Econometrics 1
Lecture 2: Simple Linear Regression
ENSAE 2014/2015

Michael Visser (CREST-ENSAE)

Definition of SLR model

Suppose we wish to investigate the relation between

- An *outcome* variable y
- One *explanatory* variable x

The Simple Linear Regression (SLR) model assumes that the following is true *in the population*

Linear in parameters

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

Remarks:

- β_0 (the constant) and β_1 (slope parameter) are unknown parameters to be estimated
- Other *unobserved* factors determining y are captured by the error term u.
- y, x, u are all viewed as random variables.
- y is called the dependent variable, or explained variable, or regressand.

Definition of SLR model (cntd)

- x is the explanatory variable, or independent variable, or regressor, or control variable, or covariate ...
- The model reflects a *population* relationship. A sample from the population will be used to learn (i.e. estimate) something about the parameter values in the population.
- The model is linear in the parameters. $y = \beta_0 + \beta_1 x^2 + u$ or $y = \beta_0 + \beta_1 \sqrt{x} + u$ are linear models as well.
- Interpretation of parameters: β_0 is the value of y when x and u equal zero. β_1 is the causal effect of a one-unit change in x on y, i.e., it measures the marginal effect of x on y keeping u constant: $\triangle y = \beta_1$ if $\triangle x = 1$ and $\triangle u = 0$.

Wage/education example

In the model

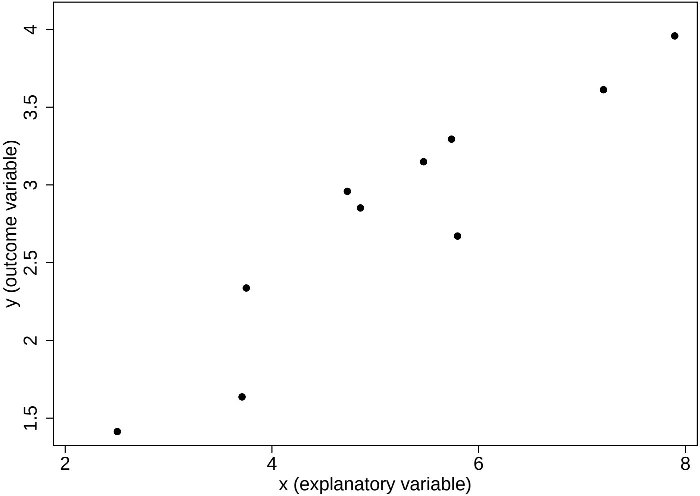
$$wage = \beta_0 + \beta_1 educ + u$$

- The intercept β_0 is the value of *wage* when *educ* and u are 0.
- The slope parameter β_1 measures how much *wage* changes for a one-unit change in *educ*, keeping u constant, i.e., fixing all other factors that might influence wages (ability, experience, gender, etc.).

Remarks:

- We are usually not interested in the intercept β_0.
- The linearity implies that a one-unit change in *educ* has the same effect on *wage* regardless of the initial value of *educ*. This is often unrealistic (not conform the model) and we will see how to tackle this later.
- This "keeping constant" is important: We will need to make assumptions about u and how it relates to *educ* in order to be able to reliably estimate β_0 and β_1.

Data on two variables



Data

In what follows we will assume that we have a sample from the population which satisfies the following condition

Random Sampling

$\{(x_i, y_i) : i = 1, \ldots, n\}$, where (x_i, y_i) are *independent and identically distributed (i.i.d.)*

This means that

- the pair (y_i, x_i) is independent of (y_j, x_j) for $i \neq j$
- each observation is a random draw from the same probability distribution

We will relax this assumption later in the course.

Main model assumption

Since there is a constant in $y = \beta_0 + \beta_1 x + u$, we can assume without loss of generality that

$$E[u] = 0.$$

More important is the following assumption:

Zero Conditional Mean

$$E[u|x] = E[u]\,(= 0).$$

This states that the unobservables u are (mean) independent of x.

The mean-independence assumption is stronger than the assumption that u and x are not correlated, i.e., $E[ux] = 0$.

What does $E[u|x] \neq 0$ mean?

It is easiest to think about this in the context of an example:

$$wage = \beta_0 + \beta_1 educ + u$$

we need to consider what factors will be captured by u. $E[u|x] = 0$ implies for example

$$E[u|\text{high school drop-out}] = E[u|\text{college graduate}]$$

where u can be

- ambition
- intelligence
- local labor market conditions (pay-off to education)
- health
- etc. etc.

