Intermediate Econometrics

Markus Haas

LMU München

Summer term 2011

15. Mai 2011

The Simple Linear Regression Model

- Considering variables x and y in a specific population (e.g., years of education and wage rate of full–time employees in Germany), we are interested in

  – investigating how y changes (on average) with changes in x,

  or (just put differently)

  – explaining (part of the variation of) y in terms of x.

- The **simple linear model** states that

$$y = \beta_0 + \beta_1 x + u \tag{1}$$

  describes the relation between y and x in our population of interest.

- In this framework, we will call

  – y the **dependent** or **explained** variable, and
  – x the **independent** or **explanatory** variable.
  – More technical language: x is the **regressor** and y the **regressand**

The Simple Linear Regression Model: $y = \beta_0 + \beta_1 x + u$

- Random variable u is called the **error term** or **disturbance**.

- Clearly we cannot expect that the relationship $y = \beta_0 + \beta_1 x$ holds exactly for each element in the population.

- The term u accounts for the *neglected* factors which have an impact on x but may not be observable.

- β_0 and β_1 are the parameters we are going to estimate.
  - The **slope parameter** β_1 measures the change in y in response to a one–unit change in x, provided the neglected factors in u are held fixed (the **ceteris paribus** effect).
  - The **intercept** β_0 may or may not be subject to meaningful interpretation, as discussed below.

The Simple Linear Regression Model: $y = \beta_0 + \beta_1 x + u$

- Our desire to figure out the partial effect of x on y raises questions about the nature of the factors determining u.

- In particular, is it reasonable to assume that u and x do not display systematic comovement?

- For example, consider an agricultural economist interested in the effect of a new fertilizer on crop yields.

- This economist may be able to randomly assign different amounts of the new fertilizer to several plots of land, i.e., independently of other plot features that affect crop yield.

- The data thus generated is **experimental data** and we can be sure that x and the neglected factors in u are independent.

- This is why, in the experimental sciences, x is also known as the "control variable" (since it is under control of the investigator).

The Simple Linear Regression Model: $y = \beta_0 + \beta_1 x + u$

- Unfortunately, this kind of data is rather untypical in economics and the social sciences in general, where we are (usually) dealing with **observational data**, where the researcher is just a passive collector of the data.

- For instance, to puzzle out the relation between education and wages, we cannot (and don't want to) randomly divide children into groups, allocate different amounts of education to these groups, and then observe the children's wage path after they mature and enter the labor force.

- In a setting more realistic than the above, education is affected by personal abilities and attitudes towards work, which, likewise, might have an impact on wage.

The Simple Linear Regression Model: $y = \beta_0 + \beta_1 x + u$

- To formalize this discussion, we shall make the assumption that the joint distribution of x and u is characterized by

$$\mathsf{E}(u|x) = \mathsf{E}(u) = 0, \tag{2}$$

i.e., u is **mean independent** of x.

- In this case, we also have

$$\mathsf{E}(y|x) = \beta_0 + \beta_1 x + \underbrace{\mathsf{E}(u|x)}_{=0} = \beta_0 + \beta_1 x, \tag{3}$$

i.e., a one–unit change in x leads to a change of β_1 units of the *expected value* of y, or, put differently, the regression tells us how *on average* (in the population) y changes in response to a change in x.

The Ordinary Least Squares (OLS) Estimator

- Assume that we observe a sample of size n,

$$\{(x_i, y_i) : i = 1, \ldots, n\}, \tag{4}$$

from the population generated by our simple linear regression model. We can write

$$y_i = \beta_0 + \beta_1 x_i + u_i, \quad i = 1, \ldots, n. \tag{5}$$

- Our estimates of β_0 and β_1 will be denoted by $\widehat{\beta}_0$ and $\widehat{\beta}_1$, and we may define the fitted values for y when $x = x_i$ as

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i, \tag{6}$$

and the **regression residuals**

$$\widehat{u}_i = y_i - \widehat{y}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i, \quad i = 1, \ldots, n. \tag{7}$$

The Ordinary Least Squares (OLS) Estimator

- Clearly, we would like to choose $\widehat{\beta}_0$ and $\widehat{\beta}_1$ such that the residuals $\widehat{u}_i$ are "small".

