cautious when interpreting empirical results, especially when making causal statements!

Of course, such an extreme outcome will only in exceptional cases materialise, but the mere fact that it can happen should make us cautious.

To analyse this result in a regression context, we only need three dummy variables: admitted yes/no, female yes/no, psychology yes/no. The reference category is therefore men, or, for the second equation, men who study maths.

As each of the three dummies has two possible values, we need a total of 6 combinations. These are generated in the script 2.5 with the *replication* function of R (`rep()`, e.g. `rep(1, 80)` generates a vector of length 80 with all ones) and pack them into a data.frame (e.g. 80 women were admitted to the psychology programme). In the second data.frame, the 20 women who were *not* admitted are appended to the first data.frame using the rbind() function, etc.

The regression in table 2.11 shows the result.

Recall that a regression on dummy variables provides the proportions, and since this is a saturated dummy variable model, we get the exact proportions.

Table 2.11: Simpson's paradox with dummy variables

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

We also recall that the *omitted variables bias* in this simple model can be calculated as follows

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

Script 2.5: R Example of Simpson's paradox

```
## Simpson's Paradox, Example

# Psychology      Women Men
# Applicants        100 400
# of which admitted  80 280

# Mathematics     Women Men
# Applicants        400 100
# of which admitted 120  20

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

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

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

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


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

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

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

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

for example:

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

Table 2.12: Panel data in *'wide'* format.

2.7.6 Example: The LSDV and *'Fixed Effects'* Model

What we know so far already gives us insights into one of the most important models in applied econometrics, the *'Fixed Effects'* model for panel data.

We often observe several individuals (countries, firms, persons, ...) over several time periods, e.g. GDP of all OECD countries from 2005 – 2016, gross wages of all employees of a firm over the last four years, mean daily temperature at different measuring stations over the last 200 years.

Table 2.12 shows an example for 3 individuals, 4 time periods, and 2 regressors (x_2, x_3).

If the data have two dimensions (e.g. countries and time periods) we need two indices (*'identifier'*) to identify an observation; for example, y_{it} denotes the value of y for individual i in period t, where $i = 1, \ldots, n$ runs over individuals and $t = 1, \ldots, T$ runs over time.

This two-dimensional data structure (individuals and time) enables various evaluations. For example, one could calculate a regression over time for each individual, but in most cases this would be of little use, e.g. if we have data for several thousand individuals. In the same way, we could calculate a cross-sectional regression for each period, but this information is also rarely of interest.

A third possibility would be to calculate the time averages of all variables and to calculate a cross-sectional regression over these averages. If we note the average over time with $\bar{y}_{i\bullet} = 1/T \sum_{t=1}^{T} y_{it}$ (cf. Figure2.12) we get the so-called *'between'* model

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

in general: $(n = 3,\ T = 4)$

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

Example:

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

Table 2.13: Panel data in *'long'* format (*'stacked'*). This format can be generated in R e.g. with the `reshape2` package from the *wide* format (there is also a reshape command in Stata).

The name *'between'* model comes from the fact that only the heterogeneity *between* individuals is modelled, the dispersion over time 'within' individuals is not taken into account.

An alternative solution with maximum information compression would be to simply calculate a regression over all observations. To do this, the data must first be rearranged, i.e. they must first be transformed into the *'long'* format by arranging the data accordingly: one simply 'stacks' the observations for the individual individuals on top of each other (see Table2.13.[17]

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

This model implies that the coefficients b_1^p or b_2^p have the same value for all countries and time periods and is also called the *pool model* (hence the superscript p).

This 'pooled' model can be estimated normally with OLS.

For 3 individuals and 4 time periods and two explanatory variables, the 'pooled' model with the '*stacked data*' would look like this in vector notation ($i = 1, \ldots, 3$, $t = 1, \ldots, 4$).

[17] This arrangement of the data does not have to be done manually, of course; all programs have special commands for this reorganisation of the data. In Stata there are the commands `xtset` and `reshape`; in R, for example, with the package `reshape2`, which provides the very flexible commands `melt` and `dcast`. Finally, with the package `plm` all kinds of panel models can be estimated.

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

The assumption that the coefficients are the same for all individuals and periods is of course quite restrictive.

A somewhat more general and flexible model, which is most commonly used in practice, allows for individual-specific intercepts but assumes the same slope coefficients for all countries. This can easily be done with the help of appropriate individual-specific dummy variables. For example, we would estimate the following model with OLS

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

where we now denote the coefficients of the dummies and the intercept with a for better readability (in most cases we are not interested in these coefficients of the dummies, which is why they are often suppressed in publications).

Here, the first individual serves as the reference category.

This model is also called a *'Least Squares Dummy Variable'* (LSDV) model and can be written for a total of n individuals and T time periods as follows.

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

with $i = 1, \ldots, n$ and $t = 1, \ldots, T$. Note that the dummies d_i do not require a time index (they are *time invariant*)!

.

Mostly, however, (especially the *'fixed effects'* model) is written in a shorter way

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

where the a_i are symbolic of the individual effects (e.g. country effects), thus symbolising the individual dummies (i.e. $a_i := a_1 + \sum_{i=2}^{n} a_i d_i$).

Imagine you want to estimate this model for a panel with several thousand individuals, then you would need several thousand dummy variables. This would even challenge the processing power of modern computers.

Fortunately, there is an alternative that is as simple as it is elegant. Let us recall the Frisch-Waugh-Lovell (FWL) theorem: we can 'eliminate' the linear influence of variables by computing auxiliary regressions in a first step, and then using the residuals of these auxiliary regressions.

So we could regress y and x on the dummy variables in a first step, and then use the residuals of these auxiliary regressions to calculate the slope coefficients of interest b_h.

At first glance, this does not seem to gain much; we still need all the dummy variables for the auxiliary regressions. But let's consider what we get from a regression of the saturated dummy variable model, exactly, *the group-specific mean values* (e.g. mean values of women and men, country-specific mean values, etc.)! And the residuals of the auxiliary regressions are simply the deviations from these group-specific means.

It is therefore sufficient to form the group means over time for each individual and for all variables, and to calculate the individual-specific deviations from these group means. Such an individual-specific mean transformation can be performed very efficiently and quickly by computers.

So, instead of computing a regression with potentially several thousand dummy variables, we can also perform individual-specific mean transformations and compute a simple OLS regression on the data transformed in this way

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

where $\bar{y}_{i\bullet}$, $\bar{x}_{2i\bullet}$ and $\bar{x}_{3i\bullet}$ are individual-specific time-mean values. This means that we only need to perform the individual-specific mean transformation, and we can run a normal OLS regression on the transformed data.

The model estimated by this method is called a 'fixed effects model' (the individual effects, e.g. country effects, do not change over time, so they are 'fixed').

Due to the FWL theorem, this method numerically leads to the exact same estimates for the slope coefficients as the LSDV model, but is much easier to calculate. But beware, this only applies to the slope coefficients and residuals, the standard errors will differ for these methods because the *'fixed effects model'* does not account for the loss of degrees of freedom in the individual-specific mean transformation. All computer programs that support fixed effects models automatically take this

into account. models automatically take this into account and output the correct standard errors.

One loses the individual effects a_i (i.e., the coefficients of the individual dummies) with this *fixed effects* method, but these are rarely of interest anyway, and could be recalculated ex post to boot.

In summary, due to the FWL theorem, we can interpret the coefficients of the *'fixed effects'* model can be interpreted in the same way as the coefficients of a least squares dummy variable model (LSDV)!

Since this model only takes into account the dispersion over time *"within'* the individuals, the *"fixed effects'* model is also called the *'within'* model.

The particular appeal of the *'fixed effects'* model is that the individual effects (or the dummies for the individuals) control for everything that does not change over time, i.e. for all *time-invariant* effects (such as gender, colonial past, ...).

The individual dummies, in a sense, 'swallow' anything that is time-invariant, whether we can observe it or not, or whether we care about it or not. The consequence of this is that with the help of the *'fixed effects'* model we cannot calculate partial effects of time-invariant variables!

From a purely technical point of view, this is already evident from the fact that individual-specific mean transformations for time-invariant variables always yield the value zero. Most computer programs automatically suppress such variables, other programs abort with an error message.

For example, suppose we have panel data with hourly wages, completed education level, work experience and gender of many individuals over many years.

If we estimate a *'fixed effects'* model, the (implicit) person dummies control for *all* individual-specific effects, e.g. also for the unobservable 'emotional intelligence' (if this does not change over time!), but since gender and completed education are also time-invariant, we cannot measure their influence either, they are, so to speak, all 'stuck' together in the individual dummies. If, on top of that, the work experience for all individuals increases by one year every year, we lose the *'between'* information, leaving only a *'within'* trend, which could equally measure the impact of inflation and the like.

However, if we are interested in variables *with* time variation, the *'fixed effects'* model is extremely powerful, as it automatically controls for *all* time-invariant effects, whether observed or not.

2.7.7 Categories that are not mutually exclusive

In the previous section we examined mutually exclusive cases, each observation could be assigned exactly one expression of the categorical variable; for example, a car cannot be both two and four years old.

Now let's look at the case with several categorical variables that do not have to be mutually exclusive; for example, a person can be female and be married or unmarried as another characteristic.

Let us return again to the example with hourly wages, see 2.7 (page 54). As expected, a regression on the dummy variable $v_i = 1$ for married and 'zero else' as the intercept yields the mean hourly wage of the reference category, i.e. unmarried, and the coefficient of the dummy v shows that married people earn on average 0.6 euros more,

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

.

One might perhaps assume that a regression on both dummy variables ($m_i = 1$ for male and zero otherwise, and $v_i = 1$ for married and zero otherwise) measures the two deviations from the reference category ($m_i = 0$ and $v_i = 0$, i.e. an unmarried woman), but this is not so, as the result shows

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

What happened? We simply made a thinking error, because the two dummy variables define *four* categories, not two! The following table shows the four categories and the respective mean values of wage for this example

		male	
		yes (1)	no (0)
married	yes (1)	15.5	12
	no (0)	14	13

If we only regress on the two categories m and v, we estimate a model that is too short, and the problem of not taking into account relevant variables (*omitted variables'*) arises again.

Only if we choose a model that takes *all possible* categories into account do we obtain as coefficients the mean values of the respective categories.

Such a model is called *saturated*, and only for such saturated dummy variable models is it true that the intercept measures the mean of the reference category, and the coefficients of the dummy variables the corresponding average deviations of the respective category from the reference category.

The simplest way to estimate such a model is to generate a dummy variable for all categories except the chosen reference category. For example, if we choose unmarried men ($m = 1$ and $v = 0$) as the reference category, we generate a dummy variable mv for male married, with $mv = m \times v$, for female unmarried $wu = w \times (1 - v)$, and for female married $wv = (1 - m) \times v$. The regression gives

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

This gives us the expected result, the average hourly wage of unmarried men is 14 euros, married men earn on average 1.5 euros more, unmarried women earn on average one euro less than unmarried men, and married women earn two euros less.

This parameterisation directly provides an output that is very easy to interpret. In the literature, however, one often finds an alternative parameterisation that gives exactly the same result in a different representation, namely a regression on both

dummy variables *and* on the interaction effect (i.e. the product) of the two dummy variables. If we choose as reference category 'female' ($m = 0$) and 'unmarried' ($v = 0$) we get

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

We can think of the dummy variables here as *'on-off switch'* in a sense, if the dummy variable has the value one the coefficient is 'on', otherwise 'off', and the interaction effect is only 1 ('on') if *both* dummy variables have the value 1.

1. Female unmarried: (reference category).

$$\widehat{\text{wage}}|(m = 0, v = 0) = 13.0 + 1.0 \times 0 - 1.0 \times 0 + 2.5(0 \times 0) = 13$$

2. Female married:

$$\widehat{\text{wage}}|(m = 0, v = 1) = 13.0 + 1.0 \times 0 - 1.0 \times 1 + 2.5(0 \times 1) = 12$$

3. Male unmarried:

$$\widehat{\text{wage}}|(m = 1, v = 0) = 13.0 + 1.0 \times 1 - 1.0 \times 0 + 2.5(1 \times 0) = 14$$

4. Male married:

$$\widehat{\text{wage}}|(m = 1, v = 1) = 13.0 + 1.0 \times 1 - 1.0 \times 1 + 2.5(1 \times 1) = 15.5$$

As you can easily convince yourself, this parameterisation gives exactly the same result in a slightly different representation, by taking the interaction effect into account the model is saturated again, so formally it does not matter which of these parameterisations you choose, it is more a matter of expediency.

However, anyone who wants to map all these combinations of characteristics with dummy variables quickly ends up in the *curse of dimensionality*. If, for example, you want to distinguish six levels of education, 10 industries and four regions in addition to (binary) gender and marital status, you will end up with 960 categories to distinguish. Even if the sample were large enough to estimate all the parameters, it might not be easy to present such a result clearly.

2.7.8 Example: *'Difference-in-Differences'* models

Imagine that a new bypass road has been built in a city and you are asked to estimate what impact this has had on property prices *in the affected region*.

This assignment presents you with a typical "what-if" question, because if the road was built, the counterfactual is missing (what would prices be if the road had not been built).

Suppose you had data on land prices *before* the bypass road was built. In this case, you could simply compare the mean of the land prices *before* the construction of the bypass road with the land prices *after* the construction of the bypass road.

However, such a comparison is difficult because if property prices in general changed during the construction of the bypass, one would falsely attribute this price change to the bypass.

In this case, one could compare the prices *before* and *after* the construction of the bypass road with the property prices of a very similar but *not affected by the intervention* region of the city. This is exactly the basic principle of the *"Difference-in-Differences"* approach.

Since this type of analysis used to be applied mainly in medicine, medical terminology has become common in the literature. One calls a group that has been given a treatment (*Intervention*) (or has been affected by a change) *'Treatment Group'*, and the control group unsurprisingly *'Control Group'*.

Where the term *'Difference-in-Differences'* comes from becomes immediately clear when we return to the example. We denote the mean of the land prices of the *'Treatment Group'* (i.e. the group that was affected by the construction) *before* the construction of the bypass road as T_B, the mean of the *'Treatment Group' after* the construction of the bypass road with T_A, and the mean values of the prices of the control group with C_B and C_A, respectively, i.e.

	Treatment Group	Control Group
Before	T_B	C_B
After	T_A	C_A

In order to estimate the '*caused*' by the construction of the bypass road price change we can simply calculate the 'difference of the difference' of the mean values, i.e.

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

But this has only *almost* solved our problem, because we will hardly find enough *comparable* property prices in the groups. Properties differ in terms of size, location, amenities, etc., so it is difficult to compare them.

Fortunately, this "difference-in-differences" approach can very easily be transformed into a regression model, and as is well known, a regression allows for the consideration of several explanatory x variables (such as size, location, amenities).

Specifically, we can estimate the following regression equation

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

with the dummies

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

and one (or more) explanatory variables x.

In the following table, one can easily see that the coefficient of the *interaction term* between the treatment and after dummy is exactly the difference-in-difference estimator.

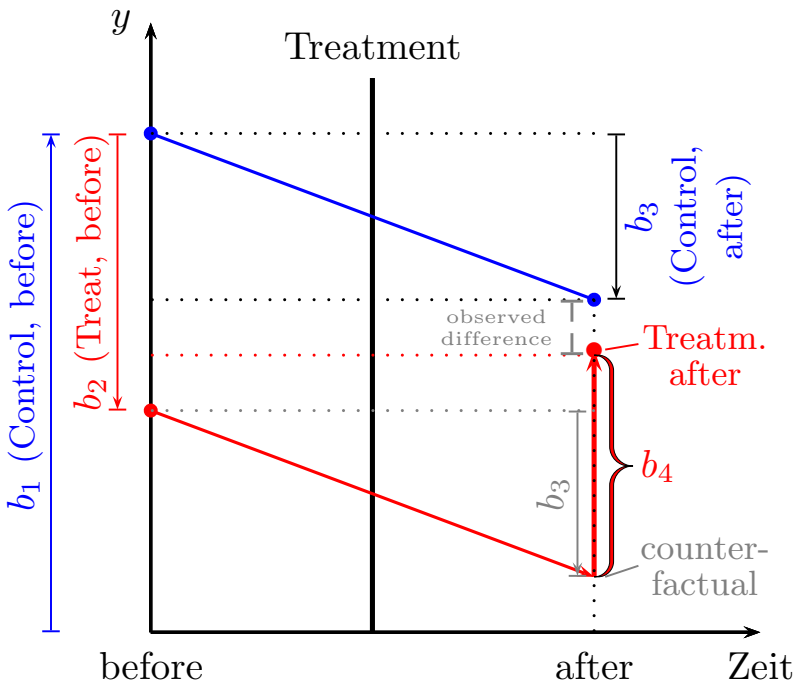


Figure 2.27: Difference-in-Differences, the treatment effect is measured by the coefficient of the interaction term b_4;
$\widehat{y}_i = b_1 + b_2\text{treat} + b_3\text{after} + b_4(\text{treat} \times \text{after})$
(here with $b_1, b_4 > 0$ and $b_2, b_3 < 0$) The 'counterfactual' tells us what value $\widehat{y}$ would have taken if the treatment group had developed the same as the control group. Note that this counterfactual is not directly observable, but is based on the identifying assumption that in the absence of the treatment both groups would have developed in the same way (*Parallel Trends Assumption*).

	Treatment Group	Control Group	Difference
Before	$b_1 + b_2 + b_5x$	$b_1 + b_5x$	b_2
After	$b_1 + b_2 + b_3 + b_4 + b_5x$	$b_1 + b_3 + b_5x$	$b_2 + b_4$
Difference	$b_3 + b_4$	b_3	b_4

In Figure 2.27 this is shown graphically.

Exercise: What do you think, does Figure 2.27 show more of the impact of a waste incinerator or that of a leisure facility on property prices in the area? Which graph would you expect in the other case?

Identifying Assumptions: Can we really be sure that we have measured the effect of the treatment correctly? And what does 'correctly' mean?

We are of course interested in how prices would have developed if, for example, the road had not been built (counterfactual) compared to the actual development of prices after the road was built (fact). Since we can never observe the counterfactual and the fact at the same time, this comparison is not feasible.

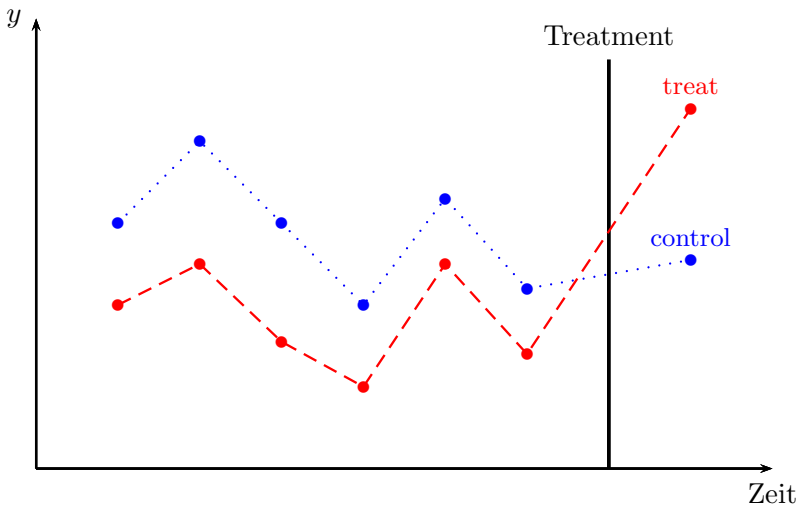


Figure 2.28: A causal interpretation of a difference-in-differences model becomes more 'trustworthy' if *prior to treatment* the two groups have developed in parallel (but this is *not* proof!).

In the 'difference-in-differences' model, of course, the control group serves as an 'approximation' for the unobservable counterfactual. And the question is how good this 'approximation' actually is.

A central assumption is that *without the treatment* in the control and treatment group the same development would have occurred, the so-called *Parallel Trends Assumption*.

This assumption implies that the regression model was correctly specified, e.g. that all relevant variables were taken into account (i.e. no *omitted variables bias*).

But of course we can never be sure that in the hypothetical case *without treatment* the same would have happened in the treatment group as in the control group.

Above all, there is no way to prove *Parallel Trends Assumption*, at best we can make good arguments why we suspect that in the hypothetical case without treatment the same thing would have happened in both groups.

Even if the *Parallel Trends Assumption* cannot be proven, one can at least look for good arguments for its validity. One of these arguments is that – with repeated observations – the trends in both groups *before treatment* should have developed approximately in parallel, see Figure 2.28.

Example: The classic example of an application of the difference-in-differences model in economics is a study on the effects of an increase in the minimum wage by Card and Krueger (1994). Not least for this study, D. Card was awarded the Alfred Nobel Memorial Prize in Economic Sciences in 2021.

On 1 April 1992, New Jersey (NJ) raised the minimum wage from US$4.25 to US$5.05. Card and Krueger (1994) surveyed the number of employees in a telephone survey of 320 fast food restaurants in New Jersey and, as a control group, 77 fast food restaurants in neighbouring Pennsylvania. Each firm was surveyed twice, once before (Feb) and once after (Nov) the introduction of the minimum wage. Fast

Table 2.14: Average number of employees in fast food restaurants before and after the implementation of a minimum wage on April 1, 1992 in New Jersey (NJ). Neighbouring Pennsylvania (PA) serves as a control group. Siehe Card and Krueger (1994).

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

food restaurants were chosen because they have a particularly high proportion of low-paid workers.

To determine the employment effect of the minimum wage increase, they conducted a difference-in-differences analysis. The simple means and their differences can be found in Table 2.14.

To the surprise of many economists, fast food restaurants in the Treatment group (New Jersey) employed *relatively* more people after the minimum wage was raised than in the Pennsylvania control group.

As shown earlier, this result can also be obtained simply by using a regression on the dummies NJ (= treatment group) and Nov (= after) and their interaction (Empl denotes the number of employees).

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

$$R^2 = 0.007, \quad n = 794$$

(Standard errors in parentheses)

However, Card and Krueger (1994) did not estimate this model, but used firm fixed effects instead of the NJ dummy. The use of firm fixed effects is numerically equivalent to considering firm dummies (each firm was surveyed twice, and all but one reference firm received a dummy). This gave them the following result

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

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

(firm-fixed effects, standard errors in parentheses)

(NJ is a dummy for New Jersey, and Nov for November). Since the model was estimated with *fixed effects*, the coefficients of the firm dummies and the intercept are not output; the R code for estimating this equation is