So $E[u|x] \neq 0$ may mean, for instance, that the average ambition or intelligence of a person varies by level of education.

Example: treatment effect model

Let us consider again Rubin's counter-factual framework (see lecture 1, last week).

Assume that the treatment effect is the same for everyone: $Y_{1i} - Y_{0i} = \rho$ for all i. The equation $Y_i = Y_{0i} + D_i(Y_{1i} - Y_{0i})$ can be rewritten (check this) as the following SLR model

$$Y_i = \alpha + \rho D_i + \eta_i$$

with $\alpha = E[Y_{0i}]$ and $\eta_i = Y_{0i} - E[Y_{0i}]$. Now since

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

we see that the error term η_i is mean-independent of the regressor D_i if there is no selection bias.

Alternative interpretation under $E[u|x] = 0$

Under $E[u|x] = E[u] = 0$ we can take the expected value of y given x:

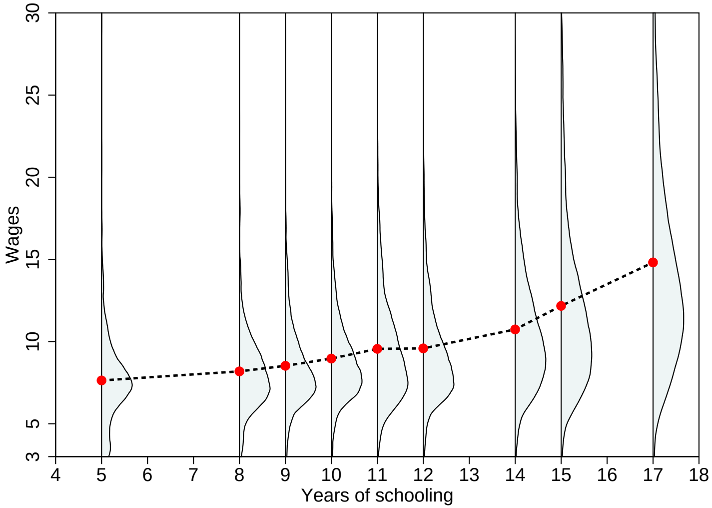
$$E[y|x] = \beta_0 + \beta_1 x$$

This gives us the following interpretation of the slope parameter:

- The *expected value* of y changes with β_1 for a one-unit change in x

The following graph plots the net hourly wage rate for different levels of schooling (defined as number of years of schooling beyond the age of 6), using the Enquête Emploi 2007. The dots represent the empirical means of hourly wage given the schooling level. The curved lines represent the empirical density functions of hourly wage.

Wages and education (Enquête Emploi, 2007)

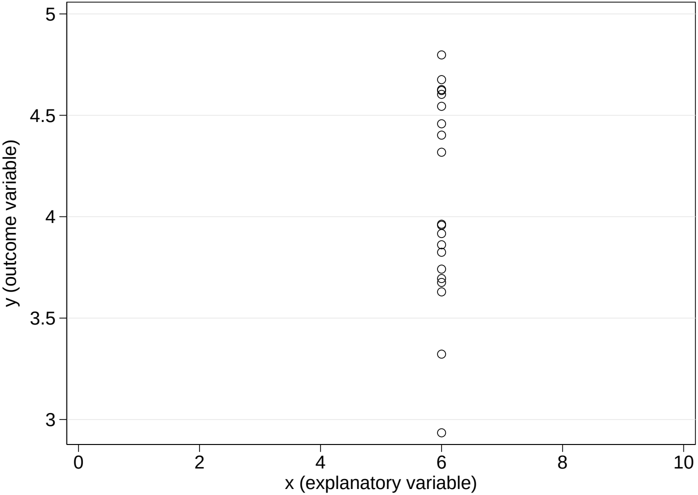


Summary of assumptions

- SLR.1 Linear In Parameters: $y = \beta_0 + \beta_1 x + u$
- SLR.2 Random Sample: $\{(x_i, y_i) : i = 1, \ldots, n\}$, where $\{x_i, y_i\}$ are *i.i.d.*
- SLR.3 Zero Conditional Mean: $E[u|x] = 0$
- SLR.4 Sample Variation in x: $x_i \neq c$

Only the last assumption has not been discussed yet. Intuitively, it is clear that we cannot identify and estimate the effect of x on y if there is no variation in the regressor in the sample. Sample data with no variation in the explanatory variable look like those depicted in the next graph:

No sample variation in x (SLR.4 violated)



Using assumptions about u

$E[u|x] = 0$ (SLR.3) implies two restrictions on the joint probability distribution of (x, y) in the population:

$$
\begin{aligned}
E[u] &= E[y - \beta_0 - \beta_1 x] = 0 \\
E[xu] &= E[x(y - \beta_0 - \beta_1 x)] = 0
\end{aligned}
$$

where the first restriction follows from the fact that $E[u] = E[E[u|x]] = 0$, and the second restriction from $E[xu] = E[E[xu|x]] = E[xE[u|x]] = 0$. We can use the sample analogues of these two moment conditions to estimate our parameters:

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) &= 0 \\
\frac{1}{n}\sum_{i=1}^{n}x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) &= 0
\end{aligned}
$$

This is a system of 2 equations and 2 unknowns.

Estimators

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = \overline{y} - \hat{\beta}_0 - \hat{\beta}_1\overline{x} = 0$$

and therefore we have an estimate of the intercept

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x}$$

Substitute in the second moment condition

$$\frac{1}{n}\sum_{i=1}^{n}x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = \frac{1}{n}\sum_{i=1}^{n}x_i(y_i - (\overline{y} - \hat{\beta}_1\overline{x}) - \hat{\beta}_1 x_i) = 0$$

the right-hand side can be rearranged

$$\sum_{i=1}^{n}x_i(y_i - \overline{y}) = \hat{\beta}_1\sum_{i=1}^{n}x_i(x_i - \overline{x})$$

and the estimated slope is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}x_i(y_i - \overline{y})}{\sum_{i=1}^{n}x_i(x_i - \overline{x})} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Ordinary Least Squares

We arrived at our estimators using $E[ux] = E[u] = 0$. Alternatively we can choose $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the sum of squared residuals

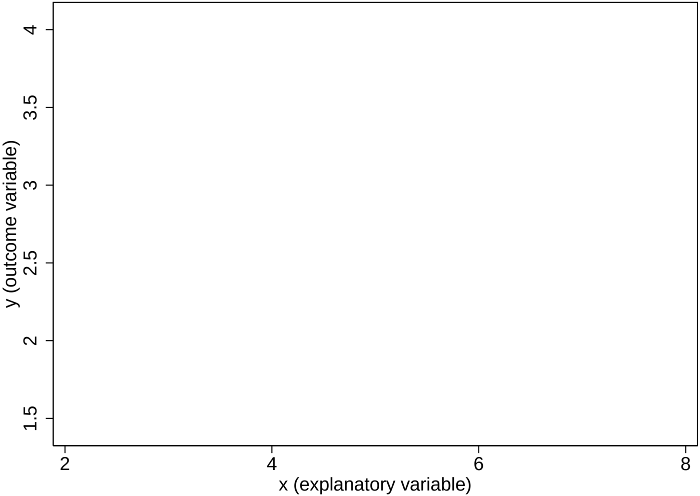
$$\sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Minimizing this criterion with respect to the two parameters gives the same expressions for $\hat{\beta}_0$ and $\hat{\beta}_1$ as above.
Once we have our estimates we can calculate the

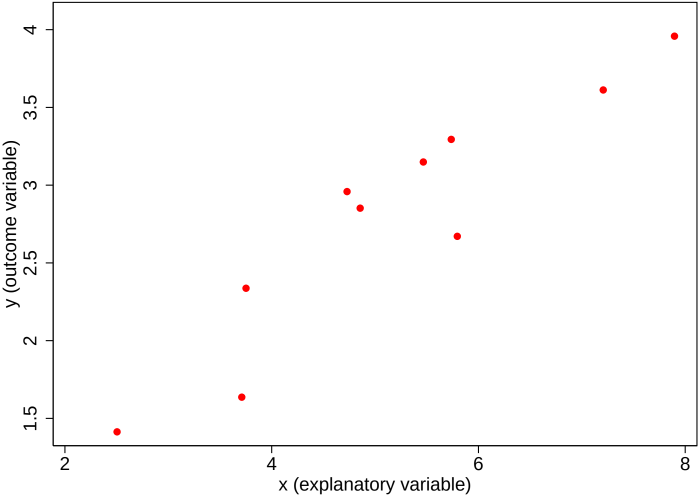
- Predicted/fitted value: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
- Residual: $\hat{u}_i = y_i - \hat{y}_i$
- OLS regression line: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

We can decompose y_i into an explained part and a residual
$y_i = \hat{y}_i + \hat{u}_i$