- The OLS approach is taken if $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are chosen such that the sum of squared differences is as small as possible, i.e., by minimizing

$$S(\widehat{\beta}_0, \widehat{\beta}_1) = \sum_{i=1}^{n} \widehat{u}_i^2 = \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2. \tag{8}$$

- The first–order conditions are

$$\frac{\partial S(\widehat{\beta}_0, \widehat{\beta}_1)}{\partial \widehat{\beta}_0} = -2 \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) = 0 \tag{9}$$

$$\frac{\partial S(\widehat{\beta}_0, \widehat{\beta}_1)}{\partial \widehat{\beta}_1} = -2 \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) x_i = 0. \tag{10}$$

Note that (9) and (10) imply

$$\sum_{i=1}^{n} \widehat{u}_i = 0, \quad \text{and} \quad \sum_{i=1}^{n} x_i \widehat{u}_i = 0, \tag{11}$$

respectively.

- Solving the first condition shows

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}, \tag{12}$$

where

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i, \quad \text{and} \quad \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{13}$$

are the sample means of y and x, respectively.

- (Rearranging (12) shows

$$\bar{y} = \widehat{\beta}_0 + \widehat{\beta}_1 \bar{x},$$

i.e., the point $(\bar{x}, \bar{y})$ will always be on the OLS regression line.)

- Inserting the solution for $\widehat{\beta}_0$ into the second condition, we get

$$\sum_{i=1}^{n}(y_i - \bar{y} - \widehat{\beta}_1(x_i - \bar{x}))x_i = 0 \tag{14}$$

$$\Rightarrow \widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})x_i}{\sum_{i=1}^{n}(x_i - \bar{x})x_i} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{15}$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2} \tag{16}$$

$$= \sum_{i=1}^{n} w_i y_i, \quad w_i = \frac{x_i - \bar{x}}{n s_x^2}, \tag{17}$$

where

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})y_i,$$

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})x_i.$$

- Note that we require $s_x^2 > 0$: If there is no variation in x, it is impossible to figure out how changes in x affect y.

Coefficient of Determination (R^2)

- Using $\overline{\widehat{y}} = n^{-1}\sum_{i=1}^{n}\widehat{y}_i = \bar{y}$ (since $\sum_i \widehat{u}_i = 0$), write

$$
\begin{aligned}
\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{=\text{SST}} &= \sum_{i=1}^{n}\{(\underbrace{y_i-\widehat{y}_i}_{=\widehat{u}_i})+(\widehat{y}_i-\bar{y})\}^2 \\
&= \underbrace{\sum_{i=1}^{n}\widehat{u}_i^2}_{=\text{SSR}} + \underbrace{\sum_{i=1}^{n}(\widehat{y}_i-\bar{y})^2}_{=\text{SSE}} + 2\underbrace{\sum_{i=1}^{n}\widehat{u}_i(\widehat{y}_i-\bar{y})}_{=0} \\
\text{SST} &= \text{SSR}+\text{SSE},
\end{aligned}
$$

where

- **SST** means **total sum of squares**, measuring the total sample variation in the y_is,

- **SSE** means **explained sum of squares**, i.e., the part of variation in the y_is that is explained by the fitted regression line,

- **SSR** means **residual sum of squares**, i.e., the part of the variation that is not explained by the fitted line.

- The result follows from

$$\begin{aligned}\sum_i \widehat{u}_i(\widehat{y}_i - \bar{y}) &= \sum_i \widehat{u}_i(\widehat{\beta}_0 + \widehat{\beta}_1 x_i - \widehat{\beta}_0 - \widehat{\beta}_1 \bar{x}) \qquad (18)\\ &= \widehat{\beta}_1 \underbrace{\sum_i \widehat{u}_i x_i}_{=0} - \widehat{\beta}_1 \bar{x} \underbrace{\sum_i \widehat{u}_i}_{=0} = 0,\end{aligned}$$

which says that the sample correlation between the fitted values ($\widehat{y}_i$) and the regression residuals is zero.

- Then the coefficient of determination, R^2,

$$R^2 = \frac{\mathsf{SSE}}{\mathsf{SST}} = 1 - \frac{\mathsf{SSR}}{\mathsf{SST}}, \qquad (19)$$

which can be interpreted as the *fraction of the sample variation in y that is explained by x* (via the fitted linear regression line).

- Clearly
$$0 \le R^2 \le 1.$$

- The term R is used because R^2 is actually just the squared correlation between the sample values of x and y.

- Write

$$R^2 = \frac{\mathsf{SSE}}{\mathsf{SST}} = \frac{\frac{1}{n}\sum_i(\widehat{y}_i - \bar{y})^2}{s_y^2} = \frac{\widehat{\beta}_1^2 \overbrace{\frac{1}{n}\sum_i (x_i - \bar{x})^2}^{=s_x^2}}{s_y^2}$$

$$= \underbrace{\widehat{\beta}_1^2}_{=(s_{xy}/s_x^2)^2} \frac{s_x^2}{s_y^2} = \left(\frac{s_{xy}}{\sqrt{s_x^2 s_y^2}}\right)^2,$$

and $s_{xy}/\sqrt{s_x^2 s_y^2}$ is just the definition of the sample correlation coefficient between x and y.

