Econometrics Lecture Notes I

Simple Regression (close to chapters 1-2, Wooldridge)

M Loecher

Simple Regression, Examples

CEO Salary

We fit a simple model of CEO salaries versus sales of a company:

- *salary*: 1990 salary, thousands \$
- *sales*: 1990 firm sales, millions \$

$$\begin{aligned} \widehat{salary_i} = & \; 1174 \;\; + 0.0155 \; sales_i \\ & \; (113) \quad\;\; (0.00891) \\ & \; n = 209, R^2 = 0.01 \end{aligned} \tag{1}$$

Education and Wages

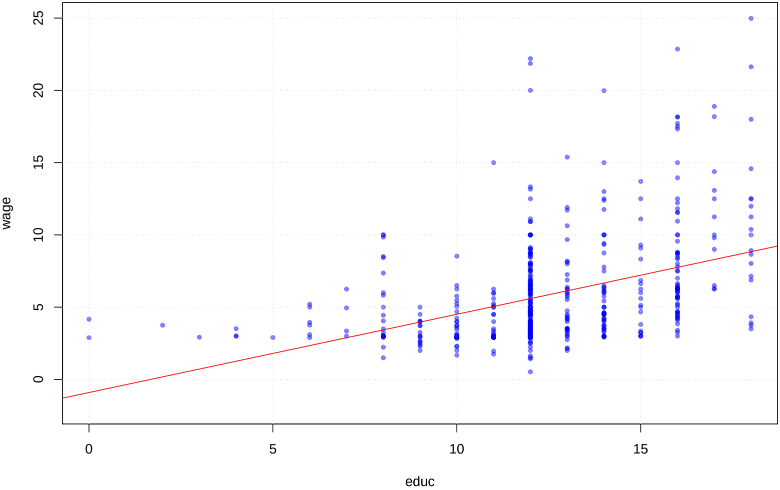
For the population of people in the workforce in 1976, let y be wage measured in dollars per hour. Let $x = educ$ denote years of schooling; for example, $educ = 12$ correspond to a complete high school education

$$wage_i = \beta_0 + \beta_1 \, educ_i + u_i \tag{2}$$

OLS yields

$$\begin{aligned} [2.27] \quad \widehat{wage_i} &= -0.905 + 0.541 \, educ_i \\ &\quad\; (0.685) \qquad (0.0532) \\ &\quad\; n = 526, R^2 = 0.16 \end{aligned} \tag{3}$$

Graphical View



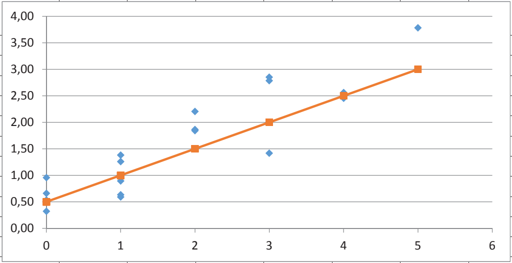
From Population to Sample

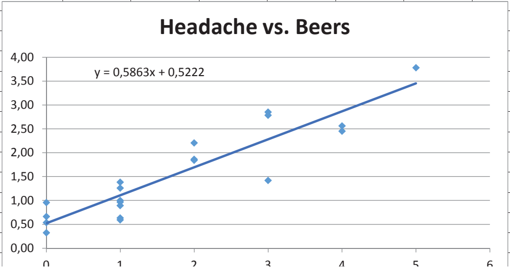
Regression Hat Notation

$$
\begin{aligned}
[2.9] \quad y_i &= \beta_0 + \beta_1 x_i + u_i && (4) \\
[2.20] \quad \hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i && (5) \\
[2.21] \quad \hat{u}_i &= y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i && (6)
\end{aligned}
$$