```
rm(list = ls())
d <- read.csv2("https://www.uibk.ac.at/econometrics/data/cardkrueger94.csv",
           dec = ".")
```

```
library(plm)
pd <- pdata.frame(d, index = c("Firm", "Period"))
eq1 <- plm(Emp ~ Nov*NJ, model = "within", data = pd)
summary{eq1}
```

With this specification, the coefficient of the treatment effect is positive and significantly different from zero at the 5% level, which was interpreted to mean that the minimum wage increase had *positive* employment effects.

This result is still very controversial today, see for example. NZZ of 23 April 2014. For a more detailed discussion of the 'difference-in-differences' approach as well as this example, see also Angrist and Pischke (2008, 228), for a more up-to-date and freely available overview of the effects of minimum wages see Manning (2021).

2.8 Logarithmic transformations

> *"If people do not believe that mathematics is simple, it is only because they do not realize how complicated life is."*
>
> (John von Neumann, 1947)

So far, we have dealt almost exclusively with linear models of the type $\widehat{y} = b_1 + b_2x + b_3x_3$, in which the variables are additively linked.[18] However, we are often interested in models in which the variables are multiplicatively linked, such as in the well-known Cobb-Douglas function $Q = AK^{\alpha}L^{\beta}$, probably the best-known example of an exponential function. Such multiplicative models also play a major role in modelling growth processes. We will see in a moment that such models can be easily linearised and estimated with the help of logarithmic functions. But since possibly already the term 'logarithm' causes slight goose bumps for some, and is not infrequently displaced with other unpleasant memories of school days, we start with a short recapitulation.

2.8.1 Review exponential and logarithm functions

A power is a term of the kind a^x, where a is called the base and x is called the exponent (superscript). The familiar rules of calculation apply to powers, such as $a^x a^y = a^{x+y}$. An *exponential function* has the form $y = ca^{bx}$ (or, in alternative notation $x \mapsto ca^{bx}$) with $x \in \mathbb{R}$, where the base a as well as c and b are fixed numbers.

Such exponential functions are suitable, among other things, for modelling growth processes in which a quantity y changes by a *constant factor* in time intervals of equal length. Suppose the value of y is y_0 in the initial period, and this value increases by 5% in each time period. If we express the time period by the subindex, it follows $y_0 = y_0(1 + 0.05)^0, y_1 = y_0(1 + 0.05)^1, y_2 = y_1(1 + 0.05) = y_0(1 + 0.05)^2, \ldots, y_T =$

[18] Since this section is concerned exclusively with functional processes, we will dispense here with the observation index i.

$y_0(1+0.05)^T$ with $t = 0, 1, \ldots, T$. So this type of growth can be described by a simple exponential function with base $(1 + 0.05)$ and exponent t.

For this example, we have assumed discrete periods of equal length. If we let the period length go to zero we get a continuous growth function $y_t = y_0 \mathrm{e}^{rt}$, where $\mathrm{e} = 2.718281828459\ldots$ denotes Euler's number [19] and r denotes the *continuous* growth rate (see Appendix). Because of these and some other properties, e is often called *natural basis* and also written as $\exp(\cdot)$ (i.e. $\exp(x) := \mathrm{e}^x$). In econometrics, the natural basis e is used almost exclusively, so we will restrict ourselves to this basis in the remainder of this paper.

The *natural logarithm* is the solution of the exponential function $y = \mathrm{e}^x$ to x, i.e. the logarithm function is the inverse function to the exponential function.

Since most computer programs use the log operator for the natural logarithm, we follow this notation here, i.e., log denotes the natural logarithm in the following. Therefore, $x = \log(y)$ and $\log(\mathrm{e}^x) = x$.

In somewhat casual terms, $\log(y)$ tells us how many times we have to multiply the base e by itself to get y. Note that the exponential function $x \mapsto \mathrm{e}^x$ maps the set of real numbers $\mathbb{R}$ into the positive real numbers $\mathbb{R}^+$, since $\mathrm{e}^{-x} = 1/\mathrm{e}^x$. Therefore, the logarithm function is only defined for the positive real numbers, it maps $\mathbb{R}^+ \mapsto \mathbb{R}$; or more simply, the logarithm of negative numbers is not defined!

The importance of the logarithmic transformation for econometric applications follows essentially from three properties:

1. *Multiplicative relations can be represented additively by logarithmisation*, respectively, exponential functions become linear functions by logarithmisation.

$$\log(xy) = \log(x) + \log(y) \qquad \text{for } x, y > 0$$

Why? To show this, we define two numbers a and b such that $\log(x) = a$ and $\log(y) = b$. It follows that $x = \mathrm{e}^a$ and $y = \mathrm{e}^b$ with $xy = \mathrm{e}^a\mathrm{e}^b = \mathrm{e}^{a+b}$ because of the rules of calculation for powers. Therefore $\log(xy) = \log(\mathrm{e}^{a+b}) = a + b := \log(x) + \log(y)$.

The most important calculation rules are

$$\begin{aligned}
\log(xy) &= \log(x) + \log(y) \qquad \text{for } x, y > 0 \\
\log\left(\frac{x}{y}\right) &= \log(x) - \log(y) \\
\log(x^a) &= a\log(x) \\
\log(1/x) &= -\log(x)
\end{aligned}$$

Therefore, for example, the Cobb-Douglas function $Q = AK^\alpha L^\beta$ can be linearised to $\log(Q) = \log(A) + \alpha\log(K) + \beta\log(L)$.

[19] note the difference between Euler's number e and the residuals e.

2. The difference between two logarithmised values corresponds approximately to a *relative* change in the original values, i.e.

$$\log(x_2) - \log(x_1) \approx \frac{x_2 - x_1}{x_1} := \frac{\Delta x}{x} \qquad \text{for } x_2, x_1 > 0$$

We will see in a moment that this property greatly simplifies the interpretation of regression coefficients of logarithmic variables.

For an intuitive understanding, let us recall the derivation rule for the natural logarithm

$$\frac{d\log(x)}{dx} = \frac{1}{x}$$

We can think of this rewritten as

$$d\log x = \frac{dx}{x}$$

and recall that we can interpret $d(\log x)$ as an infinitesimally small change in $\log(x)$. Similarly, we can think of dx as an infinitesimally small change in x, which is why dx/x represents an (infinitesimally small) *relative* change of x.

By analogy, we would expect *approximate* ('≈') to hold for discrete cases.

$$\log(x + \Delta x) - \log(x) \approx \frac{\Delta x}{x}$$

Note: If we have two concrete points x_0 and x_1 in mind we can also write $[x_0+(x_1-x_0)] = x_1$ for $(x+\Delta x)$, i.e. for the above expression $\log(x_1)-\log(x_0) \approx (x_1 - x_0)/x_0$. □

This relation actually holds if $\Delta x/x$ is 'relatively' small (see 'Excursus: Logarithmic difference and relative rates of change', page 86).

So we hold: the difference of a logarithmised variable approximately measures the relative change of the original variable.

If you multiply a *relative change* by 100 you get the percentage change.

Note that this is a property of the log-log functional form, and it follows from this property, as we will explain in more detail in a moment, that not as with the linear functional form $\widehat{y} = b_1 + b_2 x$ is the slope $d\widehat{y}/dx$ constant over the entire course of the function, but that for a log-log functional form $\widehat{\log(y)} = b_1 + b_2 \log(x)$ the ratio of *relative changes* $(d\widehat{y}/\widehat{y})/(dx/x) = b_2$ is constant over the entire course of the function!

3. Logarithmisation spreads small numerical values and compresses large numerical values, see Figure 2.29. This sometimes has the effect of reducing the influence of extreme observations on the estimate, or making skewed distributions more symmetric. A classic example of this is income data and wages, cf. Figure 2.30.

$$
\begin{aligned}
\log(0) &= -\infty \\
\log(0.000001) &= -13.816 \\
\log(0.01) &= -4.605 \\
\log(0.1) &= -2.303 \\
\log(1) &= 0 \\
\log(10) &= 2.303 \\
\log(100) &= 4.605 \\
\log(1000) &= 6.908 \\
\log(1000000) &= 13.816
\end{aligned}
$$

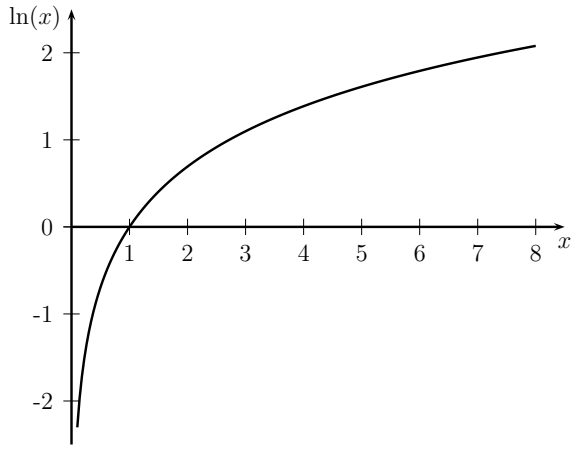


Figure 2.29: Logarithmisation spreads small numerical values and compresses large numerical values

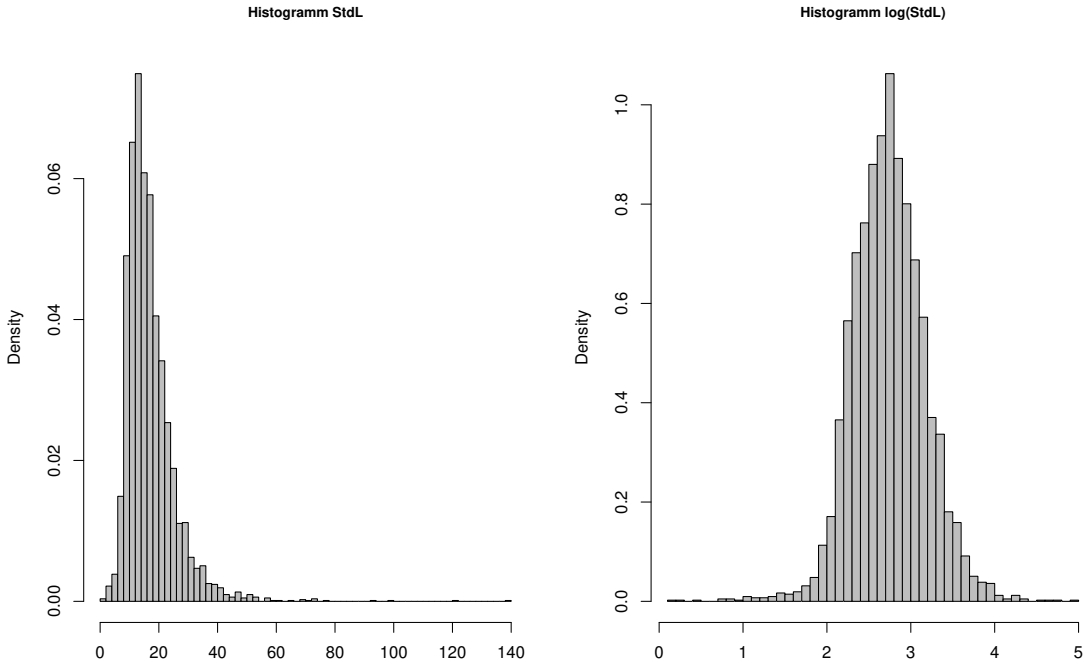


Figure 2.30: Histograms of gross hourly wage (StdL) and logarithm of gross hourly wage (hourly wages < 1 and > 200 have been removed); log(hourly wages) are more symmetrically distributed!
Source: EU-Silc 2018

Excursus: Logarithmic difference and relative rates of change

We have asserted in the text that the logarithmic difference of a variable is approximately equal to the relative change in that variable

$$\log(x + \Delta x) - \log(x) \approx \frac{\Delta x}{x} \qquad \text{for small } \frac{\Delta x}{x}$$

This can be shown in general using a Taylor expansion. With a Taylor series expansion, nonlinear (differentiable ...) functions in the neighborhood of certain points can be represented by power series. In particular

$$\log(1 + x) = \sum_{n=1}^{\infty} (-1)^{(n+1)} \frac{x^n}{n} = x - \frac{x^2}{2} + \frac{x^3}{3} - \ldots \qquad \text{for } |x| < 1$$

resp. for $x = \frac{\Delta y}{y}$

$$\log\left(1 + \frac{\Delta y}{y}\right) = \frac{\Delta y}{y} - \frac{1}{2}\left(\frac{\Delta y}{y}\right)^2 + \frac{1}{3}\left(\frac{\Delta y}{y}\right)^3 - \ldots$$

The left-hand side can also be written as a logarithmic difference, since

$$\log\left(1 + \frac{\Delta y}{y}\right) = \log\left(\frac{y + \Delta y}{y}\right) = \log(y + \Delta y) - \log(y)$$

. From this follows

$$\log(y + \Delta y) - \log(y) \approx \frac{\Delta y}{y}$$

since for small $\Delta y/y$ the consequent terms $-1/2\,(\Delta y/y)^2 + 1/3\,(\Delta y/y)^3 - \ldots$ of the series become very small and are often negligible for practical purposes.

2.8.2 Interpretation of the coefficients of logarithmised variables

For OLS estimates themselves, it does not directly matter whether the variables have been logarithmised or not. However, it does change the interpretation of the coefficients depending on whether only the dependent variable, only the explanatory variable or both have been logarithmised.

For the understanding of what follows, only two facts are important, *first* that, as emphasised earlier, a log difference is approximately equal to a relative change, i.e. $d\log(y) = \frac{dy}{y}$ or for discrete changes

$$\Delta\log(y) := \log(y + \Delta y) - \log(y) \approx \frac{\Delta y}{y}$$

and *second* that the marginal effect can usually be written as a simple difference quotient, i.e. for $\widehat{y} = b_1 + b_2 x$ the marginal effect is

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

where the second '=' sign only holds exactly for *linear* functional forms, but we would hope that this should also hold approximately for non-linear functional forms, at least for small changes.

As mentioned earlier, the nonlinear exponential function $\widehat{y} = b_0 x_2^{b_2}$ becomes linear in parameters by logarithmising, i.e.

$$\widehat{\log(y)} = b_1 + b_2\log(x_2) \quad (\text{with } b_1 = \log(b_0))$$

The marginal effect of $\log(x_2)$ is, as usual, the difference quotient

$$b_2 = \frac{d\,\widehat{\log(y)}}{d\,\log(x)} = \frac{\frac{d\widehat{y}}{\widehat{y}}}{\frac{dx}{x}} \qquad \text{or discrete} \qquad b_2 = \frac{\Delta\widehat{\log(y)}}{\Delta\log(x)} \approx \frac{\frac{\Delta\widehat{y}}{\widehat{y}}}{\frac{\Delta x}{x}}$$

(note the '≈' sign in the discrete representation).

If – as in this case – the dependent *and* the explanatory variable is logarithmised, this is called a log-linear or better log-log model. If, on the other hand, only the dependent *or* only the explanatory variable is logarithmised, one speaks of a semi-log model, or sometimes, in the case of $\widehat{\log(y)} = b_1 + b_2 x$, of an *log-level*, or, in the case of $\widehat{y} = b_1 + b_2\log(x)$, of an *level-log* model.

2.8.3 Log-log (resp. log-linear) models

As already mentioned, the exponential function

$$y_i = b x_i^{b_2}\exp(e_i)$$

be linearised by logarithmisation (cf. Figure 2.31).

$$\log y_i = b_1 + b_2\log x_i + e_i \qquad \text{with } b_1 := \log b$$

This model is *linear in parameters* and can therefore be estimated normally with OLS. The *marginal effect*

$$b_2 = \frac{d\log(y)}{d\log(x)}$$

can be interpreted as usual, namely by how many units $\log(y)$ changes when $\log(x)$ increases by *one unit*, but how would you explain this to a layman? What should you think of as one unit of $\log(x)$?

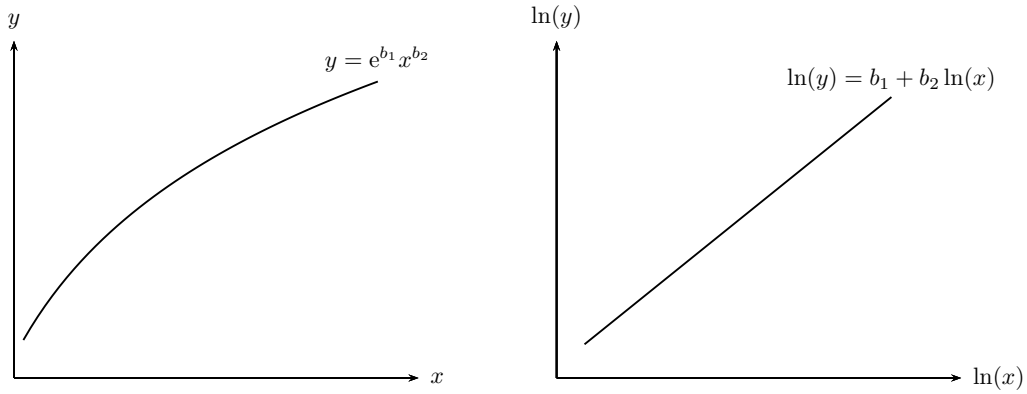


Figure 2.31: Log-log model with logarithmic and linear scaled scale

Here we are helped by the property of logarithmic functions mentioned above, that the absolute difference between two logarithmed values is approximately equal to the *relative* difference of the original values (i.e. $\Delta \log y \approx \Delta y/y$), so[20]

$$\boxed{b_2 = \frac{d\log(y)}{d\log(x)} = \frac{\frac{dy}{y}}{\frac{dx}{x}} \approx \frac{\frac{\Delta y}{y} \times 100}{\frac{\Delta x}{x} \times 100} = \frac{\text{percentage change of } y}{\text{percentage change of } x} = \text{elasticity}_{y,x}}$$

where an elasticity is defined as the *ratio of two relative (or percentage) changes*. Therefore, we can directly interpret the coefficients of log-log models as elasticities.