Statistical Properties of OLS: Assumptions (Gauß–Markov Assumptions)

- We need the following assumptions:

1) The linear model is correctly specified, i.e., y is related to x and u as $y = \beta_0 + \beta_1 x + u$. (Linearity in Parameters)
2) We observe a **random sample** of size n, $\{(x_i, y_i) : i = 1, \ldots, n\}$, generated from the linear model, i.e., $\{(x_i, u_i), i = 1, \ldots, n\}$ are drawn independently from a common distribution.
3) $s_x^2 > 0$
4) $\mathsf{E}(u|x) = 0$
5) $\mathsf{E}(u^2|x) = \mathsf{Var}(y|x) = \sigma^2$, i.e., the variance of u (and hence y) does not depend on x (**Homoskedasticity**).

- Random sampling (Assumption 2) means that observation (x_i, u_i) (and hence also y_i) is drawn independently from (x_j, u_j) for $i \neq j$ (but from the same distribution, i.e., they are *independently and identically distributed (iid)*).

- This may often be deemed realistic for **cross–sectional** data. Cross–sectional data sets consist, e.g., of a sample of individuals, households, firms, or other units, taken randomly at a given point in time.

- This is in contrast to **time series data**, where, for example, the growth rate of GDP this year is not independent of the last year's growth rate.

- Note that the assumption of random sampling implies

$$\mathsf{E}(u_i u_j) = \mathsf{Cov}(u_i, u_j) = 0, \quad i \neq j. \tag{20}$$

Statistical Properties of OLS: Unbiasedness

- Let us consider the expected value of $\widehat{\beta}_1$. We have, from (14)–(17),

$$\begin{aligned} \widehat{\beta}_1 &= \sum_i w_i y_i = \sum_i \frac{x_i - \bar{x}}{n s_x^2} y_i \\ &= \sum_i \frac{x_i - \bar{x}}{n s_x^2} (\beta_0 + \beta_1 x_i + u_i), \end{aligned} \tag{21}$$

  where in the second line we made use of Assumption 1 (correct model), i.e., we just substituted $\beta_0 + \beta_1 x_i + u_i$ for y_i; note that this involves the true (but unknown and unobservable) β_0, β_1, and u_is.

- Now,

$$\begin{aligned} \sum_i \frac{x_i - \bar{x}}{n s_x^2} \beta_0 &= \frac{\beta_0}{n s_x^2} \sum_i (x_i - \bar{x}) = 0, \\ \sum_i \frac{x_i - \bar{x}}{n s_x^2} \beta_1 x_i &= \beta_1 \frac{1}{s_x^2} \underbrace{\frac{1}{n} \sum_i (x_i - \bar{x}) x_i}_{s_x^2} = \beta_1. \end{aligned}$$

- Thus,

$$\widehat{\beta}_1 = \beta_1 + \sum_{i=1}^{n} w_i u_i, \tag{22}$$

and taking the expectation, employing Assumption 4, leads to

$$\mathsf{E}(\widehat{\beta}_1) = \beta_1 + \sum_{i=1}^{n} w_i \underbrace{\mathsf{E}(u_i)}_{=0} = \beta_1. \tag{23}$$

- For $\widehat{\beta}_0$, we write, substituting $\beta_0 + \beta_1 \bar{x} + \bar{u}$ for $\bar{y}$ (Assumption 1)

$$\begin{aligned} \widehat{\beta}_0 &= \bar{y} - \widehat{\beta}_1 \bar{x} \\ &= \beta_0 + \beta_1 \bar{x} + \bar{u} - \widehat{\beta}_1 \bar{x} \\ &= \beta_0 + \bar{u} + (\beta_1 - \widehat{\beta}_1)\bar{x}. \end{aligned}$$

- Taking the expectation, and using that $\mathsf{E}(\widehat{\beta}_1) = \beta_1$, we get

$$\mathsf{E}(\widehat{\beta}_0) = \beta_0. \tag{24}$$

- Equations (23) and (24) imply that the OLS estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are **unbiased** for β_0 and β_1.

- Note that homoskedasticty (Assumption 5) and (20) are not required for establishing unbiasedness.

Statistical Properties of OLS: Unbiasedness

- Recall what unbiasedness means:

- Unbiasedness means that if we could infinitely draw random samples from our population, then *on average* we would obtain the correct estimates of the parameters we are interested in.

- It does not mean that, for example, $\widehat{\beta}_1 = \beta_1$, or even that, *for our particular sample*, $\widehat{\beta}_1$ is very close to β_1.