Regression Hat Notation II

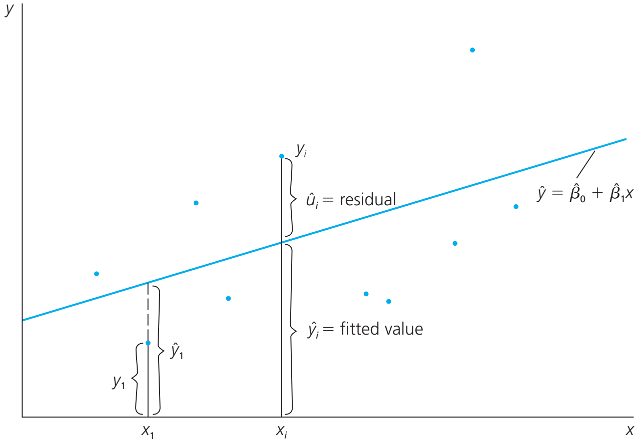
Beers	Head ache	y - u	u	ŷ	û
1	0,63	1	-0,37		
2	1,86	1,5	0,36		
1	0,89	1	-0,11		
3	2,85	2	0,85		
4	2,56	2,5	0,06		
0	0,32	0,5	-0,18		
1	1,00	1	0,00		
3	2,79	2	0,79		
1	1,38	1	0,38		
1	0,97	1	-0,03		
4	2,45	2,5	-0,05		
1	0,59	1	-0,41		
2	1,84	1,5	0,34		
2	2,21	1,5	0,71		
0	0,66	0,5	0,16		
3	1,42	2	-0,58		
1	1,26	1	0,26		
0	0,96	0,5	0,46		
5	3,78	3	0,78		
0	0,54	0,5	0,04		





Regression Hat Notation III

FIGURE 2.4 Fitted values and residuals.



Sum of Squares

Define the **total sum of squares (SST)**, the **explained sum of squares (SSE)**, and the **residual sum of squares (SSR)** (also known as the sum of squared residuals), as follows:

$$[2.33] \quad SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad (= SS_y) \tag{7}$$

$$[2.34] \quad SSE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \quad (= SS_{\hat{y}}) \tag{8}$$

$$[2.35] \quad SSR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \hat{u}_i^2 \quad (= SS_{\hat{u}} = RSS) \tag{9}$$

Terminology

Some words of caution about SST, SSE, and SSR are in order. There is no uniform agreement on the names or abbreviations for the three quantities defined in equations (7), (8), and (9). The total sum of squares is called either SST or TSS, so there is little confusion here. Unfortunately, the explained sum of squares is sometimes called the "regression sum of squares." If this term is given its natural abbreviation, it can easily be confused with the term "residual sum of squares." Some regression packages refer to the explained sum of squares as the "model sum of squares." To make matters even worse, the residual sum of squares is often called the "error sum of squares." This is especially unfortunate because the errors and the residuals are different quantities. Thus, Wooldridge always calls (2.35) the residual sum of squares or the sum of squared residuals.

Relationship to variance

Note that "sum of squares" and variances are closely related. In fact, the variance was introduced simply as the **average sum of squared deviation from the mean**, so $\sigma_Y^2 = SST/n$.

If we are careful about the exact *degrees of freedom*, we can generalize this relationship to $\sigma_X^2 = SS_X/df$, e.g.

$$\sigma_u^2 = SSR/df = SSR/(n - k - 1)$$

In a way, the notion of "sum of squares" frees us from constantly having to keep the degrees of freedom around!

Exercises

In the gretl output below, identify SST, SSE and SSR.

Model 1: OLS, using observations 1–526
Dependent variable: wage

	Coefficient	Std. Error	t-ratio	p-value
const	−0.904852	0.684968	−1.3210	0.1871
educ	0.541359	0.0532480	10.1667	0.0000

Mean dependent var	5.896103	S.D. dependent var	3.693086
Sum squared resid	5980.682	S.E. of regression	3.378390
R^2	0.164758	Adjusted R^2	0.163164
$F(1, 524)$	103.3627	P-value(F)	2.78e–22
Log-likelihood	−1385.712	Akaike criterion	2775.423
Schwarz criterion	2783.954	Hannan–Quinn	2778.764

Goodness-of-Fit

The **R-squared** of the regression, sometimes called the **coefficient of determination**, is defined as

$$[2.38] \quad R^2 = SSE/SST = 1 - SSR/SST \tag{10}$$

$R^2 \in [0; 1]$ is the ratio of the explained variation compared to the total variation; thus, it is interpreted as the fraction of the sample variation in y that is explained by x.