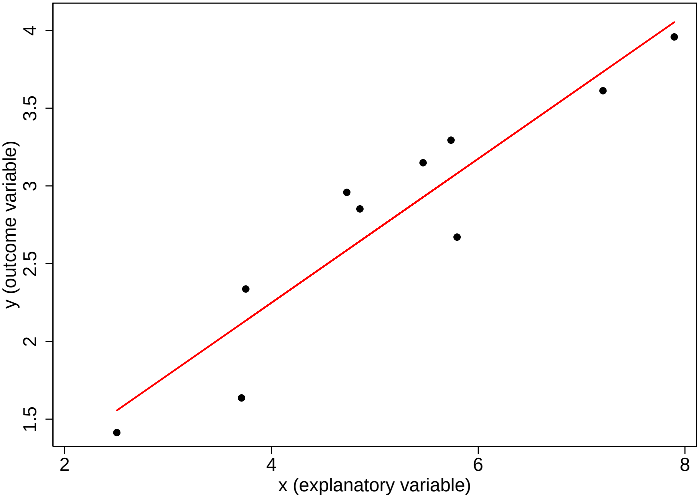
OLS



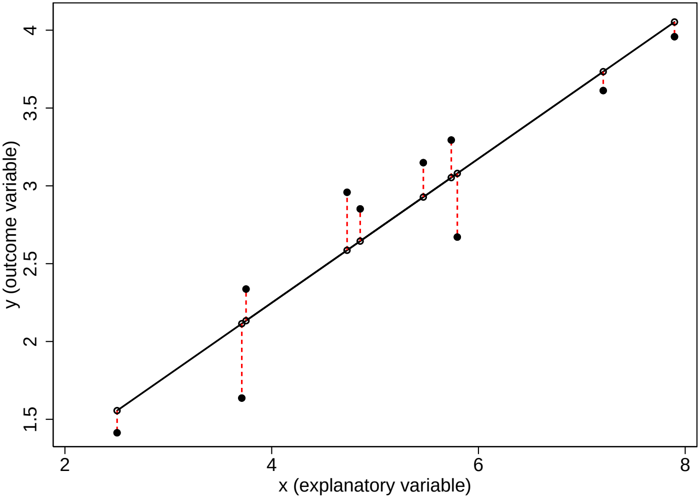
OLS: data points



OLS: regression line



OLS: fitted values and residuals



Goodness of fit

Similarly we can decompose the total sum of squares (SST) into the explained sum of squares (SSE) and the residual sum of squares (SSR):

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

The zero term follows from the first order conditions (the two conditions that determine the OLS estimators).

SST divided by n can be seen as the sample variance of the dependent variable y.

Goodness of fit (cntd)

Since the average of the $\hat{u}$ in the sample is zero, SSR divided by n can be seen as the sample variance of the residuals.

Since the average of the $\hat{y}$ in the sample equals the sample mean $\bar{y}$, SSE divided by n can be seen as the sample variance of the fitted values.

We can rewrite

$$SST = SSR + SSE$$

as follows

$$1 = \frac{SSE}{SST} + \frac{SSR}{SST}$$

The so-called R^2 is defined as

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

Goodness of fit (cntd)

R^2 is the fraction of the sample variation in y that is explained by x

The R-squared is always between zero and one. It equals 1 if all y_i are on the regression line (x perfectly explains y), and 0 if $\hat{\beta}_1 = 0$ (x has no effect at all on y).

A low R-squared is quite common, especially in cross sections. It means that there are a lot of other variables (besides the included regressor x) that explain y. The estimators can nevertheless have good properties (unbiased estimators of the causal parameters).

If we only care about prediction, it is important though to have a model with a high R-squared.

Unit of measurement

What happens when we change the unit of measurement of y: $y' = c \cdot y$?

$$\begin{aligned} y' &= \alpha_0 + \alpha_1 x + u' \\ &= c\beta_0 + c\beta_1 x + cu \end{aligned}$$

Similarly, what happens when we change the unit of measurement of x: $x' = c \cdot x$?

$$\begin{aligned} y &= \alpha_0 + \alpha_1 x' + u' \\ &= \alpha_0 + \alpha_1 cx + u' \\ &= \beta_0 + \beta_1 x + u \end{aligned}$$

So multiplying the dependent variable by a value c leads to a model where the parameters are also multiplied by c. Multiplying the regressor x by c leaves the constant unchanged and the slope parameter is divided by c.