Statistical Properties of OLS: Unbiasedness

- The objections to the unbiasedness criterion can be summarized by the following story.[1]

- Three econometricians go duck hunting. The first shoots about a foot in front of the duck, the second about a foot behind; the third yells "We got it!"

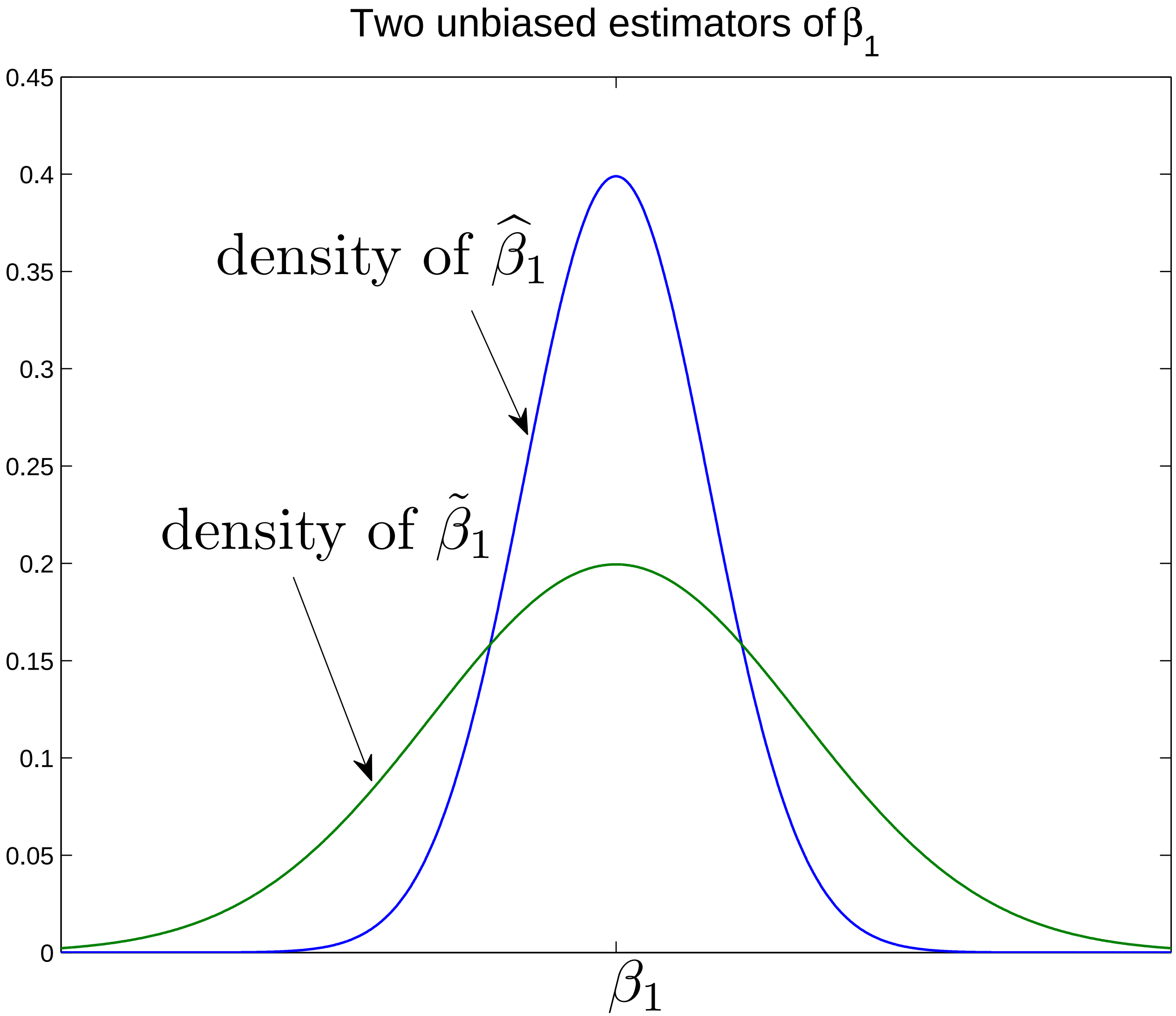
[1] See, e.g., Peter Kennedy, *A Guide to Econometrics*, 6e, p. 30.

Statistical Properties of OLS: Unbiasedness

- Still, however, unbiasedness is often deemed desirable, since we might prefer an estimator with a sampling distribution centered over the true parameter rather than one with a sampling distribution centered over some other value.

- However, it is still possible to have an “unlucky” sample and thus a bad estimate.