It can be shown that R^2 is equal to the square of the sample correlation coefficient between y_i and $\hat{y}_i$. This is where the term "R-squared" came from.

Regression Coefficients

Covariance

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

Algebraic Identity

Show that if either one (sample) mean is exactly zero, the covariance can be written much more concisely:

$$\hat{\sigma}_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} x_i \cdot y_i$$

Coefficient Estimates

Minimizing the squared residuals (hence the name "Least Squares") yields

$$[2.19] \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{11}$$

Equation (2.19) is simply the sample covariance between x and y divided by the sample variance of x: $\hat{\beta}_1 = cov_{xy}/var_x$. (why?)

Rewriting:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}\hat{r}_{x,i}y_i}{\sum_{i=1}^{n}\hat{r}_{x,i}^2} \tag{12}$$

with the "x residual" defined as simply: $\hat{r}_{x,i} \equiv (x_i - \bar{x})$

Properties of Coefficients

-The slope and correlation coefficient are closely related:

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2} = \frac{s_y}{s_x} \cdot \frac{s_{xy}}{s_x \cdot s_y} = \frac{s_y}{s_x} \cdot r$$

-The slope changes if we rescale variables but the correlation coefficient does not:

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2} = \frac{s_y}{s_x} \cdot \frac{s_{xy}}{s_x \cdot s_y} = \frac{s_y}{s_x} \cdot r$$

- The point $(\bar{x}, \bar{y})$ is always on the OLS regression line.

$$[2.16] \quad \bar{y}(= \bar{\hat{y}}) = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} \Leftrightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad (13)$$

Algebraic Properties of OLS Statistics

1. The sum, and therefore the sample average of the OLS residuals, is zero

$$[2.30] \quad \sum_{i=1}^{n} \hat{u}_i = 0 \tag{14}$$

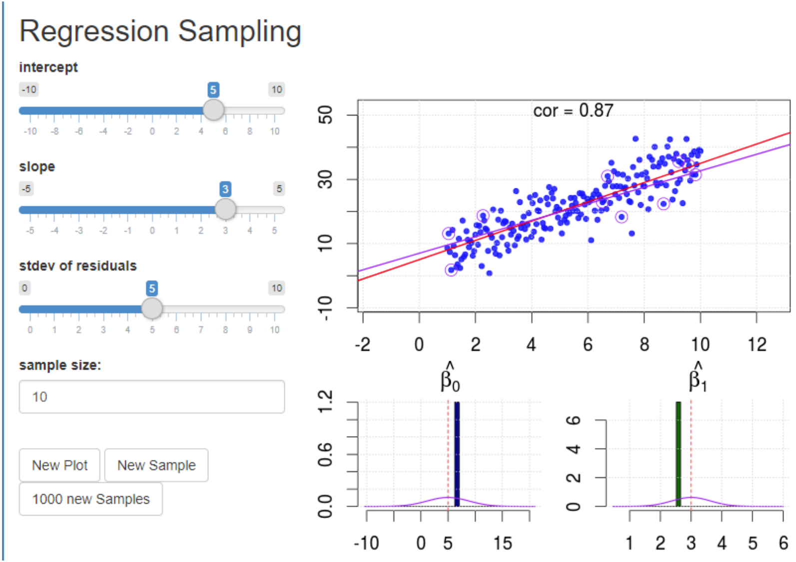
2. The sample covariance between the regressors and the OLS residuals is zero.