> $\Rightarrow$ The coefficient in a log-log model indicates by how many *percent* the dependent variable y changes (ceteris paribus) when the explanatory variable x increases *by one percent*.

where, of course, the ceteris paribus clause only applies to multiple regressions with respect to the other regressors considered.

[20] Remember that $de^x/dx = e^x$ and $d\log(x)/dx = 1/x$.
We can now differentiate either $y = \exp(b_1 + b_2 \log(x))$ to x by applying the chain rule

$$\frac{dy}{dx} = b_2 \frac{1}{x} \underbrace{[\exp(b_1 + b_2 \log(x))]}_{y} \quad \Rightarrow \quad b_2 = \frac{dy}{dx}\frac{x}{y} = \frac{\frac{dy}{y}}{\frac{dx}{x}}$$

or alternatively we can differentiate $\log(y) = b_1 + b_2 \log(x)$ totally

$$\frac{1}{y}\, dy = b_2 \frac{1}{x}\, dx \quad \Rightarrow \quad b_2 = \frac{dy}{dx}\frac{x}{y} = \frac{\frac{dy}{y}}{\frac{dx}{x}}$$

So for infinitesimal changes this relation holds exactly.

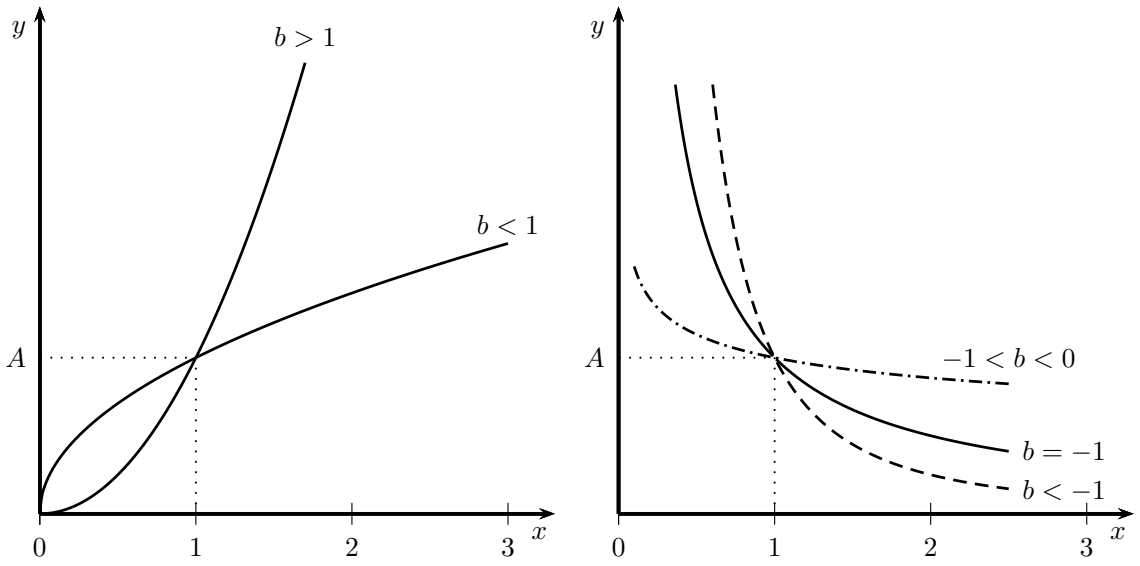


Figure 2.32: Log-log (or log-linear) models: progressions of the exponential function $y = Ax^b$ for different b.

Example: Suppose we have the function

$$\log(y) = 1 + 0.2\log(x)$$

. How does y change when x increases by *one percent*? Based on the discussion above, we expect y to increase by 0.2 *percent*. Is this really true? We can easily check this by rewriting the above function as $y = \exp[1+0.2\log(x)]$ and substituting two values for x, e.g. 5 and 5.05.

x	$(\Delta x)/x$	$\%\Delta x$	$y = \exp(1 + 0.2\log(x))$	$(\Delta y)/y$	$\%\Delta y$
5.00			3.750494		
5.05	$\frac{5.05-5}{5} = 0.01$	1%	3.757965	0.001992	$\approx 0.2\%$

We see that this does not hold exactly, since a change of one percent is not an infinitesimally small change, but for practical purposes this approximation is more than sufficient in the vast majority of cases. □

Figure 2.32 shows possible courses for positive (left) and negative b (right) of the exponential function $y = Ax^b$, which becomes linear by logarithmisation $\log(y) = \log(A) + b\log(x)$.

Example Cobb and Douglas (1928) estimated the following production function for the USA from 1899–1922.

$$\begin{array}{rcccccc} \log(Q_i) & = & -0.177 & + & 0.807\log(L_i) & + & 0.233\log(K_i) + e_i \\ & & (0.434) & & (0.145) & & (0.064) \end{array}$$

$$R^2 = 0.957, \quad n = 24$$

(standard error in parentheses)

Both slope coefficients have the expected sign and are significantly different from zero. The coefficient of $\log(L_i)$ indicates the output elasticity of the factor labour, i.e. if labour input is increased by 1% we expect ceteris paribus an increase in output of 0.807%. Similarly, if capital input increases by 1% we expect output to increase by 0.233% ceteris paribus.

caution: The relationship $y = \alpha x^{\beta}$ can be econometrically modelled in different ways, i.e. the perturbation terms can enter the model in different ways

$$\begin{array}{llllllll} 1) & y_i & = & \alpha x_i^{\beta} \exp(e_i) & \Rightarrow & \log y_i & = & \log\alpha + \beta \log x_i + e_i \\ 2) & y_i & = & \alpha x_i^{\beta} e_i & \Rightarrow & \log y_i & = & \log\alpha + \beta \log x_i + \log e_i \\ 3) & y_i & = & \alpha x_i^{\beta} + e_i & \Rightarrow & \log y_i & = & \log(\alpha x_i^{\beta} + e_i) \end{array}$$

Only the first equation can be estimated directly by OLS! be estimated!

The second equation can be estimated, but this usually has undesirable implications for the residuals e_i (not to be confused with Euler's constant e), because if $\log(e_i)$ is normally distributed, ie. i.e. $\log(e_i) \sim \mathrm{N}(0, \sigma^2)$, then the residuals e_i are log-normally distributed with a positive expected value $\exp(\sigma^2/2)$ (see excursus page 90)! In particular, this also has implications for forecasts when the fitted values $\widehat{\log(y)}$ have been estimated but we are interested in $\widehat{y}$, see e.g. Wooldridge (2012, 204ff).

Finally, the third equation $\log y_i = \log(\alpha x_i^{\beta} + e_i)$ is not linear in the parameters, since $\log(A + B) \neq \log A + \log B$, and therefore cannot simply be estimated with OLS!

2.8.4 Log-level (or log-lin) models

In the log-level model, only the dependent variable y is logarithmised, but not the explanatory x variable.

$$\widehat{\log(y)} = b_1 + b_2 x$$

Possible function curves for a positive and negative slope coefficient are shown in Figure 2.33.

The marginal effect of x on $\widehat{\log(y)}$ is, as usual, the derivative $\frac{d\widehat{\log(y)}}{dx} = b_2$, but this is a bit difficult to interpret, since not many people probably have an idea of a change in $\widehat{\log(y)}$.

We can of course calculate the usual marginal effect dy/dx of $d\log(y) = b_1 + b_2 x$

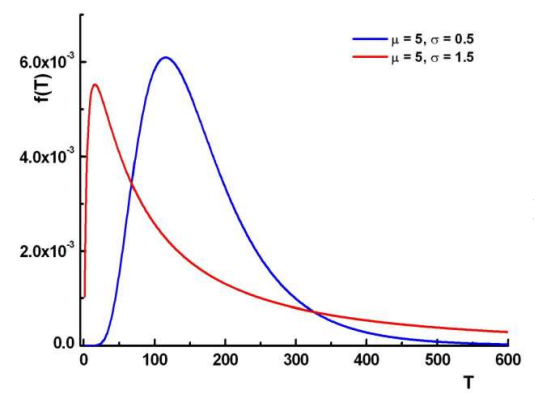
$$\frac{d\log(y)}{dx} = \frac{1}{y}\frac{dy}{dx} = b_2 \quad \Rightarrow \quad \frac{dy}{dx} = b_2 y = b_2(b_1 + b_2 x)$$

but obviously the magnitude of this usual marginal effect is not constant, but depends on the expression of x! Therefore, the specification of this marginal effect is quite uncommon in log-level models.

Excursus: The Log Normal Distribution (Logarithmic Normal Distribution

A random variable whose natural logarithm is normally distributed is log-normally distributed.

That is, if $\log(X) \sim \mathrm{N}(\mu, \sigma^2))$, then X is log-normally distributed.
The other way around, if a random variable Y is normally distributed, then $X = \exp(Y)$ is log-normally distributed.
The log-normal distribution is right skewed and can only take positive values.



par

A log-normal distribution is often used to model random variables, which can be thought of as the (multiplicative) product of many small independent factors.
Expected value and variance are

$$\begin{aligned} \mathrm{E}(X) &= \exp\left(\mu + \frac{\sigma^2}{2}\right) \\ \mathrm{var}(X) &= \exp(2\mu + 2\sigma^2)[1 - \exp(-\sigma^2)] \\ \mathrm{Median}(X) &= \exp(\mu) \end{aligned}$$

For $\mu = 0$ the mean $\exp(\sigma^2/2)$ and the variance $\exp(\sigma^2)(\exp(\sigma^2 - 1))$.

□

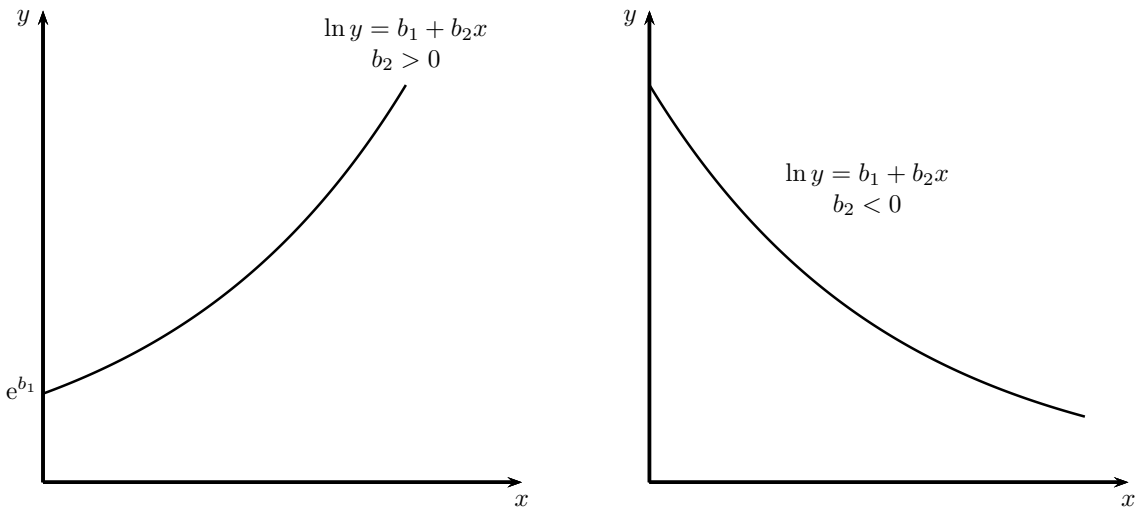


Figure 2.33: Log-level Modelle: $\log y = \alpha + \beta x$

But if we remember again that for infinitesimally small changes $d\log(y) = dy/y$ holds, we can write[21].

$$b_2 = \frac{d\log(y)}{dx} = \frac{\frac{dy}{y}}{dx} \approx \frac{\frac{\Delta y}{y}}{\Delta x}$$

where – as we will show in a moment – the approximation is sufficiently accurate only for *'small'* Δx and *small* b_2.

But even this is difficult to communicate, if x increases by a small unit, does $\Delta(y)/y$ change by b_2? It is much easier if we multiply the left *and* right side by 100, then we get a *percentage* change of y.

$$100 \times b_2 \approx \frac{\frac{\Delta y}{y} \times 100}{\Delta x}$$

or in words:

> $\Rightarrow$ If $\log(y) = b_1 + b_2 x$ and if x increases by *a small unit*, y changes (ceteris paribus) *approximately* by $100 \times b_2$ *percent*.

For the log-level functional forms, this interpretation holds for *all* x, i.e. the marginal effect expressed in this way does not depend on the expression of x! This makes the interpretation much easier.

This expression is sometimes also called a *semi-elasticity*, since the relationship between a *percentage* change of y and an *absolute* change of x is described.

However, this is only *approximately* true for discrete changes. As Figure 2.34 shows, an increase of x by *one unit* (i.e. $\Delta x = 1$) leads to a change of y by Δy units, the slope of the tangent dy/dx in the starting point describes this only *approximately*.

For highly curved curves (i.e. large b_2) and large Δx this approximation is less accurate.

But this inaccuracy can be easily corrected, to determine the effects of such discrete changes from x to y we form differences. For $\log(y) = b_1 + b_2 x$ we obtain by simple transformations

$$\begin{aligned}
\Delta\log(y) := \log(y + \Delta y) - \log(y) &= b_1 + b_2(x + \Delta x) - (b_1 + b_2 x) \\
\log\left(\frac{y + \Delta y}{y}\right) &= b_2 \Delta x \\
1 + \frac{\Delta y}{y} &= \exp(b_2 \Delta x) \\
\frac{\Delta y}{y} &= \exp(b_2 \Delta x) - 1 \\
\left(\frac{\Delta y}{y}\right) \times 100 &= [\exp(b_2 \Delta x) - 1] \times 100
\end{aligned}$$

[21] For example, by totally differentiating $\log(y) = b_1 + b_2 x$ $\frac{1}{y} dy = b_2 dx$ and solving for b_2, or alternatively by differentiating $y = \exp(b_1 + b_2 x)$ for x and rewriting

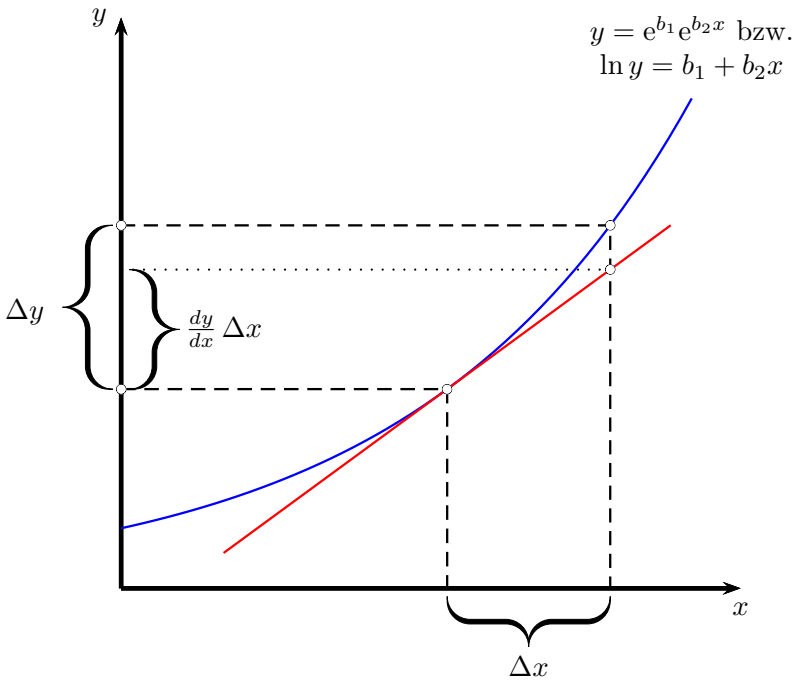


Figure 2.34: effect of a discrete change of x (i.e. Δx) on y.

resp.

$$\%\Delta y := \left(\frac{\Delta y}{y}\right) \times 100 = [\exp(b_2) - 1] \times 100 \qquad \text{for } \Delta x = 1$$

For small b_2 (e.g. $b_2 < 0.1$) $[\exp(b_2) - 1] \approx b_2$ holds (cf. Figure 2.35), so this correction is often omitted for small b_2, e.g. for $b_2 < 0.1$.

Log-level models with dummy variables When interpreting the coefficients of dummy variables in log-level models, it should be noted that dummy variables cannot change infinitesimally by definition, so the difference between the two values of the

b_2	$\exp(b_2) - 1$
−0.50	−0.393
−0.30	−0.259
−0.10	−0.095
−0.01	−0.010
0.01	0.010
0.05	0.051
0.10	0.105
0.30	0.350
0.50	0.649

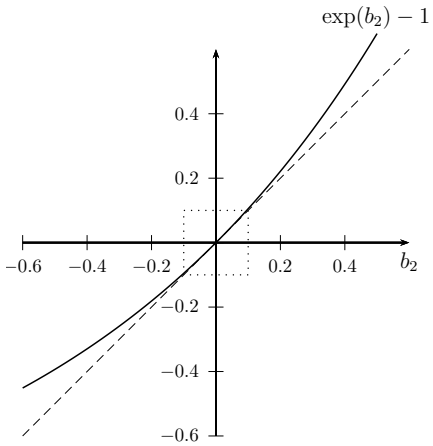


Figure 2.35: For small values of b_2 (e.g. $|b_2| \leq 0.1$) the difference between b_2 and $\exp(b_2) - 1$ is often negligible.

dummy variable must be formed. If d is a dummy variable and $\log(y_i) = b_1 + b_2 d_i$, then the differences are

$$\begin{aligned}
\Delta \log(y_i) := [\log(y_i)|d_i = 1] - [\log(y_i)|d_i = 0] &= b_1 + b_2 \times 1 - b_1 - b_2 \times 0 \\
\log\left(\frac{y_i|d_i = 1}{y_i|d_i = 0}\right) &= b_2 \\
\frac{y_i|d_i = 1}{y_i|d_i = 0} - 1 &= \exp(b_2) - 1 \\
\left[\frac{(y_i|d_i = 1) - (y_i|d_i = 0)}{(y_i|d_i = 0)}\right] \times 100 &= [\exp(b_2) - 1] \times 100
\end{aligned}$$

i.e. to find the percentage difference in $\widehat{y}$ between the two categories defined by the dummy variable we have to apply the same correction again (when b_2 is 'large'):

> $\Rightarrow$ The average *percentage* Difference between the two categories defined by a dummy variable is
>
> $$[\exp(b_2) - 1] \times 100$$

For $|b_2| < 0.1$, $[\exp(b_2) - 1] \approx b_2$ holds again.

Example 1: Based on EU-Silc data for Austria (2018), the following wage equation was estimated, where 'wage' denotes the hourly wage of employed persons, 'educ' is the potential education time in years (age at job entry minus 6), 'exper' is the work experience in years and 'female' is a dummy variable.

$$\begin{array}{rccccccc}
\log(\text{wage}) = & 2.087 & + & 0.038\ \text{educ} & + & 0.011\ \text{exper} & - & 0.139\ \text{female} \\
& (0.03)^{***} & & (0.002)^{***} & & (0.001)^{***} & & (0.013)^{***}
\end{array}$$

$$R^2 = 0.161, \qquad n = 4104$$

According to this estimate, the hourly wage increases ceteris paribus by 3.8% with each year of potential education (more precisely: $(\exp(0.038) - 1)100 = 3.87\%$). Due to the log-level functional form, this applies to all levels of education.

According to this estimate, women earn on average and ceteris paribus about 13 *percent* less than men because $(\exp(-0.139) - 1) * 100 = -13\%$.

Example 2: Consider the function

$$\log(y) = 1 + 0.2x$$

. How does y change when x increases by *one unit* from 5 to 6? Based on the discussion above, we expect y to increase by $[(\exp(0.2) - 1] \times 100\% = 22.14$ *percent*.

Again, we can easily show this numerically by rewriting the above function to $y = \exp(1 + 0.2x)$ and substituting two values for x, e.g. 5 and 6.

x	Δx	$y = \exp(1 + 0.2x)$	$(\Delta y)/y$	$\%\Delta y$
5		7.3890561		
6	1	9.0250135	0.2214	22.14%

Since in this case the coefficient $b_2 = 0.2$ is relatively large we have to make the correction $[(\exp(0.2) - 1]100\% = 22.14\%$, i.e. the increase of x by *one unit* leads to an increase of y by 22.14 *percent*.

□

Example 3: Calculating average growth rates using OLS Using a simple log-level regression on trend[22] an average growth rate can be calculated.

If g is the discrete growth rate of a variable y holds.

$$y_t = y_0(1 + g)^t \quad \Rightarrow \quad \log y_t = \log y_0 + \log(1 + g) \times t$$

This relationship should hold for each period. Therefore, to estimate the discrete growth rate g we can replace t with a trend variable trend $= 1, 2, 3, \ldots, T$. If by y_0 we denote the value of y in the initial period is

$$\log y_t = \underbrace{\log y_0}_{b_1} + \underbrace{\log(1 + g)}_{b_2} \times \text{trend}_t$$

So we can simply

$$\log y_t = b_1 + b_2 \text{trend}_t + e_t$$

and calculate the average discrete growth rate g from $b_2 = \log(1 + g)$, because from

$$b_2 = \log(1 + g) \quad \text{follow} \quad g = \exp(b_2) - 1$$

The percentage average discrete growth rate is therefore

$$\boxed{g\% := g \times 100 = [\exp(b_2) - 1] \times 100}$$

If b_2 is very small (e.g. less than 0.1) b_2 will differ only slightly from $\exp(b_2) - 1$, but for larger values the correction should be applied.

Example The following regression was estimated for China's real GDP per capita (Data source: World dataBank, WDI; Dependent variable: GDP, PPP, constant 2005 international dollar.)

$$\begin{array}{rcccc} \log(\text{GDPpc}) & = & -\ 160.669 & + & 0.084\ \text{Trend} \\ & & (2.71)^{***} & & (0.001)^{***} \end{array}$$

$$R^2 = 0.994, \qquad n = 25 \quad (1995 - 2019)$$

(Standard errors in parentheses)

Script 2.6: Average growth rate of per capita income in China, `R-Code`