- Therefore, the unbiasedness criterion needs to be complemented by a measure that tells us how probable an “unlucky” sample actually is.


Two unbiased estimators of β1
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
density of $\widehat{\beta}_1$
density of $\tilde{\beta}_1$
β_1

The Variances of the OLS Estimators

- Recall from above that

$$\widehat{\beta}_1 = \beta_1 + \sum_{i=1}^{n} w_i u_i. \tag{25}$$

- Also recall the following formula for the variance of a linear combination of random variables $y_1, \ldots, y_n$:

$$\text{Var}\left(\sum_{i=1}^{n} a_i y_i\right) = \sum_{i=1}^{n} a_i^2 \text{Var}(y_i) + 2\sum_{i=1}^{n}\sum_{j<i} a_i a_j \text{Cov}(y_i, y_j). \tag{26}$$

- However, when applied to (20), all the covariance terms in (26) turn out to be zero due to (20).

- Moreover, due to **homoskedasticity** (Assumption 5), the variance of all u_is is just σ^2.

- Hence

$$\mathrm{Var}(\widehat{\beta}_1) = \sigma^2 \sum_{i=1}^{n} w_i^2 = \sigma^2 \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n^2 s_x^4} = \frac{\sigma^2}{n s_x^2}.$$

- For $\widehat{\beta}_0$, we can use the representation

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x},$$

which gives

$$\mathrm{Var}(\widehat{\beta}_0) = \mathrm{Var}(\bar{y}) + \bar{x}^2 \mathrm{Var}(\widehat{\beta}_1) - 2\bar{x}\mathrm{Cov}(\bar{y}, \widehat{\beta}_1).$$

- For $\mathrm{Cov}(\bar{y}, \widehat{\beta}_1)$, we find

$$\begin{aligned} \mathrm{Cov}(\bar{y}, \widehat{\beta}_1) &= \mathrm{Cov}\left(\frac{1}{n}\sum_{i=1}^{n} y_i, \sum_{i=1}^{n} w_i y_i\right) \\ &= \frac{\sigma^2}{n}\sum_{i=1}^{n} w_i = 0, \end{aligned}$$

and so

$$
\begin{aligned}
\mathsf{Var}(\widehat{\beta}_0) &= \mathsf{Var}(\bar{y}) + \bar{x}^2 \mathsf{Var}(\widehat{\beta}_1) \\
&= \frac{\sigma^2}{n}\left(1 + \frac{\bar{x}^2}{s_x^2}\right) \\
&= \frac{\sigma^2 \overline{x^2}}{n s_x^2},
\end{aligned}
$$

since

$$
s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \overline{x^2} - \bar{x}^2.
$$

Gauß–Markov Theorem

- It turns out that the OLS estimator has several favorable properties when compared to alternative estimators within a particular class.

- This is the class of estimators that are
  - unbiased, and
  - linear in the y_is (or, equivalently, the u_is).

- In this class, OLS has the smallest variance.

- This property is called BLUE (**B**est **L**inear **U**nbiased **E**stimator).

- This is the Gauß–Markov Theorem, which holds under Assumptions 1–5 (Gauß–Markov Assumptions) listed above.

Estimation of the Error Variance

- To estimate the variances of $\widehat{\beta}_0$ and $\widehat{\beta}_1$, we need an estimate of σ^2, i.e., the error variance.

- It turns out that
$$\widehat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \widehat{u}_i^2 \tag{27}$$
is the appropriate estimator.

- (27) is unbiased, i.e.,
$$\mathsf{E}(\widehat{\sigma}^2) = \sigma^2.$$

Why we often assume that relationships are linear

- The "true" relationship between the variables may often be at least approximately linear over the range of relevant values.

- This point is related to why the coefficient $\widehat{\beta}$ often has no meaningful interpretation.

- For example, if, in a regression of *wage* on *years of education*, we estimate a negative slope coefficient, this is not to be interpreted in the sense that people without education receive negative wage.