$$[2.31] \quad \sum_{i=1}^{n} x_i \cdot \hat{u}_i = 0 \tag{15}$$

3. The slope and intercept are negatively correlated.

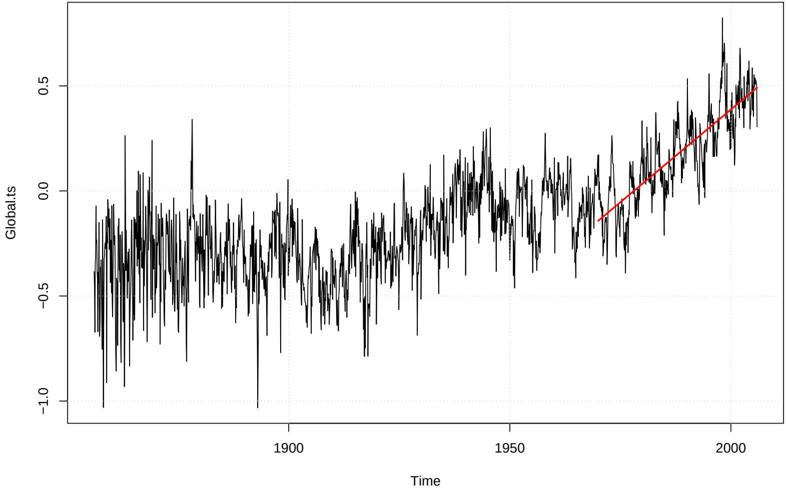
Correlation of coefficients

https://codeandstats.shinyapps.io/regression/



Global Warming

$$E(T|t) = E(T)?$$



Uncertainty of slope

OLS yields

$$\begin{aligned}\widehat{Temperature}_i &= -34.9 + 0.0177\ time_i \\ &\quad\ (1.16) \qquad (0.000586) \\ &\quad\ n = 432, R^2 = 0.68\end{aligned} \tag{16}$$

Election outcomes and campaign expenditures

Data for 173 two-party races for the U.S. House of Representatives in 1988. There are two candidates in each race, A and B. Let voteA be the percentage of the vote received by Candidate A and shareA be the percentage of total campaign expenditures accounted for by Candidate A. Many factors other than shareA affect the election outcome (including the quality of the candidates and possibly the dollar amounts spent by A and B). Nevertheless, we can estimate a simple regression model to find out whether spending more relative to one's challenger implies a higher percentage of the vote.

$$\begin{aligned} \widehat{voteA_i} = \; & \underset{(0.887)}{26.8} \quad + \underset{(0.0145)}{0.464} \; shareA_i \\ & n = 173, R^2 = 0.86 \end{aligned} \tag{17}$$

Logarithmic Transformations

Summary, Logarithms

TABLE 2.3 Summary of Functional Forms Involving Logarithms

Model	Dependent Variable	Independent Variable	Interpretation of β_1
Level-level	y	x	$\Delta y = \beta_1 \Delta x$
Level-log	y	$\log(x)$	$\Delta y = (\beta_1/100)\%\Delta x$
Log-level	$\log(y)$	x	$\%\Delta y = (100\beta_1)\Delta x$
Log-log	$\log(y)$	$\log(x)$	$\%\Delta y = \beta_1 \%\Delta x$

Example, Elasticity

We can estimate a constant elasticity model relating CEO salary to firm sales. Let sales be annual firm sales, measured in millions of dollars. A constant elasticity model is

$$[2.45] \quad log(salary)_i = \beta_0 + \beta_1 \, log(sales)_i + u_i \tag{18}$$

where β_1 is the elasticity of salary with respect to sales. Estimating this equation by OLS gives

$$\begin{aligned}[2.46] \quad log\widehat{(salary)}_i = \; & 4.82 \qquad + \; 0.257 \; log(sales)_i \\ & (0.288) \qquad (0.0345) \\ & n = 209, R^2 = 0.21 \end{aligned} \tag{19}$$

The coefficient of log(sales) is the estimated elasticity of salary with respect to sales. It implies that a 1% increase in firm sales increases CEO salary by about 0.257%.

Econometric software reports

Level-Level

Model 1: OLS, using observations 1–526
Dependent variable: wage

	Coefficient	Std. Error	t-ratio	p-value
const	5.373	0.257	20.908	0.000
exper	0.031	0.012	2.601	0.010

Mean dep var	5.9	S.D. dep var	3.69
Sum sq resid	7069.14	S.E. of regr	3.67
R^2	0.01	Adjusted R^2	0.01
$F(1, 524)$	6.77	P-value(F)	0.01
Log-likelihood	−1429.686	Akaike criterion	2863.373
Schwarz criterion	2871.9	Hannan–Quinn	2866.71

Log-Level

Model 2: OLS, using observations 1–526
Dependent variable: lwage

	Coefficient	Std. Error	t-ratio	p-value
const	1.549	0.037	41.872	0.000
exper	0.004	0.002	2.565	0.011