```
# Average growth rate of per capita income in China
# Data: World Development Indicators (WDI), World Bank

rm(list = ls())
# install.packages(WDI)
# see e.g. https://cengel.github.io/gearup2016/worldbank.html
library(WDI)
WDIsearch(string = "GNI␣per␣capita,␣PPP", field = "name",
          short = TRUE) # search variable

# download data
china <- WDI(country = "CN",
             indicator = c("NY.GNP.PCAP.PP.KD"),
             start = 1995, end = 2021)
# data.frame china, sort by year in ascending order
china <- china[order(china$year, decreasing = FALSE), ]

GDPpc <- china$NY.GNP.PCAP.PP.KD
Trend <- china$year

eq <- lm(log(GDPpc) ~ Trend)
b2 <- coef(eq)["Trend"]
WR <- (exp(b2) - 1)*100
# Ausgabe am Bildschirm
paste("Discrete␣annual␣growth␣rate␣of␣Chinese␣per␣capita␣income
␣␣␣␣␣␣income␣from", min(Trend), "␣to␣", max(Trend), ":",
      round(WR ,2), "percent")  ## 8.49%
```

According to this estimate, over the period 1995 – 2019, China's real per capita income increased on average *annually* by $(\exp(0.084) - 1) \times 100 \approx 8.81$ *percent*, see Script 2.6.

note We might ask in what time period does income double if it grows at a natural growth rate r. To do this, we just have to

$$2Y_0 = Y_0 \mathrm{e}^{rt}$$

to t. Logarithmising gives $\log(2) = rt$ or $t = \log(2)/r \approx 0.7/r$. Multiplying the numerator and denominator by 100 gives us a percentage growth rate, so[23]

$$\text{doubling time} \approx \frac{70}{r\%}$$

So with a growth rate of 10%, income would double approximately every seven years!

[22] A trend variable increases by one unit with each observation, e.g. trend $= 1, 2, 3, \ldots, T$.

[23] Of course, this is also approximately true for a discrete growth rate g. We solve $2Y_0 = Y_0(1+g)^t$ after the doubling time t and get $t = \log(2)/\log(1+g)$. For small g, $\log(1+g) \approx g$ holds, since as shown in Excursus page 86

$$\log(x + \Delta x) - \log(x) = \log\left(\frac{x + \Delta x}{x}\right) = \log\left(1 + \frac{\Delta x}{x}\right) \approx \frac{\Delta x}{x}$$

Therefore, for $g := \Delta x/x$ it follows $\log(1 + g) \approx g$, i.e. this relation holds *approximately* also for discrete growth rates.

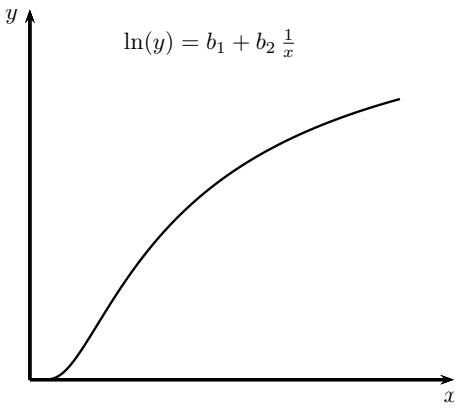


Figure 2.36: Log-Reziproke Transformationen

Example 4: A special specification is

$$\log(y) = b_1 + b_2 \frac{1}{x}$$

This log-inverse model allows the modelling of first increasing and then decreasing marginal effects, cf. Figure 2.36. Such a model could be used, for example, to explain sales turnover as a function of advertising expenditure. The S-shaped functional form allows first increasing marginal returns from advertising expenditure, and from the turning point at $b_2/2$ decreasing marginal returns, and finally asymptotically approaches a horizontal course.

2.8.5 Level-log (or lin-log) models

In the level-log model, the explanatory variable is logarithmised rather than the dependent variable. A graphical representation of the level-log model

$$y_i = b_1 + b_2 \log x_i$$

can be found in Figure 2.37.
We can write this down again in changes[24]

$$\Delta y = b_2 \Delta \log(x) \quad \Rightarrow \quad b_2 = \frac{\Delta y}{\Delta \log(x)} \approx \frac{\Delta y}{\frac{\Delta x}{x}} = \frac{\text{absolute change of } y}{\textit{relative} \text{ change from } x}$$

To get a percentage change of x in this case we need to *divide* the left and right sides by 100

$$\frac{b_2}{100} \approx \frac{\Delta y}{\frac{\Delta x}{x} \times 100} = \frac{\text{absolute change of } y}{\textit{percentage} \text{ change from } x}$$

i.e. an increase of x by *one per cent* leads ceteris paribus to an absolute change of y by $0.01 \times b_2$ *units*.

[24]For infinitesimal changes we differentiate $y = b_1 + b_2 \log(x)$ total and get $dy = b_2 \frac{1}{x} dx$ from which follows

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

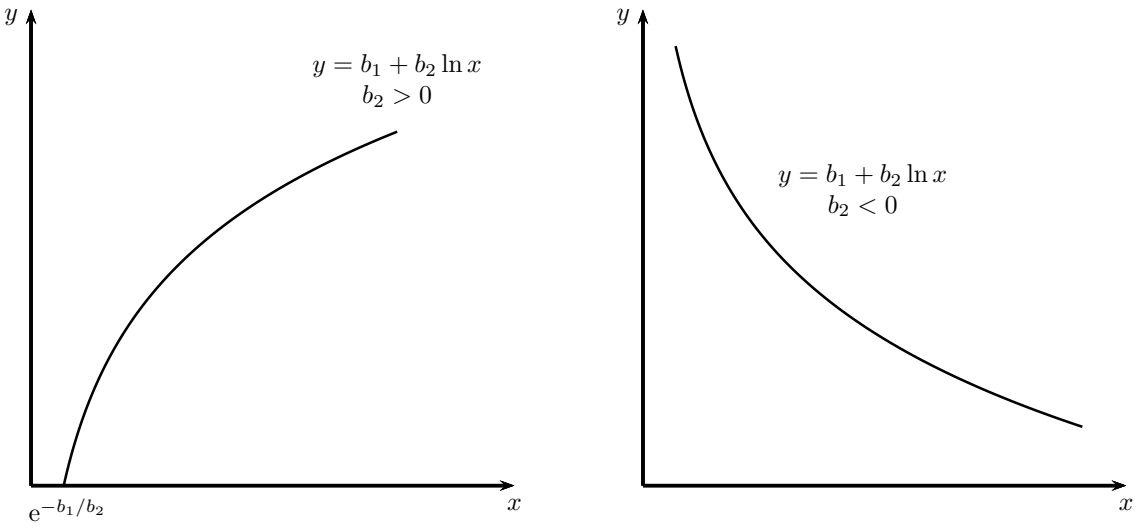


Figure 2.37: Lin-log Modell: $y = b_1 + b_2 \log x$

Example 1: Suppose we have the function

$$y = 1 + 0.2 \log(x)$$

How does y change when x increases by *one percent*? Based on the discussion above, we expect y to increase by 0.002 *units*.

We check this again by substituting two values for x into the above function, e.g. 5 and 5.05.

x	$(\Delta x)/x$	$\%\Delta x$	$y = 1 + 0.2 \log(x)$	Δy
5.00			1.3218876	
5.05	$\frac{5.05-5}{5} = 0.01$	1%	1.3238777	0.00199

Since a change of one percent is not an infinitesimally small change, this does not apply exactly, but for practical purposes this approximation is usually sufficient. □

Example 2: Figure 2.38 shows the relationship between life expectancy (LE; Life expectancy at birth, total (years)) and per capita income (GNIpC; GNI per capita, PPP (current international $)) for a cross-section of countries and the year 2018. The left panel shows the level-level graph, obviously a linear regression describes this relationship much worse than a level-log regression. The right panel shows the relationship when log per capita income is plotted on the x axis.

Table 2.15 shows the associated regressions.

Interpretation: If per capita income increases by *one percent* we expect, based on this estimate (ceteris paribus), an increase in life expectancy of about. 0.057 *years*, which is about 21 days.

Attn.: We must not interpret this result causally, since both life expectancy and per capita income depend on many other variables, and thus there is almost certainly an *'omitted variable bias'*!!!!

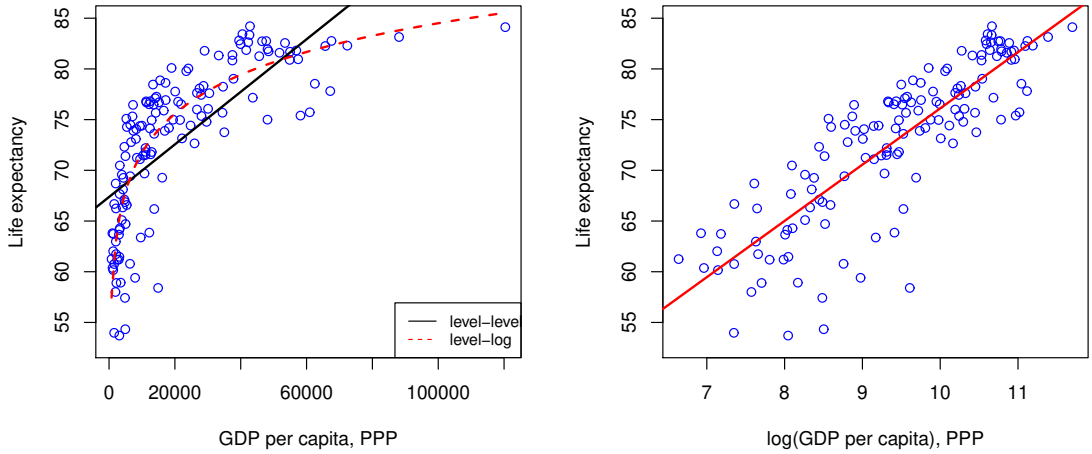


Figure 2.38: Life expectancy at birth, total (years) vs. GNI per capita, PPP, constant 2011 international $, 2018; Source: World Bank, WDI, `https://databank.worldbank.org/`.

2.8.6 When to logarithmise?

Logarithmic transformations are frequently used in econometrics, especially for variables that are measured in monetary terms or some other level, e.g. GDP, population or area. In particular, macroeconomic variables that can plausibly be expected to grow approximately exponentially in the long run are often logarithmised because, as we have seen, the logarithm of such variables increases linearly.

As a simple guide, you can consider whether you think of changes in a variable more in terms of *percentage* or changes by an *absolute amount*.

Variables that have a time dimension, on the other hand, are rarely logarithmised, e.g. age, work experience or years of education.

In addition, logarithmisation sometimes offers advantages when the distribution of y is skewed or the variances of e_i are not constant (heteroskedasticity), because often the distribution of a logarithmised variable better satisfies the assumptions of the regression model than the distribution of a non-logarithmised variable. Since logarithmisation 'compresses' large numerical values, logarithmically specified regressions are also often less susceptible to outliers (*'outliers'*).

Another reason is that the standard deviation of many economic variables is approximately proportional to the level of these variables, therefore the standard deviation of logarithmised variables is approximately constant. In other words, logarithmising economic variables often leads to a kind of 'stabilisation' of the standard errors of the coefficients.

A major advantage of logarithmic transformations is that the slope coefficients can be interpreted as elasticities, or semi-elasticities, and are therefore independent of the original unit of measurement. However, this should not tempt one to carelessly choose a logarithmic functional form.

Script 2.7: Life expectancy at birth, total (years) vs. log(GNI per capita), `R-Code`

```
# WDI, Grafik: Life Exp. vs. log(GDP)
rm(list = ls())
# install.packages(WDI)
# see e.g. https://cengel.github.io/gearup2016/worldbank.html
library(WDI)

# download data
wdi_dat <- WDI(country = "all",
               indicator = c("NY.GNP.PCAP.PP.KD", "SP.DYN.LE00.IN"),
               start = 2020, end = 2020, extra = TRUE, cache = NULL)

# remove country aggregates
wdi_dat <- subset(wdi_dat, region != "Aggregates")
# rename
names(wdi_dat)[names(wdi_dat) == "NY.GNP.PCAP.PP.KD"] <- "GDPpc"
names(wdi_dat)[names(wdi_dat) == "SP.DYN.LE00.IN"] <- "LifeExp"
eqlog <- lm(LifeExp ~ log(GDPpc), data = wdi_dat)

# Graph
x11(width=10,height=5) # only for Windows, for Mac's use quartz()
par(mfrow = c(1,2)) # two graphs, 1 row, 2 columns
plot(wdi_dat$LifeExp ~ wdi_dat$GDPpc, col = "blue",
     ylab = "Life␣expectancy", xlab = "GDP␣per␣capita,␣PPP")
abline(lm(wdi_dat$LifeExp ~ wdi_dat$GDPpc), lwd = 2, col = "black")
curve(expr = (coef(eqlog)[1] + coef(eqlog)[2]*log(x)),
      add = TRUE, col = "red", lwd = 2, lty = 2)
legend("bottomright", legend = c("level-level", "level-log"),
       col=c("black", "red"), lty=1:2, cex=0.8)