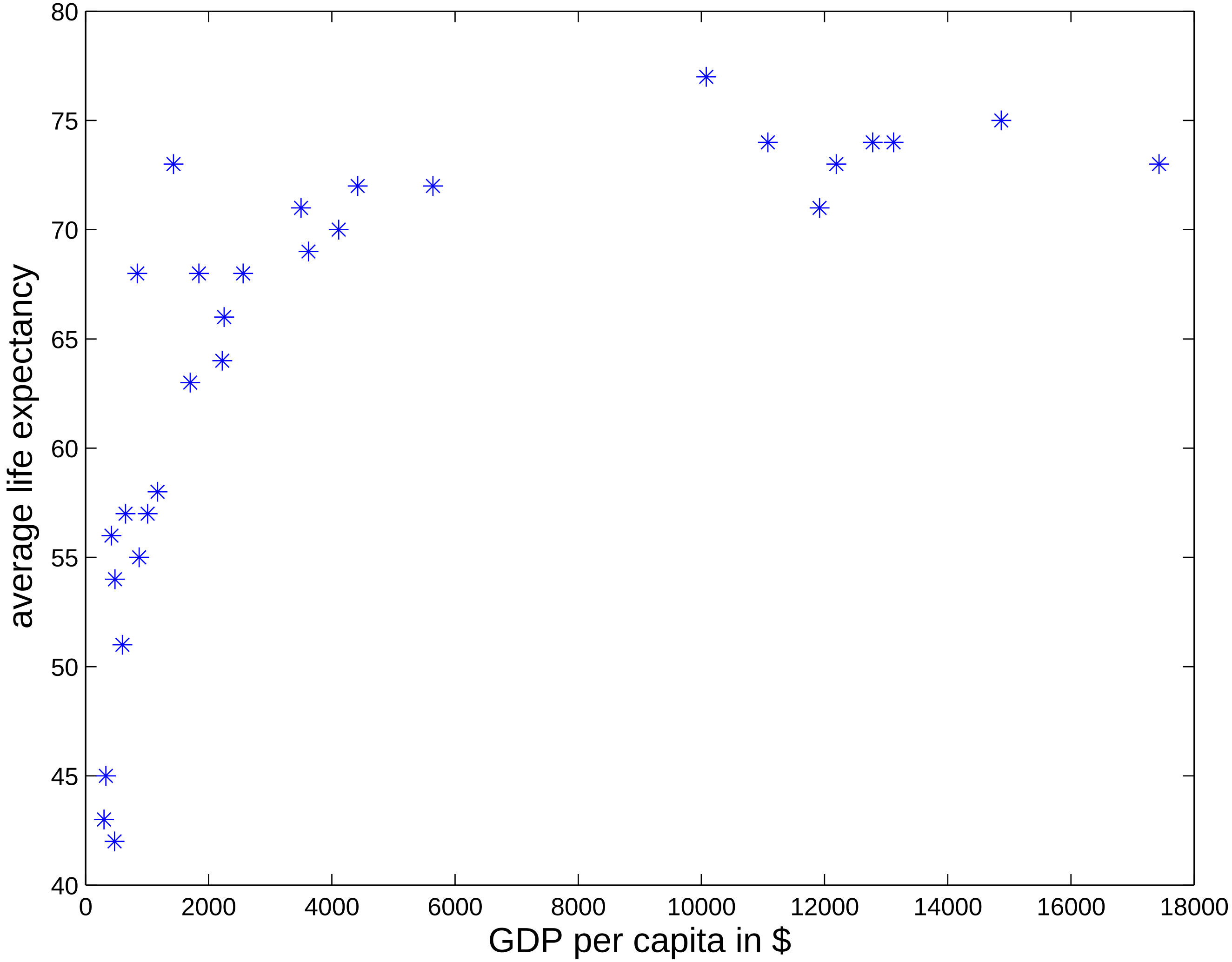
- Realistically, most people have at least 8–10 years of education, but the line that exhibits the best fit to the observations over the relevant range does not go through the origin.