Mean dep var	1.62	S.D. dep var	0.53
Sum sq resid	146.49	S.E. of regr	0.53
R^2	0.01	Adjusted R^2	0.01
$F(1, 524)$	6.58	P-value(F)	0.01
Log-likelihood	−410.1570	Akaike criterion	824.3140
Schwarz criterion	832.84	Hannan–Quinn	827.65

Level-Log

Model 3: OLS, using observations 1–526
Dependent variable: wage

	Coefficient	Std. Error	t-ratio	p-value
const	4.12016	0.387587	10.6303	0.0000
l_exper	0.741691	0.147911	5.0144	0.0000

Mean dependent var	5.896103	S.D. dependent var	3.693086
Sum squared resid	6832.550	S.E. of regression	3.610986
R^2	0.045789	Adjusted R^2	0.043968
$F(1, 524)$	25.14452	P-value(F)	7.29e–07
Log-likelihood	−1420.734	Akaike criterion	2845.467
Schwarz criterion	2853.998	Hannan–Quinn	2848.807

Log-Log

Model 4: OLS, using observations 1–526
Dependent variable: lwage

	Coefficient	Std. Error	t-ratio	p-value
const	1.34346	0.0555185	24.1984	0.0000
l_exper	0.116858	0.0211870	5.5156	0.0000

Mean dependent var	1.623268	S.D. dependent var	0.531538
Sum squared resid	140.1908	S.E. of regression	0.517242
R^2	0.054871	Adjusted R^2	0.053067
$F(1, 524)$	30.42148	P-value(F)	5.47e–08
Log-likelihood	−398.5976	Akaike criterion	801.1952
Schwarz criterion	809.7258	Hannan–Quinn	804.5353

The Gauss-Markov Assumptions for Simple Regression

SLR.1-SLR.2

Only SLR.1 through SLR.4 are needed to show that $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased.

Assumption SLR.1 (Linear in Parameters)

In the population model, the dependent variable, y, is related to the independent variable, x, and the error (or disturbance), u, as

$$y_i = \beta_0 + \beta_1 x_i + u_i \tag{20}$$

where β_0 and β_1 are the population intercept and slope parameters, respectively.

Assumption SLR.2 (Random Sampling)

We have a random sample of size n, $(x_i, y_i) : i = 1, 2, \ldots, n$, following the population model in Assumption SLR.1.

SLR.3-SLR.4

Assumption SLR.3 (Sample Variation in the Explanatory Variable)

The sample outcomes on x, namely, $(x_i) : i = 1, 2, \ldots, n$, are not all the same value.

Assumption SLR.4 (Zero Conditional Mean)

The error u has an expected value of zero given any value of the explanatory variable. In other words,

$$E(u|x) = E(u) = 0 \tag{21}$$

SLR.5

The homoskedasticity assumption, SLR.5, leads to the OLS variance formulas [2.57] and [2.58].

Assumption SLR.5 (Homoskedasticity)

The error u has the same variance given any value of the explanatory variable. In other words,

$$Var(u|x) = Var(u) = \sigma^2 = \sigma_u^2 \tag{22}$$

Gauss Markov

Under Assumptions SLR.1 through SLR.5, all our estimates are **unbiased**

$$[2.53] \quad E(\hat{\beta}_i) = \beta_i \tag{23}$$

Furthermore:

$$E(\hat{\sigma_u^2}) = \sigma_u^2 \tag{24}$$

Variance of the Estimators

$$[2.57] \quad Var(\hat{\beta}_1) = \sigma_u^2 / SST_x = \sigma_u^2 \cdot \Big(1/n + \bar{x}^2 / SST_x\Big) \tag{25}$$

$$[2.58] \quad Var(\hat{\beta}_0) = \sigma_u^2 \cdot \overline{x^2} / SST_x \tag{26}$$

$$cov(\hat{\beta}_0, \hat{\beta}_1) = -\sigma_u^2 \cdot \bar{x} / SST_x \tag{27}$$