plot(wdi_dat$LifeExp ~ log(wdi_dat$GDPpc), col = "blue",
     ylab = "Life␣expectancy", xlab = "log(GDP␣per␣capita),␣PPP")
abline(lm(wdi_dat$LifeExp ~ log(wdi_dat$GDPpc)),
       lwd = 2, col = "red")
```

The decisive argument for the choice of functional form should be which model is consistent with the theory and better represents the data. In principle, there are also tests for the choice of functional form (e.g. test by MacKinnon, White & Davidson), but these are often not very powerful. Sometimes it is enough to look at a scatterplot or the histogram of the estimated residuals to see which functional form is more appropriate.

Percentages versus *percentage points*: Care should be taken when using *growth rates* and/or *shares* in regressions. For the interpretation of the coefficients of such regressions it is essential to distinguish between percentages and percentage points. For example, if the unemployment rate increases from 5% to 6%, this is an increase of one *percentage point*, but an increase of 20 percent ($= (6-5)/5 \times 100 = 20$) over the original level. Note that the difference in logarithms approximates a relative change, e.g. $\log(6) - \log(5) = 0.1823 \approx 0.2$ (note that $(\exp(0.1823) - 1) \times 100 = 20$).

Assume that in the regression $\widehat{\log(y)} = b_1 + b_2 A$, A is a proportion, i.e. is a number

Table 2.15: Relationship between life expectancy and per capita income in a cross-section of countries (in the level-level specification, GDPpc was measured in 1000 $. As can be seen in Figure 2.38, the level-level specification provides a very poor fit to the data

	Dependent variable:	
	LifeExp	
	(1)	(2)
Constant	19.303***	66.820***
	(2.936)	(0.724)
log(GDPpc)	5.683***	
	(0.314)	
GDPpc/1000		0.266***
		(0.026)
Observations	149	149
R^2	0.691	0.425
Note:	*p<0.1; **p<0.05; ***p<0.01	

between zero and one ($0 \leq A \leq 1$), what makes it a log-level model. Therefore, $100b_2$ approximates how much *percent* $\widehat{y}$ changes when A increases by *one unit*.

But what should we think of by 'one unit' of a proportion? If we multiply A by 100 we get percent, or an increase by one *percentage point*, because in the log-level model $\widehat{\log(y)} = b_1 + b_2 A$ is

$$b_2 \approx \frac{\frac{\Delta\widehat{y}}{\widehat{y}}}{\Delta A} = \frac{\frac{\Delta\widehat{y}}{\widehat{y}} \times 100}{\Delta A \times 100} = \frac{\text{percentage change of } \widehat{y}}{\text{increase of } A \text{ by one } \textit{percentage point}}$$

Therefore, b_2 approximates the percentage change in $\widehat{y}$ when A increases by *one percentage point* (as always in log-level models, we get a slightly more accurate value with $[\exp(b_2) - 1]$, and this correction should at least be made when $b_2 > 0.1$).

In a log-log model $\widehat{\log(y)} = b_1 + b_2 \log(A)$, b_2 is again interpreted as elasticity as usual, that is, by how many percent $\widehat{y}$ changes when the share A increases by *one percent*.

Observations with zeros: As has been pointed out several times, the logarithm of zero and negative values is not defined, so variables that contain (or can take on) negative values or the value zero must not be logarithmed!

A bit of a question of faith is what to do when a variable y contains only a few zeros. If for some reason you still want to logarithmise y, and it really is only a few zeros, which are also of no great importance in terms of the content interpretation, some authors recommend simply using $\log(1 + y)$ instead of $\log(y)$. The usual percentage interpretation is often at least approximately preserved, with the exception of changes near $y = 0$, where it is not even defined (cf. Wooldridge, 2005, 199). In such

cases, however, it is generally advisable to use more suitable estimation methods, e.g. Tobit or Poisson models.

Review:

1. **Linear model**: $\widehat{y} = b_1 + b_2 x_2 + \cdots + b_h x_h + \cdots + b_k x_k$

   Marginal effect:

   $$b_h = \left.\frac{\Delta \widehat{y}}{\Delta x_h}\right|_{\text{ceteris paribus}}$$

   by ceteris paribus we mean that *all other* x variables are assumed constant ($\Delta x_1 = \ldots = \Delta x_{h-1} = \Delta x_{h+1} = \ldots = \Delta x_k = 0$) *Interpretation:* An ceteris paribus increase of x by *unit* is accompanied by a change of y by b_2 *units*.

2. **Log-log model**: $\widehat{\log y} = b_1 + b_2 x_2 + \cdots + b_h \log x_h + \cdots + b_k x_k$.

   Marginal effect:

   $$b_h = \left.\frac{\Delta \widehat{\log y}}{\Delta \log x_h}\right|_{\text{c.p.}} \approx \frac{\frac{\Delta \widehat{y}}{\widehat{y}}}{\frac{\Delta x_h}{x_h}} = \frac{\frac{\Delta \widehat{y}}{\widehat{y}} \times 100}{\frac{\Delta x_h}{x_h} \times 100} \qquad \text{(Elasticity)}$$

   Interpretation: An ceteris paribus increase of x by *one per cent* is accompanied by a change of y by b_2 *per cent*, i.e. b_2 can be interpreted as elasticity.

3. **Log-level model**: $\widehat{\log y} = b_1 + b_2 x_2 + \cdots + b_h x_h + \cdots + b_k x_k$

   Marginal effect:

   $$b_h = \left.\frac{\Delta \widehat{\log y}}{\Delta x_h}\right|_{\text{c.p.}} \approx \frac{\frac{\Delta \widehat{y}}{\widehat{y}}}{\Delta x_h} \quad \text{oder} \quad 100 \times b_h \approx \frac{\frac{\Delta \widehat{y}}{\widehat{y}} \times 100}{\Delta x_h}$$

   Interpretation: An ceteris paribus increase of x by *one unit* (e.g. one euro) is accompanied by a change of y by *approximately* $100 \times b_2$ *percent*, or more precisely, to a change of $[\exp(b_2) - 1] \times 100$ *percent* (for $|b_2| > 0.1$).

4. **Level-log model** $\widehat{y} = b_1 + b_2 x_2 + \cdots + b_h \log x_h + \cdots + b_k x_k$

   Marginal effect:

   $$b_h = \left.\frac{\Delta \widehat{y}}{\Delta \log x_h}\right|_{\text{c.p.}} \approx \frac{\Delta \widehat{y}}{\frac{\Delta x_h}{x_h}} \quad \text{oder} \quad 0.01 \times b_h = \frac{\Delta \widehat{y}}{\frac{\Delta x_h}{x_h} \times 100}$$

   Interpretation: An increase of x by *one per cent* is accompanied by a change of y by $0.01 \times b_2$ *units* (e.g. euros).

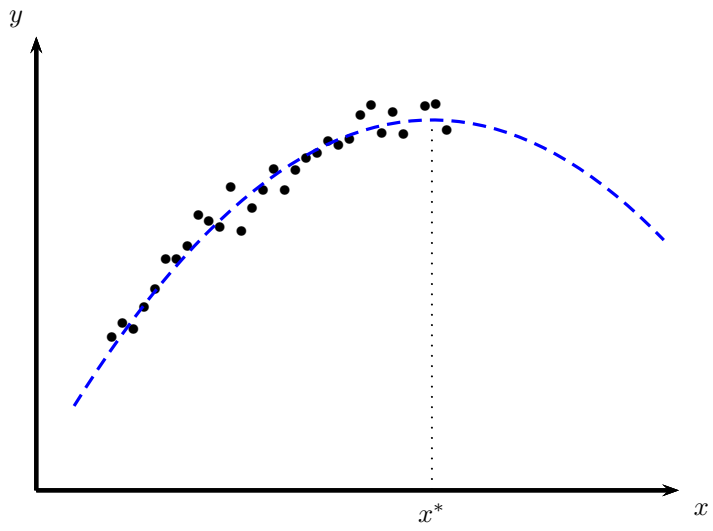


Figure 2.39: Quadratic models $\widehat{y} = b_1 + b_2 x + b_3 x^2$ assume a symmetric process, so they may have a very good fit in the sample but give very poor predictions. In this example, $b_2 > 0$ and $b_3 < 0$.

2.9 Quadratic models

Linear functional forms are practical and easy to interpret, but unfortunately reality is not always so simple, sometimes the marginal effects are not constant but depend on the level.

For example, we know from microeconomics that the short-run average cost function of a firm often has a U-shaped shape.

A very simple – albeit rather restrictive – method for modelling such non-linearities is to use polynomials. In what follows we will restrict ourselves to quadratic functional forms.

$$\widehat{y}_i = b_1 + b_2 x + b_3 x^2$$

. Note that this function is linear in the parameters b_1 and b_2, and therefore can be easily estimated with OLS, but is *nonlinear in the explanatory variable* x.

The *marginal effect* can be calculated as a derivative as usual

$$\frac{d\widehat{y}}{dx} = b_2 + 2b_3 x$$

Obviously, the marginal effect for this quadratic model is not constant, but depends on the level of x. This is also evident in Figure 2.39; there $b_2 > 0$ and $b_3 < 0$, therefore the slope is large at very small x, decreases with increasing x, and eventually becomes negative.

Caution: since $d\widehat{y}/dx = b_2 + 2b_3 x$, b_2 measures the marginal effect only at the point $x = 0$ (abscissa intercept), and this is rarely of interest. Most of the time it is more appropriate to specify the marginal effect in the mean of x or in another point that can be interpreted well (e.g. quartiles). If enough space is available, it can also be represented graphically.

Caution: Should the influence of x on $\widehat{y}$ be tested statistically, the *common* significance of b_2 *and* b_3 must be tested using F-statistics; more on this later.

Quadratic functions have a maximum or minimum, which can be easily calculated by setting the derivative $d\widehat{y}/dx = b_2 + 2b_3x$ equal to zero and solving for x.

$$x^* = \frac{-b_2}{2b_3}$$

Before interpreting this extreme value x^*, however, one should check what proportion of the observations lie to the left or to the right of the extreme value, compare Figure2.39.

Square log-level models: Suppose we want the marginal effect of x in the model.

$$\widehat{\log(y)} = b_1 + b_2x + b_3x^2$$

calculate.

The usual marginal effect of x is not constant in this case, but depends on the expression of x

$$\frac{d\,\widehat{\log(y)}}{d\,x} = b_2 + 2b_3x \approx \frac{\frac{\Delta\widehat{y}}{\widehat{y}}}{\Delta x}$$

If we want to indicate by how many *percent* $\widehat{y}$ changes when x increases by *one unit*, we could calculate this for the mean $\bar{x}$, for example

$$\%\Delta\widehat{y} := \frac{\frac{\Delta\widehat{y}}{\widehat{y}} \times 100}{\Delta x} \approx (\exp(b_2 + 2b_3\bar{x}) - 1) \times 100$$

However, even this does not hold exactly, if we make differences for discrete changes we get [25]

$$\begin{aligned}\Delta\widehat{\log(y)} := \log(\widehat{y} + \Delta\widehat{y}) - \widehat{\log(y)} &= b_1 + b_2(x + \Delta x) + b_3(x + \Delta x)^2 - \\ &\quad b_1 - b_2x - b_3x^2 \\ &= b_2\Delta x + b_3[2x\Delta x + (\Delta x)^2]\end{aligned}$$

and for $\Delta x = 1$

$$\log\left(1 + \frac{\Delta\widehat{y}}{\widehat{y}}\right) = b_2 + b_3(2x + 1) = b_2 + 2b_3(x + 0.5) \approx \frac{\Delta\widehat{y}}{\widehat{y}}$$

und

$$\%\Delta\widehat{y} := \frac{\Delta \log \widehat{y}}{\Delta x} \times 100 \approx \frac{\frac{\Delta\widehat{y}}{\widehat{y}} \times 100}{\Delta x} = [\exp(b_2 + 2b_3(x + 0.5)) - 1] \times 100$$

How important this correction is again depends on the concrete numerical values.

[25] Thanks to Klaus Nowotny for the hint!

Example: for Austrian EU-Silc data (2018) we obtain[26]

$$\begin{array}{rcccccc} \log(\text{wage}) = & 1.871 & + 0.037\,\text{educ} & + 0.028\,\text{exper} & - 0.000366\,\text{exper}^2 \\ & (0.035)^{***} & (0.002)^{***} & (0.002)^{***} & (0.00005)^{***} \end{array}$$

$$R^2 = 0.15, \qquad n = 4104$$

Accordingly, the mean hourly wage (wage) increases ceteris paribus by approx. 3.7% with each additional year of potential education (educ).

$$\frac{\Delta\widehat{\log(\text{wage})}}{\Delta\text{educ}} = 0.037$$

or more precisely

$$(\exp(0.037) - 1) * 100 = 3.77\%$$

On the other hand, the average effect of work experience (exper) is not constant

$$\frac{\Delta\widehat{\log(\text{wage})}}{\Delta\text{exper}} = 0.028 - 2 * 0.000366\,\text{exper}$$

it increases first, reaching a maximum after about 38 years, [27]

$$\text{exper}^{\max} = \frac{b_3}{-2b_4} = \frac{0.028}{-2(-0.000366)} = 38.25$$

and decreases thereafter.

The marginal effect of work experience for a beginner (i.e. for exper = 0) is ceteris paribus

$$\%\Delta\widehat{\log(\text{wage})} = [\exp(0.028 - 2 \times 0.000366(0 + 0.5)) - 1]100 = 2.8$$

i.e. in the first year of employment the mean hourly wage increases ceteris paribus by ca. 2.8%, ...

$$\%\Delta\widehat{\log(\text{wage})} = [\exp(0.028 - 2 \times 0.000366(20 + 0.5)) - 1]100 = 1.31$$

after 20 years (i.e. exper = 20), the fitted hourly wage increases by only approx. 1.31 percent with one more year of work experience, ceteris paribus.