Functional Form

In the SLR model studied so far the effect of x does *not* depend on the initial value of x:

$$\frac{\Delta y}{\Delta x} = \beta_1$$

Another important case is when the relative effect is constant (and independent of x):

$$\frac{\Delta y / y}{\Delta x} = \beta_1$$

To account for this possibility we can consider the model

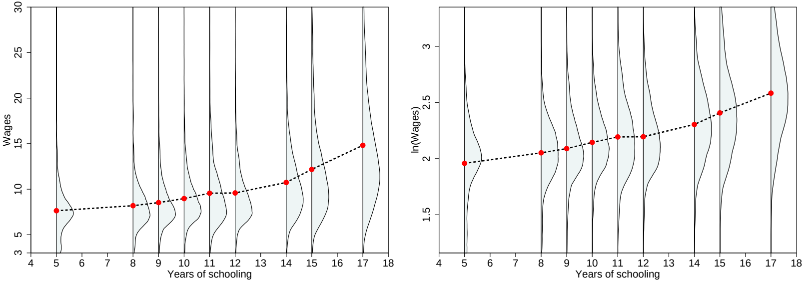
$$\log y = \beta_0 + \beta_1 x + u$$

since

$$\frac{\triangle \log y}{\triangle x} = \frac{\triangle y}{\triangle x} \frac{\triangle log y}{\triangle y} = \frac{\triangle y / y}{\triangle x} = \beta_1$$

The next graph shows that a logarithmic transformation of the dependent variable may make the linearity assumption more credible.

Log transformation of wage (Enquête Emploi, 2007)



Summary of functional forms with logs

Model	Dep. Var.	Indep. Var.	Interpretation
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	$\log y$	$\log x$	$\%\Delta y = \beta_1 \%\Delta x$

We call β_1 in the

- log-level model: semi-elasticity
- log-log model: elasticity

Unbiasedness of OLS estimators

The slope estimate

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \overset{SLR.1}{=} \frac{\sum_{i=1}^n (x_i - \bar{x})(\beta_0 + \beta_1 x_i + u_i)}{s_x^2} \\
&= \beta_0 \frac{\sum_{i=1}^n (x_i - \bar{x})}{s_x^2} + \beta_1 \frac{\sum_{i=1}^n (x_i - \bar{x}) x_i}{s_x^2} + \frac{\sum_{i=1}^n (x_i - \bar{x}) u_i}{s_x^2} \\
&= \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x}) u_i}{s_x^2}
\end{aligned}
$$

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

Unbiasedness of OLS estimators (cntd)

The slope estimate

Now

$$
\begin{aligned}
E[\hat{\beta}_1 | x_1, \dots, x_n] &= E[\beta_1 + (1/s_x^2) \sum_{i=1}^{n} (x_i - \bar{x}) u_i | x_1, \dots, x_n] \\
&= \beta_1 + (1/s_x^2) \sum_{i=1}^{n} (x_i - \bar{x}) E[u_i | x_1, \dots, x_n] \\
&\overset{SLR.2}{=} \beta_1 + (1/s_x^2) \sum_{i=1}^{n} (x_i - \bar{x}) E[u_i | x_i] \overset{SLR.3}{=} \beta_1
\end{aligned}
$$

and unbiasedness follows directly from the law of total expectations

$$
E[\hat{\beta}_1] = E[E[\hat{\beta}_1 | x_1, \dots, x_n]] = E[\beta_1] = \beta_1
$$

Unbiasedness of OLS estimators (cntd)

The intercept estimate

Similarly

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

so

$$\begin{aligned}E[\hat{\beta}_0|x_1,\dots,x_n] &= \beta_0 + E[(\beta_1 - \hat{\beta}_1)\bar{x}|x_1,\dots,x_n] + E[\bar{u}|x_1,\dots,x_n] \\ &= \beta_0 + E[\beta_1 - \hat{\beta}_1|x_1,\dots,x_n]\bar{x} + E[\bar{u}|x_1,\dots,x_n] \\ &\overset{SLR.2}{=} \beta_0 + E[\beta_1 - \hat{\beta}_1|x_1,\dots,x_n]\bar{x} + \frac{1}{n}\sum_{i=1}^{n} E[u_i|x_i] \\ &\overset{SLR.3}{=} \beta_0 + E[\beta_1 - \hat{\beta}_1|x_1,\dots,x_n]\bar{x} \\ &= \beta_0\end{aligned}$$

where the last equality follows from the unbiasedness of $\hat{\beta}_1$.
Note that the above implies that $E[\hat{\beta}_0] = \beta_0$