Zero Conditional Mean

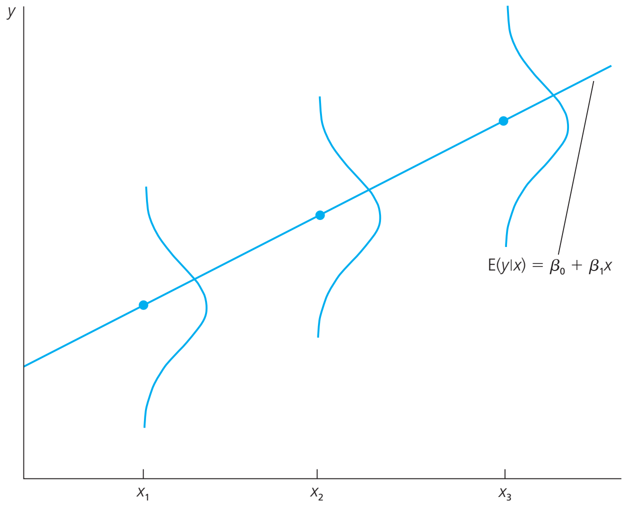
Important: note that assumption SLR.4 $E(u|x) = 0$ is by the far the most important one and its violation has far reaching consequences. We will later see that the most common reason for $E(u|x) \neq 0$ are variables left out of the regression that correlate both with x and y.

We often say that x is an **exogenous** explanatory variables if SLR.4 (22) holds true. If x is correlated with u for any reason, then x is said to be an **endogenous** explanatory variable

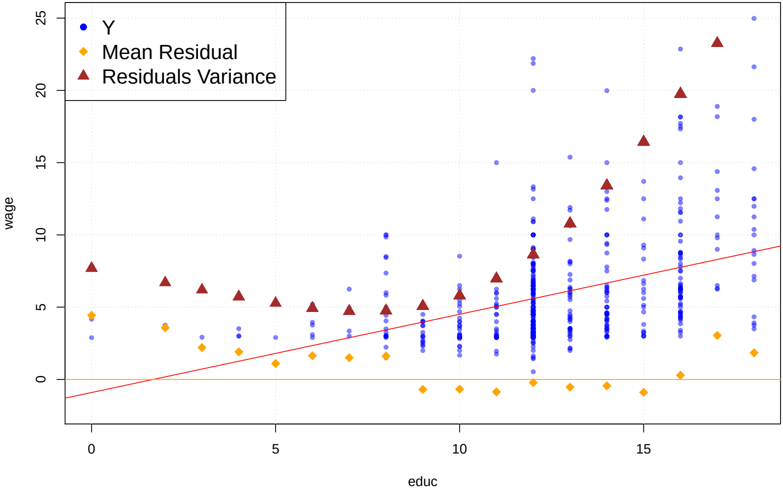
Visualization, Assumptions on u

We are close to the stronger assumption $u \sim N(0, \sigma)$

FIGURE 2.1 E(y|x) as a linear function of x.

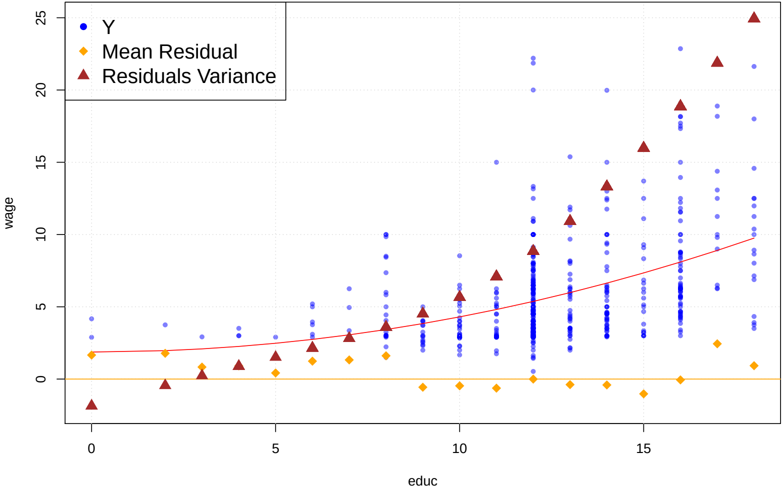


Violation of which assumptions?