[26] These types of wage equations are often called Mincer income equations in honour of Jacob Mincer (1922 – 2006, one of the founders of modern empirical labour economics). For an overview on estimating and measuring the impact of human capital on income, see e.g. the freely available contributions by Deming (2022) and Abraham and Mallatt (2022).

[27] We set the first derivative after exper equal to zero and solve for exper.

2.10 Interaction models

We can also use *products* of individual explanatory variables as explanatory variables, e.g. in the following model the product of x_2 and x_3

$$\widehat{y} = b_1 + b_2 x_2 + b_3 x_3 + b_4 (x_2 x_3)$$

In these models, x_2 and x_3 are called main terms or main effects and the product $x_2 x_3$ is called the interaction term.

If interaction terms are considered in a model, the main effects should in any case also be taken into account, as otherwise the risk of misspecification due to missing relevant variables is extremely high (the interaction term $x_2 x_3$ is practically always correlated with x_2 and x_3 (cf. Balli and S rensen, 2012)).

The marginal effect of x_2 depends on the level of x_3,

$$\frac{\partial \widehat{y}}{\partial x_2} = b_2 + b_4 x_3$$

i.e. the ceteris paribus effect of a change from x_2 to $\widehat{y}$ also depends on the absolute value of x_3 at the point in question.[28]

One can graphically represent the dependence of the marginal effect of x_2 on the value of x_3, for example, by plotting $\partial y / \partial x_2$ against x_3 in a graph.

Attention: The coefficient b_2 measures the marginal effect of x_2 only in the point $x_3 = 0$!

Analogously, for a change of x_3, if x_2 is kept constant,

$$\frac{\partial \widehat{y}}{\partial x_3} = b_3 + b_4 x_2$$

From the fact that the coefficient of the interaction term in the linear model is

$$\widehat{y} = b_1 + b_2 x_2 + b_3 x_3 + b_4 (x_2 x_3)$$

simply the second derivative is

$$b_4 = \frac{\frac{\partial \widehat{y}}{\partial x_2}}{\partial x_3} = \frac{\frac{\partial \widehat{y}}{\partial x_3}}{\partial x_2} = \frac{\partial^2 \widehat{y}}{\partial x_2 \partial x_3}$$

another important observation follows:

> The coefficient of the interaction term b_4 indicates how the marginal effect of x_2 changes when x_3 increases by one (infinitesimal) unit.

Note that the functional form enforces a symmetry of the marginal effects (this follows from the symmetry of the second derivatives, cf. Young's theorem), i.e. b_4 can also be interpreted as the change in the marginal effect of x_3 when x_2 increases by one unit.

[28] If it is to be tested later whether x_2 has an effect on y, the simple t test for the coefficient of x_2 must not be used, but e.g. with a F test the simultaneous null hypothesis $\beta_2 = 0$ and $\beta_4 = 0$ must be tested.

Example: The effectiveness of development aid has always been highly controversial; simple regressions of indicators of development aid received on the growth rate of recipient countries regularly showed no relationship or produced contradictory results.

In a influential paper on development aid, Burnside and Dollar (2000) *"We find that aid has a positive impact on growth in developing countries with good fiscal, monetary, and trade policies but has little effect in the presence of poor policies."* You justified your statement with a somewhat more complex OLS regression, which looks (strongly) simplified as follows

$$G = b_1 + b_2 A + b_3 P + b_4 (A \times P) + \cdots + e$$

where G (*growth*) is the growth rate of recipient countries, A (*aid*) is an indicator of development aid received, and P (*policy*) is an indicator of 'good' economic policy based on macroeconomic variables such as inflation rate, budget deficit, etc.

The argument of Burnside and Dollar (2000) that development aid only works in countries with 'good' economic policies was based on the coefficient of the interaction variable, which in their model was positive and statistically significantly different from zero. This means that development aid has a positive impact on the growth rate in countries with 'good' economic policies.

$$\frac{\partial \widehat{G}}{\partial A} = b_2 + b_4 P$$

This paper was extremely influential politically, as it fitted well into the worldview of many concerned actors, but it was soon shown that this result was very much dependent on country selection and time period. For a larger number of countries and later years, the result could often not be reproduced (Jensen and Paldam, 2006).

2.10.1 Alternative parameterisation of interaction models*

A simple 'reparameterisation' can be used to estimate an alternative interaction model whose coefficients directly measure the marginal effect in the mean of the variable in question.

Let us start with the simple interaction model

$$\widehat{y} = b_1 + b_2 x_2 + b_3 x_3 + b_4 x_2 x_3 \tag{2.18}$$

For example, if we are interested in the marginal effect of x_2 measured in the *mean* of x_3 we can simply calculate it as follows

$$\frac{\partial \widehat{y}}{\partial x_2} = b_2 + b_4 \bar{x}_3 \tag{2.19}$$

We will now show that we can estimate this marginal effect *in the mean of the other variables* even more simply by a simple variable transformation.

Suppose that instead of equation (2.18) we estimate a reparametrised model

$$y = a_1 + a_2 x_2 + a_3 x_3 + a_4 (x_2 - \bar{x}_2)(x_3 - \bar{x}_3) \tag{2.20}$$

where $\bar{x}_2$ and $\bar{x}_3$ denote the mean values as usual, then a_2 measures the marginal effect of x_2 *in the mean of* x_3, i.e. gives exactly the same value we get from equation (2.19)!

Why this is so can be easily shown. If we multiply out the interaction term we get

$$\begin{aligned} y &= a_1 + a_2 x_2 + a_3 x_3 + a_4[x_2 x_3 - x_2 \bar{x}_3 - x_3 \bar{x}_2 + \bar{x}_2 \bar{x}_3] \\ &= a_1 + a_4 \bar{x}_2 \bar{x}_3 + (a_2 - a_4 \bar{x}_3) x_2 + (a_3 - a_4 \bar{x}_2) x_3 + a_4 x_2 x_3 \end{aligned}$$

From a comparison with equation (2.18) we immediately see that $(a_2 - a_4 \bar{x}_3) = b_2$, $(a_3 - a_4 \bar{x}_2) = b_3$ and $a_4 = b_4$.

What have we gained from this? From $(a_2 - a_4 \bar{x}_3) = b_2$ and $a_4 = b_4$ follows

$$a_2 = b_2 + b_4 \bar{x}_3$$

but this is exactly the marginal effect of x_2 measured in the *mean of* x_3 we got from equation (2.19)!

Thus, if instead of equation (2.18) we estimate the reparametrised equation (2.20), the coefficient a_2 directly measures the marginal effect of x_2 *at the mean of* x_3, with the associated correct standard error of the marginal effect at the mean. The associated t statistic can thus be used directly to test the extent to which the marginal effect *at the mean of* x_3 is different from zero. The same is true for a_3.

This is obviously more interesting than estimating b_2 from equation (2.18), because this coefficient measures the marginal effect only at the point $x_3 = 0$, which is unlikely to ever be relevant.

However, this does not change the non-linearity of the relationship, the marginal effect of x_2 is different for each value of x_3!

In summary, some final recommendations for dealing with interaction effects of Brambor et al. (cf. 2006):

- Never interpret the coefficients of interaction variables as unconditional marginal effects! If you are interested in marginal effects, calculate them for values of interest of the other variables, or plot them.
- Calculate the marginal effects for relevant values of the variables, e.g. mean or median. A simple way to do this is to use the alternative parameterisation explained above.
- For significance test: test the *joint* significance of the relevant coefficients! (e.g. using an F-test, comes later in the chapter on hypothesis testing in the multiple regression model).

For a recent and more detailed discussion of multiplicative interaction effects with practical advice, see Hainmueller et al. (2019).

Example: Table 2.16 shows two estimates for wage equations with interaction effects.

Table 2.16: Wage equation for Austria with interaction effects (data: EU-Silc 2018)

Dependent Var.: log(wage)	(1)	(2)
Constant	2.31315***	1.88338***
	(0.05849)	(0.03429)
educ	0.00851**	0.04287***
	(0.00342)	(0.00176)
exper	0.00024	0.01909***
	(0.00379)	(0.00252)
exper^2	−0.00015***	−0.00015***
	(0.00005)	(0.00005)
(educ × exper)	0.00151***	
	(0.00016)	
$(\text{educ} - \overline{\text{educ}}) \times (\text{exper} - \overline{\text{exper}})$		0.00151***
		(0.00016)
Observations	4,104	4,104
R^2	0.16744	0.16744
Adjusted R^2	0.16663	0.16663
F Statistic (df = 4; 4099)	206.09320***	206.09320***
Note:		*p<0.1; **p<0.05; ***p<0.01

Column (1) of Table2.16 shows the direct interaction effect as in equation (2.18), column (2) the estimate for the reparametrised model as in equation (2.20).

All coefficients have the expected sign and are highly significantly different from zero.

Column (1) of Table2.16 shows the estimate for the equation

$$\widehat{\log(\text{wage})} = b_1 + b_2\,\text{educ} + b_3\,\text{exper} + b_4\,\text{exper}^2 + b_5\,\text{educ} \times \text{exper}$$

The marginal effect of educ is

$$\frac{\partial \widehat{\log(\text{wage})}}{\partial \text{educ}} = b_2 + b_5 \text{exper}$$

This function is shown in the left panel of Figure 2.40. If we take the estimated coefficients from Table2.16 and the mean for work experience $\overline{\text{exper}} = 22.79$ inserting we get

$$\frac{\partial \widehat{\log(\text{wage})}}{\partial \text{educ}} = 0.00851 + 0.00151 \times 22.79 = 0.04287$$

i.e. exactly the value of the coefficient of educ in column (2). This is not surprising, of course, as we proved this in the previous section on reparametrisation in general. So for someone with 22.79 years of work experience, we expect that an *additional year of education* will yield approximately a 4.287% higher hourly wage.

The good news is that once education increases in value with work experience, the marginal effect of education increases with experience.

Also by a simple partial derivation we can calculate the marginal effect for work experience 'exper'.

$$\frac{\partial \widehat{\log(\text{wage})}}{\partial \text{exper}} = b_3 + 2b_4\,\text{exper} + b_5 \text{educ}$$

In this case, the marginal effect of work experience on hourly wages depends on the level of years of education *and* work experience!

Obviously, this marginal effect differs according to education and experience.

But we can calculate the marginal effects for different values of for education and experience using the coefficients from column (1) from Table2.16, e.g. the marginal effect in the mean values ($\overline{\text{educ}} = 12.51, \overline{\text{exper}} = 22.79$)

$$\frac{\partial \widehat{\log(\text{wage})}}{\partial \text{exper}} = 0.0002371 + 2 \times (-0.000154) \times 22.79 + 0.00151 \times 12.51 \approx 0.0121$$

That is, someone with average education and experience can expect a wage increase of 1.21% for an additional year of experience (of course, we are talking about a linear approximation ...).

For someone with average education and exper = 0, we get a marginal effect of experience of 0.0191, the value from column (2) of Table 2.16.

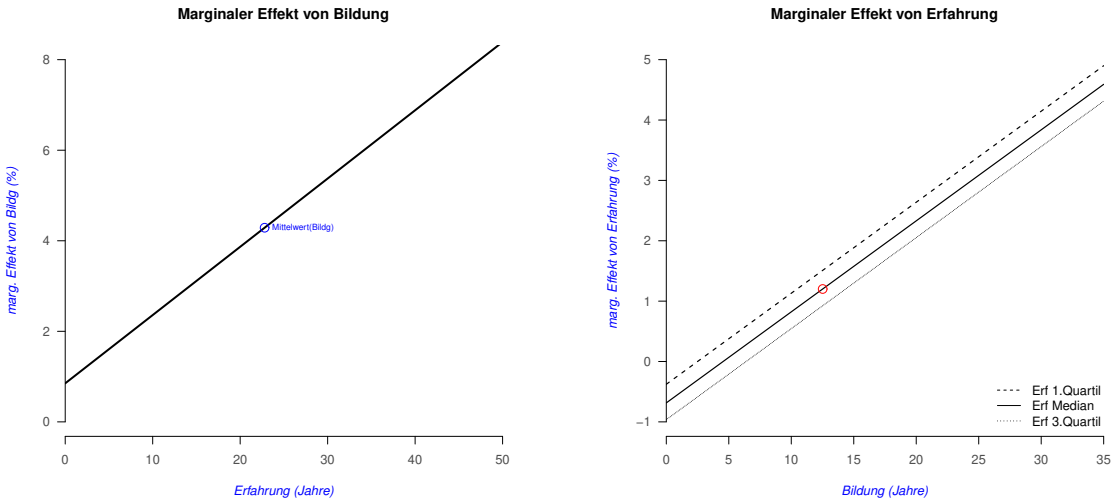


Figure 2.40: Marginal effects with interaction effects for wage equation.

Thus we can calculate a marginal effect for each educ – exper combination, but this is not very convenient. One can either restrict oneself to combinations of interest, or show the effects graphically, see Figure2.40, but this option is rarely used for such simple models.