Unbiasedness of OLS estimators (cntd)

Remarks:

- Unbiasedness fails if any one of assumptions SLR.1-4 fail: in any application we should ask ourselves whether they are valid.
- To keep notations simple, $\hat{\beta}_0$ and $\hat{\beta}_1$ stand for two different things: *estimators* (which are random variables) and *estimates* (realizations of the estimators, i.e., non-random variables).
- Unbiasedness is a feature of the *sampling* distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$. It does not mean that the OLS estimates obtained using a particular sample are equal to the true values of β_0 and β_1. Rather, if we could *indefinitely* draw random samples (each of size *n*) from the population, calculate the OLS estimates each time, and average these estimates over all samples, we would find β_0 and β_1. This is of course a thought experiment since in practice one typically has only one sample!

The last remark can be illustrated with a simple simulation study.

Simulation study

- Consider the regression model $y = 1 + 2x + \epsilon$. Assume that $\epsilon \sim N(0, 3)$, $x \sim N(1, 1)$, and x and ϵ independent random variables.
- Since x and ϵ are independent, we have $E(\epsilon|x) = E(\epsilon) = 0$.
- Let us now simulate n independent observations x_i, $i = 1, ..., n$ (statistical packages such as Stata and Gauss have random number generators that can do this job).
- Also generate n independent observations $\epsilon_i^{(1)}$, $i = 1, ..., n$.
- Using that $y_i^{(1)} = 1 + 2x_i + \epsilon_i^{(1)}$, this gives us a random sample of size n, $\{(x_i, y_i^{(1)}) : i = 1, ..., n\}$.
- Calculate the OLS estimates $\hat{\beta}_0^{(1)}$ and $\hat{\beta}_1^{(1)}$. They are reported in the next table (trial 1).
- Repeat this procedure a second time (i.e. simulate again n independent observations $\epsilon_i^{(2)}$, construct the second sample $\{(x_i, y_i^{(2)}) : i = 1, ..., n\}$, calculate the OLS estimates, etc.). The OLS estimates based on the second sample are also reported in the table (trial 2).

Simulation study (cntd)

- Let us consider 10000 repetitions of the procedure.
- Note that the observations x_i, $i = 1, ..., n$, remain fixed in all trials.
- The simulation study is performed for $n = 50$ (relatively small sample) $n = 1000$ and (large sample).
- Reporting the estimates of all trials would take too much space! Instead the table lists the results of the first ten trials, and the average of OLS estimates (over all 10000 trials).

Simulation study (cntd)

Simulating model $y = 1 + 2x + \epsilon$,
for $\epsilon \sim N(0,3)$, $x \sim N(1,1)$

	$\hat{\beta}_0^{(t)}$; $\hat{\beta}_1^{(t)}$	
Trial	n=50	n=1000
$t = 1$	1.505; 2.125	0.836; 2.090
$t = 2$	0.745; 2.157	1.012; 2.018
$t = 3$	1.501; 1.548	0.973; 2.000
$t = 4$	1.287; 1.980	0.876; 2.030
$t = 5$	1.096; 2.133	1.044; 1.964
$t = 6$	1.106; 2.056	1.022; 1.985
$t = 7$	0.700; 2.339	0.969; 1.998
$t = 8$	0.772; 2.142	0.997; 1.994
$t = 9$	0.683 1.970	1.130; 1.962
$t = 10$	1.319 1.909	1.049; 1.940
...	...	...
...	...	...
Average	1.001; 2.001	1.000; 2.000

Simulation study (cntd)

As the table shows, in some samples the estimates are indeed far from true values (e.g. the third trial for $n = 50$, or the first trial for $n = 1000$).

Note that estimates for small samples are generally less precise than those for large samples.

The average estimates are, however, quite accurate in both cases.

Variances

In addition to knowing that the OLS estimators are unbiased, it is of interest to know how far we can expect the estimators to be away from their true values.

The measure of spread in the distribution of the OLS estimators the easiest to work with is the variance.