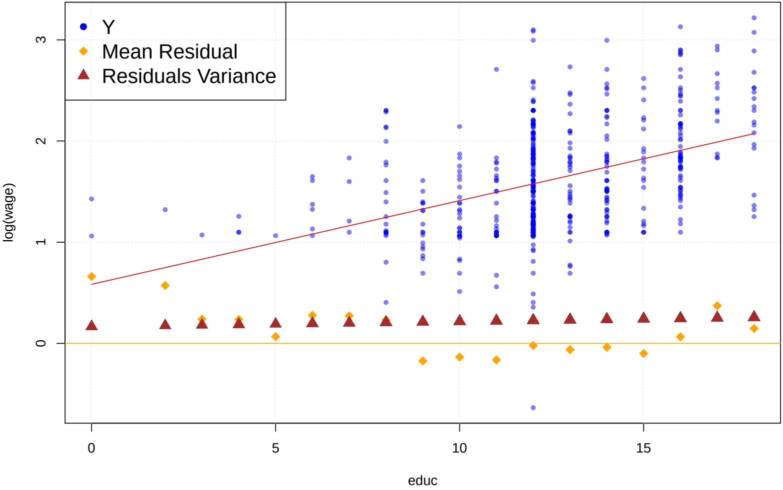
Nonlinear Fit I

$$\widehat{wage_i} = \underset{(0.39)}{1.87} + \underset{(0.00219)}{0.0243}\ educ_i^2 \qquad (28)$$
$$n = 526, R^2 = 0.19$$



Nonlinear Fit II

$$\widehat{log(wage)}_i = \underset{(0.0973)}{0.584} + \underset{(0.00757)}{0.0827}\ educ_i \qquad (29)$$
$$n = 526,\ R^2 = 0.19$$



The Selection Problem

Do hospitals make people healthier?

The National Health Interview Survey (NHIS) contains the information needed to make this comparison. Specifically, it includes a question "During the past 12 months, was the respondent a patient in a hospital overnight?" which we can use to identify recent hospital visitors. The NHIS also asks "Would you say your health in general is excellent, very good, good, fair, poor?" The following table displays the mean health status (assigning a 1 to excellent health and a 5 to poor health) among those who have been hospitalized and those who have not (tabulated from the 2005 NHIS):

Selection Problem

Group	Sample Size	Mean health status	Std. Error
Hospital	7774	2.79	0.014
No Hospital	90049	2.07	0.003

- Taken at face value, this result suggests that going to the hospital makes people sicker. It's not impossible this is the right answer: hospitals are full of other sick people who might infect us, and dangerous machines and chemicals that might hurt us.
- Still, it's easy to see why this comparison should not be taken at face value: people who go to the hospital are probably less healthy to begin with.
- Moreover, even after hospitalization people who have sought medical care are not as healthy, on average, as those who never get hospitalized in the first place, though they may well be better than they otherwise would have been.

Counterfactuals

- To describe this problem more precisely, think about hospital treatment as described by a binary random variable, $D_i = \{0, 1\}$. The outcome of interest, a measure of health status, is denoted by Y_i . The question is whether Y_i is affected by hospital care.

$$Y_i = \begin{cases} Y_{1i} & \text{if } D_i = 1 \\ Y_{0i} & \text{if } D_i = 0 \end{cases}$$

  The observed outcome, Y_i , can be written in terms of potential outcomes as

$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i})D_i$$

Selection Bias

$$\underbrace{E[Y_i|D_i = 1] - E[Y_i|D_i = 0]}_{\text{Observed difference in avg. health}} = \underbrace{E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1]}_{\text{avg. treatment effect on the treated}}$$
$$+ \underbrace{E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]}_{\text{selection bias}}$$

The term

$$E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1] = E[Y_{1i} - Y_{0i}|D_i = 1]$$

is the *average causal effect of hospitalization on those who were hospitalized*. This term captures the average difference between the health of the hospitalized, $E[Y_{1i}|D_i = 1]$; and what would have happened to them had they not been hospitalized, $E[Y_{0i}|D_i = 1]$: The observed difference in health status however, adds to this causal effect a term called **selection bias**.