The right panel of Figure 2.40 shows the dependence of the marginal effect of experience (exper) on hourly wage for different education durations educ and for three levels of exper (by quartiles) is in the right panel of Figure 2.40. Obviously, the marginal effect of work experience increases ceteris paribus with education duration, and is higher ceteris paribus the lower the work experience.

If we substitute the median of 20 years for exper instead of zero, we get the value 0.01565, which means that for someone with 13.8 years of education, we expect the hourly wage to increase by about 1.565% when the work experience increases from 20 years to 21 years. This point is plotted on the graph.

Caution with non-linear functional forms:

- Polynomial models (e.g. quadratic or cubic models) can sometimes be used to achieve a very good fit in the sample, but they are usually quite useless for forecasting, as the functional form *'out of sample'* often forces extreme runs.
- The coefficient of determination R^2 may only be used to compare models in which the dependent variable y has not been transformed *and* if both models have the same number of explanatory variables.

  This is easy to see: the coefficient of determination is defined as the proportion of the dispersion explained by x in the total dispersion of y. If y is transformed (e.g. $\log(y)$ is used instead of y), the dispersion will of course also change, and with it the R^2

$$R_y^2 = 1 - \frac{\sum_i e_i^2}{\sum_i (y_i - \overline{y})^2} \quad \Leftrightarrow \quad R_{\log(y)}^2 = 1 - \frac{\sum_i e_i^2}{\sum_i [\log(y_i) - \overline{\log(y)}]^2}$$

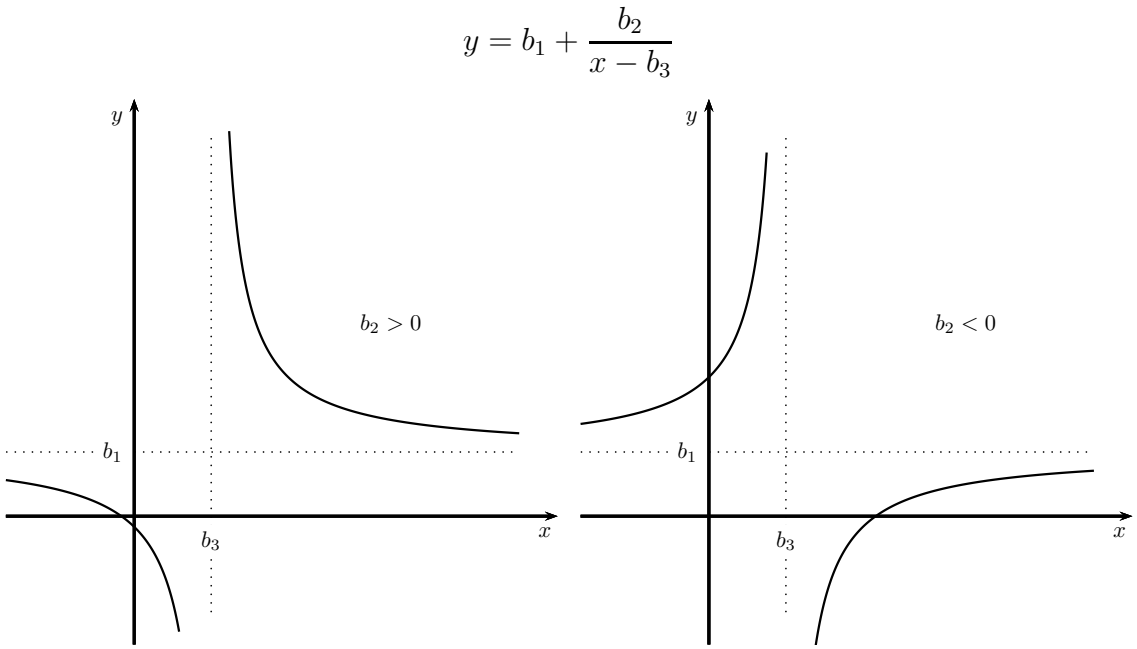


Figure 2.41: Reciprocal transformations

This also applies to the corrected coefficient of determination $\bar{R}^2$, which merely allows a comparison of models with a *different* number of explanatory variables.

Generally, neither R^2 nor $\bar{R}^2$ should be overinterpreted for assessing the quality of an estimate, as they only describe the adjustment in the sample. In most cases it is much more important whether the estimated coefficients meet theoretical expectations and are significantly different from zero.

2.11 Reciprocal transformations

Another functional form that can be used, for example, for the estimation of Phillips curves are reciprocal transformations, one simply uses the reciprocal of the variable, see Figure 2.41.

$$y = b_1 + b_2 \frac{1}{x}$$

Overview:

Model	Equation	Slope ($= \frac{dy}{dx}$)	Elasticity ($= \frac{dy}{dx}\,\frac{x}{y}$)
Linear	$y = \alpha + \beta x$	β	$\beta(x/y)$
Log-log	$\log y = \alpha + \beta \log x$	$\beta(y/x)$	β
Log-level	$\log y = \alpha + \beta x$	$\beta(y)$	$\beta(x)$
Level-log	$y = \alpha + \beta \log x$	$\beta(1/x)$	$\beta(1/y)$
Reziprok	$y = \alpha + \beta(1/x)$	$-\beta(1/x^2)$	$-\beta(1/xy)$

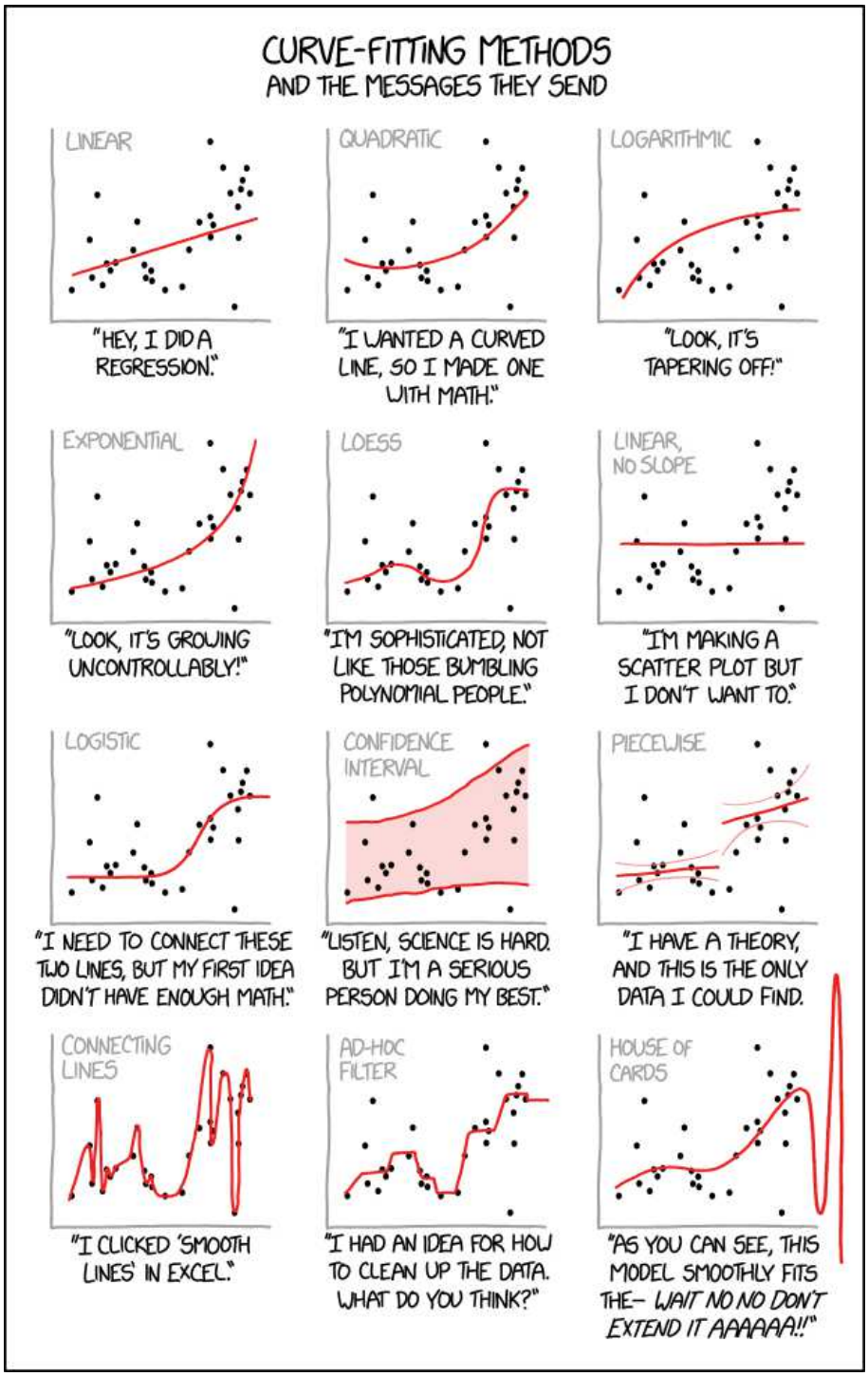


Figure 2.42: *Quelle:* XKCD, `https://xkcd.com/2048/`

2.12 Miscellaneous

2.12.1 Mean value transformations

There is a special data transformation that is often used in econometrics and will often prove useful later, namely the mean transformation.

The mean transformation simply consists of subtracting from each individual observation x_i of a data series the mean value of the same data series $\bar{x}$.

The resulting data series simply consists of deviations from the mean, hence the name mean transformation. In the following, we will mark an observation (or data series) transformed in this way with two dots above the variable name in question, e.g.

$$\ddot{x}_i := x_i - \bar{x}$$

Figure 2.43 shows a graphical interpretation of this mean transformation. This transformation "subtraction of the mean" measures the coordinates of the variable thus transformed in relation to a new coordinate system whose new zero point lies in the mean of the original variable $(\bar{x}, \bar{y})$. So, in a sense, the subtraction of the mean causes a shift in the coordinate system so that the new zero point is shifted to the mean of the data.

We will encounter such mean-transformed data repeatedly, and have already encountered it; for example, equation (2.8) for b_2 is formed from the mean-transformed variables x and y, i.e.

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

Note also that the mean of a mean-transformed variable is always zero because.

$$\bar{\ddot{y}} := \frac{1}{n}\sum_i \ddot{y}_i = \frac{1}{n}\sum_i (y_i - \bar{y}) = \frac{1}{n}\sum_i y_i - \frac{1}{n}\sum_i \bar{y} = \bar{y} - \frac{1}{n} n\bar{y} = \bar{y} - \bar{y} = 0$$

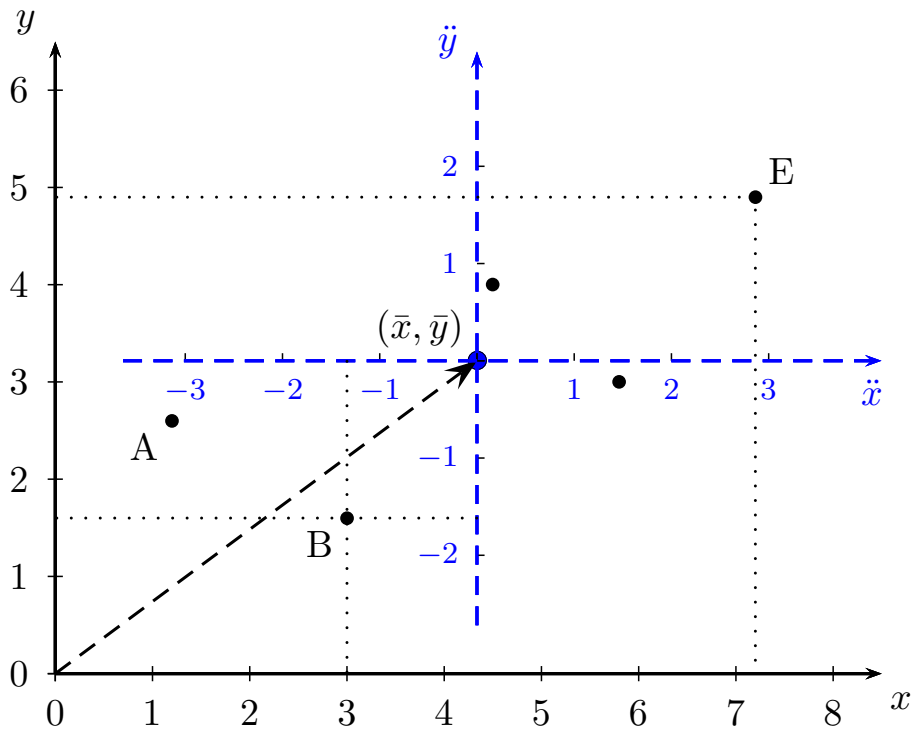
It also follows, for example, that

$$b_2 = \frac{\mathrm{cov}(y,x)}{\mathrm{var}(x)} = \frac{\mathrm{cov}(\ddot{y},\ddot{x})}{\mathrm{var}(\ddot{x})}$$

Therefore, it does not matter for the calculation of the slope coefficient b_2 whether one uses the original data series or mean-transformed data series, the OLS method gives the same result for the slope coefficient in both cases.

However, the intercept b_1 can no longer be calculated directly from the mean-transformed data series, because this is dropped in the mean transformation.

$$\begin{array}{rcll} y_i &=& b_1 + b_2 x_i + e_i & \\ \bar{y} &=& b_1 + b_2 \bar{x} + \bar{e} & /- \\ \hline y_i - \bar{y} &=& b_1 - b_1 + b_2(x_i - \bar{x}) + e_i - \bar{e} & \\ \ddot{y}_i &=& b_2 \ddot{x}_i + \ddot{e}_i & \end{array}$$



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
mean:	3.22	4.34	0.00	0.00

Figure 2.43: Mean transformation (subtraction of the mean value for each observation): The coordinates of point B in the original coordinate system are $(3.0, 1.6)$; if the mean is subtracted, the coordinates are obtained in relation to a new coordinate system whose origin lies in the mean value of the observations $(\bar{x}, \bar{y})$, for point B thus $(-1.34, -1.62)$.

This should not be surprising because, as we saw earlier, the mean transformation graphically corresponds to a shift of the zero point of the coordinate system to the mean of the variables, and there to the intercept by definition zero.

But of course the intercept can be easily recalculated from the non-transformed data with $b_1 = \bar{y} - b_2\bar{x}$.

Exercise: Using the mean-transformed data, we can write the relationship $y_i = b_1 + b_2 x_i + e_i$ shorter $\ddot{y}_i = b_2\ddot{x}_i + e_i$, because the OLS estimator b_2 is actually the same in both cases.

We can derive the OLS estimator for the mean-transformed model as an exercise. The residuals are $e_i = \ddot{y}_i - b_2\ddot{x}_i$, therefore the minimisation problem is

$$\min_{b_2} \sum_i e_i^2 = \min_{b_2} \sum_i (\ddot{y}_i - b_2\ddot{x}_i)^2$$

The first order condition is

$$\frac{d\sum_i e_i^2}{d\, b_2} = -2\sum_i (\ddot{y}_i - b_2\ddot{x}_i)(-\ddot{x}_i) = 0$$

It follows that $\sum_i \ddot{y}_i\ddot{x}_i = b_2(\sum_i \ddot{x}_i)^2$ or

$$b_2 = \frac{\sum_i \ddot{x}_i\ddot{y}_i}{\sum_i \ddot{x}_i^2} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{\mathrm{cov}(x, y)}{\mathrm{var}(x)}$$

The intercept can again be calculated as usual with $b_1 = \bar{y} - b_2\bar{x}$. As we have already mentioned, this result can be interpreted as a special case of the Frisch-Waugh-Lovell theorem.

2.12.2 Scaling

Sometimes it is necessary to change the scaling of variables, for example when prices are to be expressed in dollars instead of euros, time in months instead of years, or distances in kilometres instead of metres.

In such cases, dependent and/or explanatory variables are multiplied by a constant number (e.g. the exchange rate). What consequences does this have for estimates? Do we have to re-estimate with the linearly transformed data, or can we simply calculate the results from the estimates?

It turns out that in such cases it is sufficient to linearly transform the coefficients and standard errors of the estimates as well, the estimates need *not* to be redone.

Recall that the coefficients in the simple linear model measure the marginal effect, i.e. by how many units $\widehat{y}$ changes when x increases by one unit, ceteris paribus. If we now multiply an explanatory x variable by a constant c the new coefficient measures by how many units $\widehat{y}$ changes when the explanatory variable increases by 'one *new* unit' ($= cx$); so all we need to do is divide the original coefficient by c

$$\widehat{y} = b_1 + (\underbrace{\frac{1}{c} b_2}_{b_2^*})(cx) + e$$

where b_2^* denotes the coefficient of the scaled equation.

Let us return to our old example with the used cars, the first column of Table2.17 shows the original estimate, the second column shows what happens if we give the age in months instead of years, i.e. we multiply the age in years by 12. This is a simple linear transformation and we see that the intercept, the coefficient of km as well as the coefficient of determination R^2 do not change as a result.

Table 2.17: Scaling:

	Dependent variable:		
	Price in €		Price in 1000 €
	(1)	(2)	(3)
Constant	$22,649.880^{***}$	$22,649.880^{***}$	22.650^{***}
	(411.870)	(411.870)	(0.412)
Age in years	$-1,896.264^{***}$		
	(235.215)		
Age in months		-158.022^{***}	-0.158^{***}
		(19.601)	(0.020)
km	-0.031^{***}	-0.031^{***}	-0.00003^{***}
	(0.008)	(0.008)	(0.00001)
Observations	40	40	40
R^2	0.907	0.907	0.907
Note:			*p<0.1; **p<0.05; ***p<0.01

The coefficient of the second equation indicates that the fitted price decreases by 158,022 euros ceteris paribus when the car gets older by *one month*.

As asserted earlier, we can also simply calculate the coefficient and the standard error (the number in parentheses below the coefficient) from the original equation: for the coefficient $-1\,896,264/12 = -158,022$ and for the standard error $235,215/12 = 19,601$.

If the scaling of the *dependent variable* is changed, i.e. if y is multiplied by a constant d, the right-hand side of the equation must also be multiplied by d, i.e. for $y^* := dy$

$$y^* = dy = \underbrace{db_1}_{b_1^*} + \underbrace{db_2}_{b_2^*} x + \underbrace{de}_{e^*}$$

The third column of Table2.17 shows the estimate when the price is measured in units of one thousand *euros* instead of euros (i.e. $d = 1/1000 = 0.001$). If age increases by one *month*, the price decreases ceteris paribus by 0.158 *thousand* euros (= 158.022 euros).

Somewhat more generally, we can show this for the bivariate model by defining, for any two constants $c, d > 0$, a scaled model with

$$y^* := dy \quad \text{and} \quad x^* := cx$$

and compare the estimators of the scaled model $y^* = b_1^* + b_2^* x^* + e^*$ with the original model $y = b_1 + b_2 x + e$.
For the OLS estimators of the coefficients are

$$b_2^* = \frac{\sum \ddot{x}_i^* \ddot{y}_i^*}{\sum \ddot{x}_i^{*2}} = \frac{\sum (c\,\ddot{x}_i)(d\,\ddot{y}_i)}{\sum (c\,\ddot{x}_i)^2} = \frac{dc \sum \ddot{x}_i \ddot{y}_i}{c^2 \sum \ddot{x}_i^2} = \frac{d \sum \ddot{x}_i \ddot{y}_i}{c \sum \ddot{x}_i^2} = \frac{d}{c}\, b_2$$

$$b_1^* = \bar{y}^* - b_2^* \bar{x}^* = d\bar{y} - b_2^* c\bar{x} = d\bar{y} - \left(\frac{d}{c}\, b_2\right) c\bar{x} = d b_1$$

mit $\ddot{y}_i := y_i - \bar{y}$ und $\ddot{x}_i := x_i - \bar{x}$.
It is just as easy to see that the coefficient of determination is not affected by scaling

$$R^{*2} = 1 - \frac{\sum e^{*2}}{\sum \ddot{y}^{*2}} = 1 - \frac{\sum (de)^2}{\sum (d\ddot{y})^2} = 1 - \frac{d^2 \sum e^2}{d^2 \sum \ddot{y}^2} = R^2$$