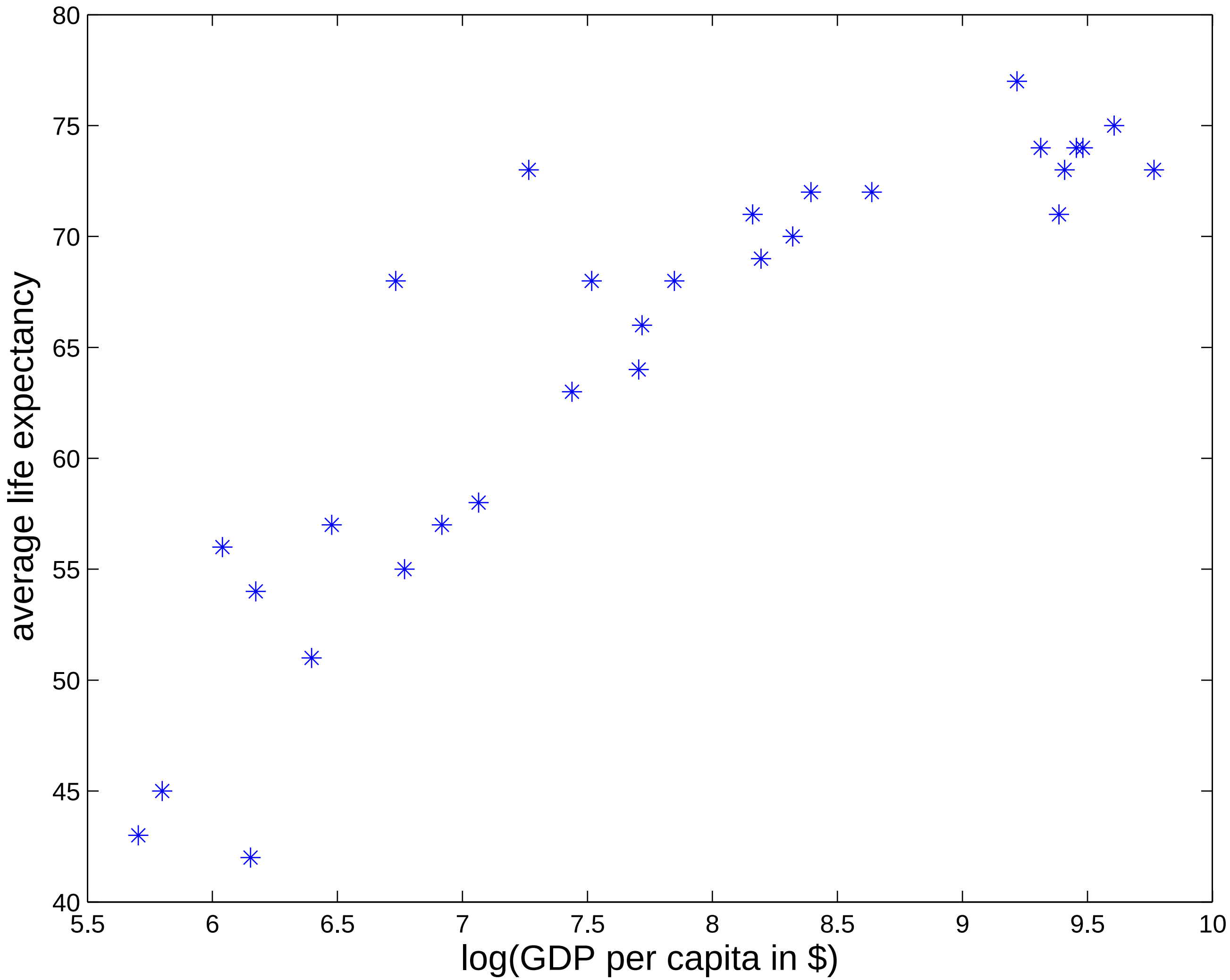
- This means, of course, that even if β_0 has no substantial interpretation, *we must not ignore it* in the applications of our model.