Random Assignment Solves the Selection Problem

Random assignment of D_i solves the selection problem because random assignment makes D_i independent of potential outcomes. To see this, note that

$$\begin{aligned} E[Y_i|D_i = 1] - E[Y_i|D_i = 0] &= E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0] \\ &= E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1] \end{aligned}$$

where the independence of Y_{0i} and D_i allows us to swap $E[Y_{0i}|D_i = 1]$ for $E[Y_{0i}|D_i = 0]$ in the second line. In fact, given random assignment, this simplifies further to

$$\begin{aligned} E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1] &= E[Y_{1i} - Y_{0i}|D_i = 1] \\ &= E[Y_{1i} - Y_{0i}] \end{aligned}$$

Random Assignment

The effect of randomly-assigned hospitalization on the hospitalized is the same as the effect of hospitalization on a randomly chosen patient. The main thing, however, is that random assignment of D_i eliminates selection bias. This does not mean that randomized trials are problem-free, but in principle they solve the most important problem that arises in empirical research.

Appendix I, Summation Math

The $\bar{x}$ notation

We defined $\bar{x}$ as the (empirical) average of x:

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

It therefore could be useful to denote $\overline{\cdots}$ as the average of "anything", where $\cdots$ could be more complex than just one variable:

$$\overline{\cdots} = \frac{1}{n} \cdot \sum_{i=1}^{n} \cdots_i$$

e.g. the average of $1/y$ is simply:

$$\overline{y^{-1}} = \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{1}{y_i}$$

Variables versus constants

Generally :

$$\overline{x+y} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i + y_i = \bar{x} + \bar{y}$$

and

$$\overline{x \cdot y} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \cdot y_i \stackrel{?}{=} \bar{x} \cdot \bar{y}$$

Often the context makes it clear, what letters are varying (“with index i”) and what are constant without an index i:

$$\overline{x+c} = \frac{1}{n} \cdot \sum_{i=1}^{n} (x+c)_i = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i + c = \bar{x} + c$$

Exercise

Average of a linear transformation of x ($m \cdot x + b$) in $\overline{\cdots}$ notation

$$\overline{m \cdot x + b} = \frac{1}{n} \cdot \sum_{i=1}^{n} (mx + b)_i = \frac{1}{n} \cdot \sum_{i=1}^{n} (mx_i + b) = m\bar{x} + b$$

Exercise:

Average of x squared in $\overline{\cdots}$ notation

$$\overline{x^2} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i^2$$

Exercise: variance von x in $\overline{\cdots}$ notation

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

Exercise: Binomial formula for the variance:

$$(a-b)^2 = a^2 - 2ab + b^2 \Leftrightarrow$$

$$s = \overline{(x-\bar{x})^2} = \overline{x^2 - 2x\bar{x} + \bar{x}^2} = \overline{x^2} - \overline{2x\bar{x}} + \bar{x}^2 = ??$$

$$s_x^2 = \overline{x^2} - 2\bar{x}\bar{x} + \bar{x}^2 = \overline{x^2} - \bar{x}^2$$

Exercise: $\overline{\cdots}$ notation for the covariance:

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

$$s_{xy} = \overline{(x - \bar{x}) \cdot (y - \bar{y})}$$

Exercise: Alternative $\overline{\cdots}$ notation for the covariance:

$$COV = s_{xy} = \overline{(x-\bar{x})\cdot(y-\bar{y})} = \overline{xy - x\bar{y} - \bar{x}y + \bar{y}\bar{x}} = ??$$

$$s_{xy} = \overline{x \cdot y} - \bar{y} \cdot \bar{x}$$

From covariance to the correlation coefficient

$$r = \frac{s_{xy}}{s_x s_y} = \frac{\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2}\sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2}} = \frac{\frac{1}{n} \ldots}{\sqrt{\frac{1}{n} \ldots}\sqrt{\frac{1}{n} \ldots}}$$

Summary

$$\text{variance: } s_x^2 = \overline{(x-\bar{x})^2} = \overline{x^2} - \bar{x}^2$$

$$\text{covariance: } s_{xy} = \overline{x \cdot y} - \bar{y} \cdot \bar{x}$$

Therefore, **if** $\bar{x} = 0$ **or** $\bar{y} = 0$

$$\Rightarrow s_{xy} = \overline{x \cdot y}$$

$$\Rightarrow s_{xy} = \overline{(x-\bar{x}) \cdot y} = \overline{x \cdot (y-\bar{y})}$$