To derive the sampling variance make the following additional assumption:

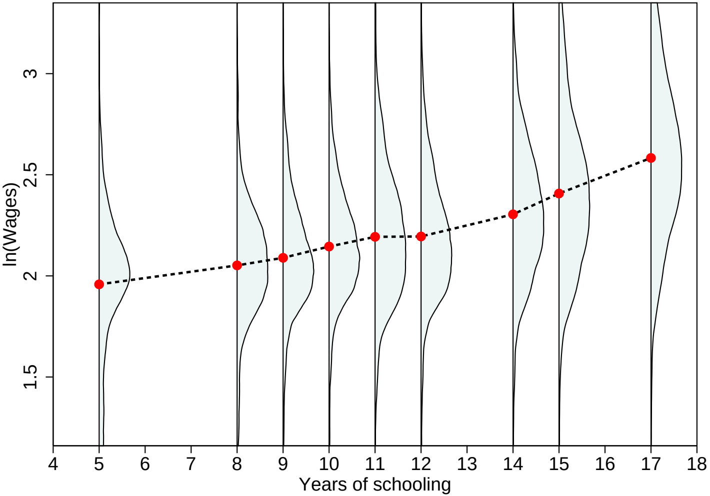
Assumption SLR.5: Homoscedasticity

$Var(u|x) = \sigma^2$

Under SLR.5 we have
$Var(y|x) = Var(\beta_0 + \beta_1 x + u|x) = Var(u|x) = \sigma^2$.

When $Var(u|x)$ depends on x the error term is said to be *heteroscedastic.* In that case $Var(y|x)$ also varies with x.

Example of heteroscedasticity



Sampling Variance of the OLS estimators

Sampling variance of $\hat{\beta}_0$ and $\hat{\beta}_1$

$$\begin{aligned} Var(\hat{\beta}_1|x_1,\ldots,x_n) &= \sigma^2/s_x^2 \\ Var(\hat{\beta}_0|x_1,\ldots,x_n) &= \sigma^2 n^{-1}\sum_{i=1}^n x_i^2/s_x^2 \end{aligned}$$

We only show the first result. Remember that $\hat{\beta}_1 = \beta_1 + (1/s_x^2)\sum_{i=1}^n (x_i - \bar{x})u_i$, therefore

$$\begin{aligned} Var(\hat{\beta}_1|x_1,\ldots,x_n) &= (1/s_x^2)^2 Var(\sum_{i=1}^n (x_i-\bar{x})u_i|x_1,\ldots,x_n) \\ &= (1/s_x^2)^2 \sum_{i=1}^n (x_i-\bar{x})^2 Var(u_i|x_1,\ldots,x_n) \\ &= (1/s_x^2)^2 \sum_{i=1}^n (x_i-\bar{x})^2\sigma^2 = \sigma^2/s_x^2 \end{aligned}$$

Sampling Variance of the OLS estimators (cntd)

Remarks about the variance of $\hat{\beta}_1$ (we are usually only interested in the slope parameter):

- The larger the error variance σ^2 the larger the variance of $\hat{\beta}_1$
- The larger the variance in x the more precisely we can estimate β_1. This means that if we can manipulate x (as in an experiment where x is the treatment) we should spread out x_i as much as possible
- $var(\hat{\beta}_1|x)$ decreases as n increases. This is because $\sum_i (x_i - \bar{x})^2$ increases with n.

Estimating the error variance

The sampling variances depend on σ^2, which is generally unknown. To estimate the sampling variances we first need an estimator of σ^2. First, write the residuals as a function of the errors:

$$\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i = \beta_0 + \beta_1 x_i + u_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

or

$$\hat{u}_i = u_i - (\hat{\beta}_0 - \beta_0) - (\hat{\beta}_1 - \beta_1) x_i$$

Averaging in our sample

$$n^{-1} \sum_{i=1}^{n} \hat{u}_i = 0 = \overline{u} - (\hat{\beta}_0 - \beta_0) - (\hat{\beta}_1 - \beta_1)\overline{x}$$

so that

$$\hat{u}_i = (u_i - \overline{u}) - (\hat{\beta}_1 - \beta_1)(x_i - \overline{x})$$

Estimating the error variance (cntd)

Using the last expression we get

$$
\begin{aligned}
\sum_{i=1}^{n} \hat{u}_i^2 &= \sum_{i=1}^{n} \Big((u_i - \overline{u}) - (\hat{\beta}_1 - \beta_1)(x_i - \overline{x}) \Big)^2 \\
&= \sum_{i=1}^{n} (u_i - \overline{u})^2 + (\hat{\beta}_1 - \beta_1)^2 \sum_{i=1}^{n} (x_i - \overline{x})^2 \\
&\quad -2(\hat{\beta}_1 - \beta_1) \sum_{i=1}^{n} (x_i - \overline{x})(u_i - \overline{u})
\end{aligned}
$$