Why we often assume that relationships are linear

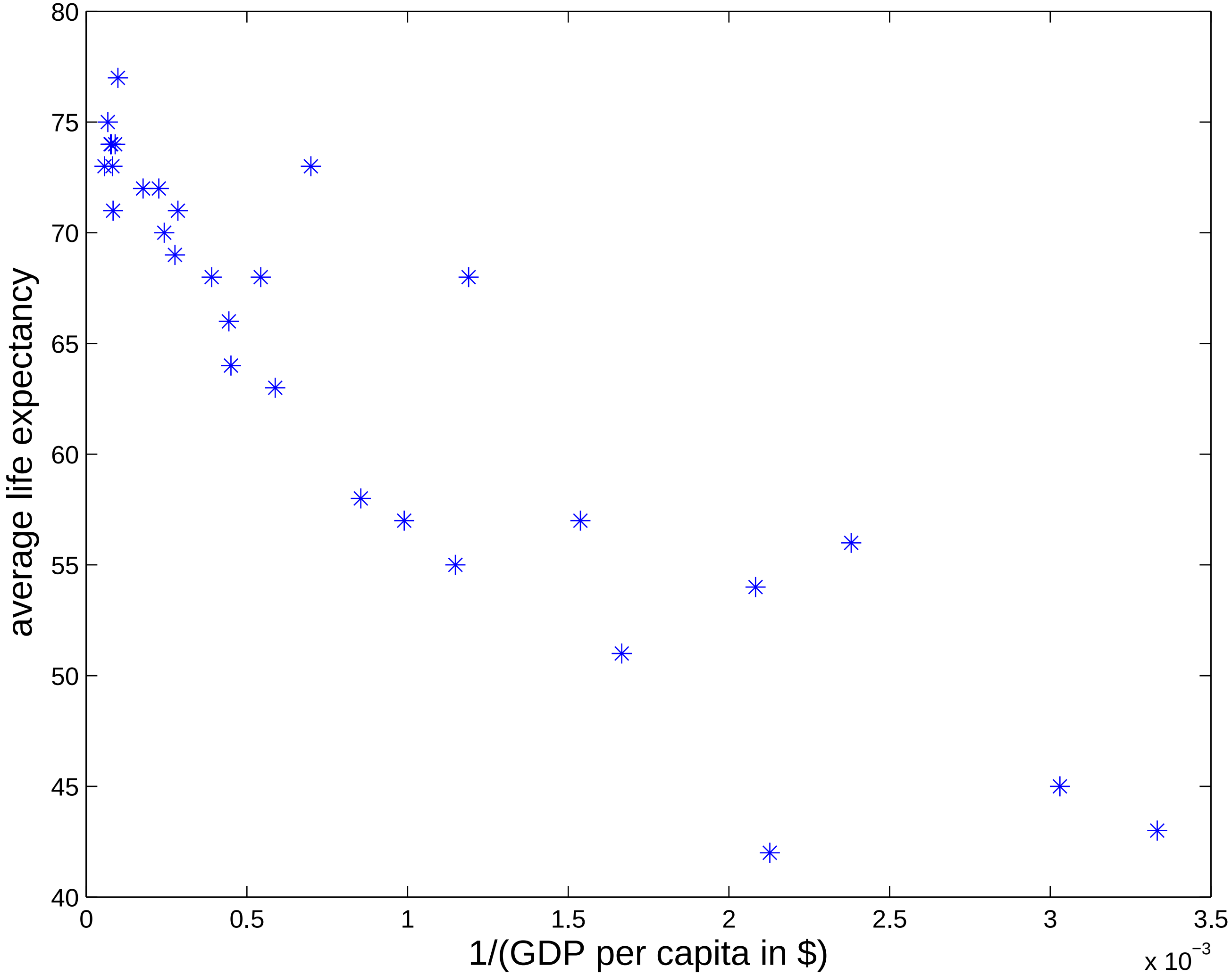
- Even if this is not the case, we can often transform the variables in such a way as to linearize the relationship.

- Consider the following example of the relationship between average GDP and average lifespan for several countries (1984).


average life expectancy
GDP per capita in $
80
75
70
65
60
55
50
45
40
0
2000
4000
6000
8000
10000
12000
14000
16000
18000



80
75
70
65
60
55
50
45
40
average life expectancy
5.5
6
6.5
7
7.5
8
8.5
9
9.5
10
log(GDP per capita in $)



80
75
70
65
60
55
50
45
40
average life expectancy
0
0.5
1
1.5
2
2.5
3
3.5
1/(GDP per capita in $)
x 10^-3

Why we often assume that relationships are linear

- In economics, variables often appear in logarithmic form.

- For example, consider the equation

$$\log(wage) = \beta_0 + \beta_1 edu + u, \tag{28}$$

where *edu* is years of education.

- In this case, the equation says that (on average) each additional year of education increases the wage by a constant percentage.

- Recall that

$$\frac{d\log y}{dy} = \frac{1}{y} \Rightarrow \Delta \log y \approx d\log y = \frac{dy}{y} \approx \frac{\Delta y}{y},$$

so that

$$\Delta \log y = \beta_1 \Delta x,$$

is an approximate measure of the percentage change of y in response to the change Δx in x, i.e.,

$$\%\Delta y = (100 \times \beta_1)\Delta x, \tag{29}$$

where the multiplication by 100 is to get real *percentages*, e.g., 10% instead of 0.1.

- If both variables are in logarithmic form,

$$\log y = \beta_0 + \beta_1 \log x + u,$$

then we have (on average)

$$\frac{d \log y}{d \log x} = \frac{dy/y}{dx/x} \approx \frac{\Delta y/y}{\Delta x/x} = \beta_1,$$

so β_1 measures the **elasticity** of y with respect to x, i.e.,

- β_1 measures the *percentage* change in y when x increases by 1%.

- This is often considered in price theory; for example, elasticity of demand with respect to price or income.

- In Equation (28), coefficient β_1 is also termed **semi–elasticity**, since it measures the percentage change in y when x increases by one *unit*.

- The third possibility is that only x comes in log–form,

$$y = \beta_0 + \beta_1 \log x + u,$$

where $\beta_1/100$ (approximately) measures the response in y in response to a one–percent increase in x, since

$$\Delta y = \beta_1 \Delta \log x = \left(\frac{\beta_1}{100}\right)(100 \times \Delta \log x) = \left(\frac{\beta_1}{100}\right) \% \Delta x,$$

where $\% \Delta x$ is measured in "real" percentages, e.g., 1% instead of 0.01.