We will discuss and calculate the standard errors ($\widehat{\mathrm{se}}(b_h)$ with $h = 1, 2$) in a later chapter, but it is worth noting here that the standard errors can be adjusted just as easily when scaling x or y (we will derive the following formulas for the standard errors later)

$$s^{*2} = \frac{\sum e_i^{*2}}{n-2} = \frac{\sum (de_i)^2}{n-2} = d^2\, s^2$$

$$\widehat{\mathrm{se}}(b_2^*) = \sqrt{\frac{s^{*2}}{\sum \ddot{x}_i^{*2}}} = \sqrt{\frac{d^2 s^2}{c^2 \sum \ddot{x}_i^2}} = \frac{d}{c}\, \widehat{\mathrm{se}}(b_2)$$

$$\widehat{\mathrm{se}}(b_1^*) = \sqrt{\frac{s^{*2} \sum x_i^{*2}}{n \sum \ddot{x}_i^{*2}}} = \sqrt{\frac{d^2 s^2\, c^2 \sum x_i^2}{n\, c^2 \sum \ddot{x}_i^2}} = d\, \widehat{\mathrm{se}}(b_1)$$

It was thus shown that for $y_i^* := d y_i$ and $x_i^* := c x_i$ holds

$$\begin{aligned}
b_1^* &= d b_1 \\
b_2^* &= \frac{d}{c}\, b_2 \\
s^{2*} &= d^2\, s^2 \\
\widehat{\mathrm{se}}(b_1^*) &= d\, \widehat{\mathrm{se}}(b_1) \\
\widehat{\mathrm{se}}(b_2^*) &= \left(\frac{d}{c}\right) \widehat{\mathrm{se}}(b_2) \\
R^{2*} &= R^2
\end{aligned}$$

This also applies more generally to the multiple regression model.

Exercise: For the bivariate regression model, show that adding a constant to the dependent and/or explanatory variable only affects the intercept, but has no effect on the slope coefficient.

Is this also true for $y_i^* := d_1 + d_2 y_i$ and $x_i^* := c_1 + c_2 x_i$?

2.12.3 Standardised (beta) coefficients

As we have just seen, the value of the regression coefficients, as well as their standard errors, depends on the units of measurement in which the variables were measured (but not the R^2).

In some applications the variables do not have natural dimensions, e.g. in psychological tests the units are often adopted arbitrarily.

Therefore, in such cases, the variables are sometimes z-transformed (standardised) before the regression is estimated; that is, the mean is subtracted from all expressions and the resulting values are divided by the standard deviation of the variable.

Through this standardisation, the new units are the standard deviations of the variable. This standardisation is of course only a scaling, so this does not change the correlations, but since all variables are now measured on the same scale, the coefficients of the individual variables can be meaningfully compared with each other.

Such coefficients of a regression with z-transformed variables are called 'standardised coefficients' or 'beta coefficients'.

If we again label the deviations from the sample mean with two points above the variable we obtain the z-transformed variables by dividing by their standard deviation

$$\ddot{y}_i^z := \frac{y_i - \bar{y}}{s_y} := \frac{\ddot{y}_i}{s_y}, \quad \ddot{x}_{i1}^z := \frac{\ddot{x}_{i1}}{s_{x_1}}, \quad \ddot{x}_{i2}^z := \frac{\ddot{x}_{i2}}{s_{x_2}}, \quad \ldots, \quad \ddot{x}_{ik}^z := \frac{\ddot{x}_{k,i}}{s_{x_k}}$$

For the original model

$$y_i = b_1 + b_2 x_{i2} + b_3 x_{i3} + \ldots + b_k x_{ik} + e_i$$

we get after the z-transformation

$$\ddot{y}_i^z = b_2^z \ddot{x}_{i2}^z + b_3^z \ddot{x}_{i3}^z + \ldots + b_k^z \ddot{x}_{ik}^z + \tilde{e}_i$$

Note that subtracting the mean values eliminates the intercept.

The independence of units of measurement now allows a direct comparison of the coefficients b_h^z $(h = 1, 2, \ldots, k)$ with each other, i.e. if, for example, the variable x_h increases by *one standard deviation* ceteris paribus, we expect the dependent variable y to change by b_h^z *standard deviations*.

Because most economic variables are measured in readily interpretable units and because these 'beta coefficients' b_h^z cannot isolate the influence of individual explanatory variables on the dependent variable y any better than the usual coefficients b_h, they are rather rarely used in econometrics.

2.12.4 Reverse Regressions

We have so far minimised the sum of squared residuals of the equation $y_i = b_1 + b_2 x_i + e_i$, i.e. the square of the vertical distances between y_i and $\widehat{y}_i$, because we want to 'explain' y with the help of the x variable. In some cases, however, the direction

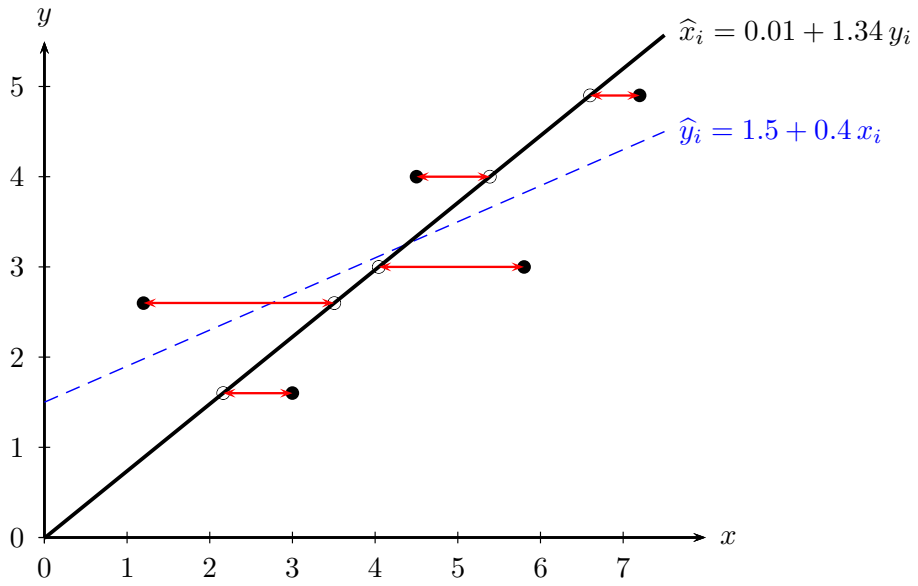


Figure 2.44: *'Reverse Regression'*: the regression from x to y ($\hat{x}_i = b_1^* + b_2{}^* y_i$), and dashed the normal regression $\widehat{y}_i = b_1 + b_2 x_i$.

of action is not clear, e.g. we can argue both ways in the case of the relationship between height x and weight y.

Ad hoc, many would expect that it does not matter whether we regress y on x or x on y, i.e.

$$y_i = b_1 + b_2 x_i + e_i \qquad \longleftrightarrow \qquad x_i = b_1^* + b_2^* y_i + e_i^*$$

because $y = b_1 + b_2 x + e$ can of course be rewritten as

$$x = -\frac{b_1}{b_2} + \frac{1}{b_2}\, y - \frac{1}{b_2}\, e$$

One might mistakenly assume that $b_1^* = -b_1/b_2$ and $b_2^* = 1/b_2$, but this is not so! The transformations are of course correct, but these are *not* the OLS estimators.

The OLS estimators of the reverse regression are

$$b_2^* = \frac{\mathrm{cov}(x,y)}{\mathrm{var}(y)}, \qquad b_1^* = \bar{x} - b_2^* \bar{y}$$

Figure 2.44 shows that in the case of the twisted regression the sums of squares of the horizontal distances are minimised. For comparison purposes, the direct regression $\widehat{y}_i = b_1 + b_2 x_i$ is also shown dashed.

2.12.5 Historical

The actual origins of the OLS method are still not fully understood. The only certainty is that it was first developed for astronomical applications, namely to calculate the most probable result for a new measurement from a series of inaccurate measurements, and that it was first used in 1805 by the French mathematician Adrien-Marie Legendre (1752-1833) in the appendix of a work on the calculation of

cometary orbits"Nouvelles méthodes pour la détermination des orbites des comètes." Paris 1805, Appendix: "Sur la Méthode des moindres quarrés", pp. 72-80 . was published. Legendre looked for a method of solving a system of equations with more equations than unknowns and showed that the 'method of least squares' (*"Méthode des moindres quarrés"*) leads to a system of equations that can be solved by *'ordinary'* methods, hence the name OLS (*'Ordinary Least Squares'*).

However, it is considered very likely that Carl Friedrich Gauss (1777-1855) developed the basics of the OLS method as early as 1795 at the age of 18. Presumably, the application of this method also contributed significantly to Gauss's early fame, because in 1801 it allowed him to calculate fairly accurately from a series of error-prone measurements the location where the recently discovered dwarf planet Ceres would re-emerge from behind the Sun. When Gauss finally published the method in 1809, he claimed the discovery of the OLS method as his own, which led to a dispute over authorship between Gauss and Legendre, who was 25 years older (cf. Singh, 2010).

The term *regression* is considerably younger and goes back to Francis Galton (1822 – 1911), a cousin of Charles Darwin. Galton, like many of his contemporaries – and especially many of the early pioneers of statistics – was concerned that the proliferation of negatively-valued hereditary traits would cause great long-term problems for Britain, and so became a founder of *eugenics*, which sought ways to increase the proportion of positively-valued hereditary traits. Galton found that in a regression of the height of children on the height of their parents, the regression coefficient was consistently less than one, so that parents who were taller than average tended to have smaller children, and parents who were shorter than average tended to have taller children. Galton (1886) called this "*Regression towards Mediocrity in Hereditary Status"*.[29] The statistical technique underlying the analysis subsequently became known as 'regression'.

Against the backdrop of the Boer War (1899 – 1902), which revealed a shortage of able recruits, Galton's findings were noted with some concern by the elites of the time. It led to fears of degeneration and long-term decline of imperial greatness.

However, it turned out that Galton's worries were unfounded; a regression coefficient smaller than one is perfectly compatible with a stable distribution of heights over time.[30]. That is why this phenomenon entered the literature as *"Galton's Fallacy"*.

[29]see `http://galton.org/essays/1880-1889/galton-1886-jaigi-regression-stature.pdf`

[30]Imagine three people with heights of 160, 180 and 200cm. Assume that each of these persons would again have three children, one 10% shorter, one the same height, and one 10% taller. Then already in the second generation the smallest child of the 160cm father would be only 144cm tall, the tallest child of the 200cm father would already be 220cm tall. Over a few generations, the smallest persons would be the size of ants, and the tallest persons would be true monsters! Such a distribution would obviously not be stable over time

Bibliography

Abraham, K. G. and Mallatt, J. (2022), 'Measuring human capital', *Journal of Economic Perspectives* **36**(3), 103–30.
URL: *https://www.aeaweb.org/articles?id=10.1257/jep.36.3.103*

Angrist, J. D. and Pischke, J.-S. (2008), *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press.

Balli, H. and S rensen, B. (2012), 'Interaction effects in econometrics', *Empirical Economics* pp. 1–21.
URL: *http://dx.doi.org/10.1007/s00181-012-0604-2*

Brambor, T., Clark, W. R. and Golder, M. (2006), 'Understanding interaction models: Improving empirical analyses', *Political Analysis* **14**(1), pp. 63–82.
URL: *http://www.jstor.org/stable/25791835*

Burnside, C. and Dollar, D. (2000), 'Aid, policies, and growth', *American Economic Review* **90**(4), 847–868.

Card, D. and Krueger, A. B. (1994), 'Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania', *The American Economic Review* **84**(4), 772–793.

Cobb, C. W. and Douglas, P. H. (1928), 'A theory of production', *The American Economic Review* **18**(1), pp. 139–165.
URL: *http://www.jstor.org/stable/1811556*

Deming, D. J. (2022), 'Four facts about human capital', *Journal of Economic Perspectives* **36**(3), 75–102.
URL: *https://www.aeaweb.org/articles?id=10.1257/jep.36.3.75*

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*

Galton, F. (1886), 'Regression towards mediocrity in hereditary stature.', *The Journal of the Anthropological Institute of Great Britain and Ireland* **15**, 246–263.
URL: *http://www.jstor.org/stable/2841583*

Hainmueller, J., Mummolo, J. and Xu, Y. (2019), 'How much should we trust estimates from multiplicative interaction models? simple tools to improve empirical practice', *Political Analysis* p. 1 30.

Jensen, P. and Paldam, M. (2006), 'Can the two new aid-growth models be replicated?', *Public Choice* **127**, 147 175.
URL: *https://doi.org/10.1007/s11127-006-0865-4*

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.

Manning, A. (2021), 'The elusive employment effect of the minimum wage', *Journal of Economic Perspectives* **35**(1), 3–26.
URL: *https://www.aeaweb.org/articles?id=10.1257/jep.35.1.3*

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

Wooldridge, J. (2005), *Introductory Econometrics: A Modern Approach*, 3 edn, South-Western College Pub.

Wooldridge, J. M. (2012), *Introductory Econometrics: A Modern Approach*, 5 edn, South-Western College Pub.