Now
$E[\sum_{i=1}^{n} (u_i - \overline{u})^2 | x_1, \ldots, x_n] = (n-1)\sigma^2$
$E[\sum_{i=1}^{n} (\hat{\beta}_1 - \beta_1)^2 (x_i - \overline{x})^2 | x_1, \ldots, x_n] = s_x^2 Var(\hat{\beta}_1 | x_1, \ldots, x_n) = \sigma^2$

Estimating the error variance (cntd)

Remember that

$$\hat{\beta}_1 = \beta_1 + \frac{1}{s_x^2}\sum_{i=1}^{n}(x_i - \bar{x})u_i$$

We therefore have

$$
\begin{aligned}
&E[-2(\hat{\beta}_1 - \beta_1)\sum_{i=1}^{n}(x_i - \overline{x})(u_i - \overline{u})|x_1, \ldots, x_n] \\
&\qquad = E[-2(\hat{\beta}_1 - \beta_1)\sum_{i=1}^{n}(x_i - \overline{x})u_i|x_1, \ldots, x_n] \\
&\qquad = E[-2(\hat{\beta}_1 - \beta_1)^2 s_x^2|x_1, \ldots, x_n] \\
&\qquad = -2s_x^2 Var(\hat{\beta}_1|x_1, \ldots, x_n) = -2\sigma^2
\end{aligned}
$$

Estimating the error variance (cntd)

We can substitute these results to obtain the expectation of the SSR:

$$E[\sum_{i=1}^{n} \hat{u}_i^2 | x_1, \ldots, x_n] = (n-1)\sigma^2 + \sigma^2 - 2\sigma^2 = (n-2)\sigma^2$$

which implies that $E[SSR/(n-2)] = \sigma^2$, i.e., $\hat{\sigma}^2 \equiv SSR/(n-2)$ is an unbiased estimator of σ.
If $\hat{\sigma}^2$ is plugged into the variance formulas, we obtain unbiased estimators of $Var(\hat{\beta}_0 | x_1, \ldots, x_n)$ and $Var(\hat{\beta}_1 | x_1, \ldots, x_n)$:

$$\begin{aligned} \widehat{Var}(\hat{\beta}_1 | x_1, \ldots, x_n) &= \hat{\sigma}^2 / s_x^2 \\ \widehat{Var}(\hat{\beta}_0 | x_1, \ldots, x_n) &= \hat{\sigma}^2 n^{-1} \sum_{i=1}^{n} x_i^2 / s_x^2 \end{aligned}$$

Estimating the error variance (cntd)

The conditional standard deviations of the estimators, i.e., $\sqrt{Var(\hat{\beta}_0|x_1,\ldots,x_n)}$ and $\sqrt{Var(\hat{\beta}_1|x_1,\ldots,x_n)}$, can be estimated by

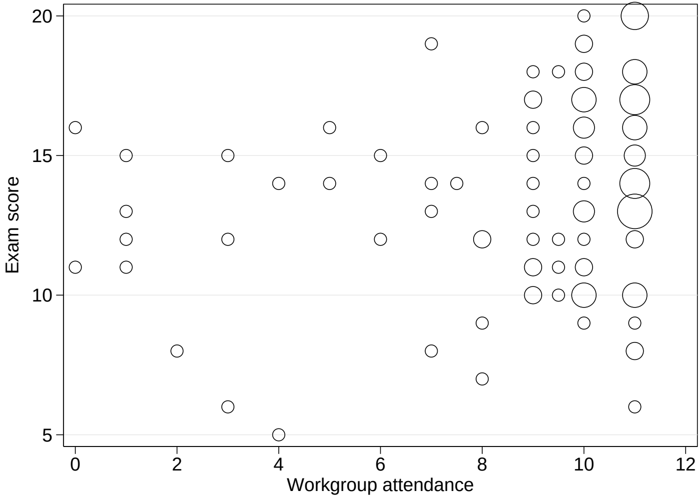
$$
\begin{aligned}
se(\hat{\beta}_1|x_1,\ldots,x_n) &= \hat{\sigma}/\sqrt{s_x^2} \\
se(\hat{\beta}_0|x_1,\ldots,x_n) &= \hat{\sigma}\sqrt{n^{-1}\sum_{i=1}^{n} x_i^2/s_x^2}
\end{aligned}
$$

These last expressions are called the standard errors of the OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$.

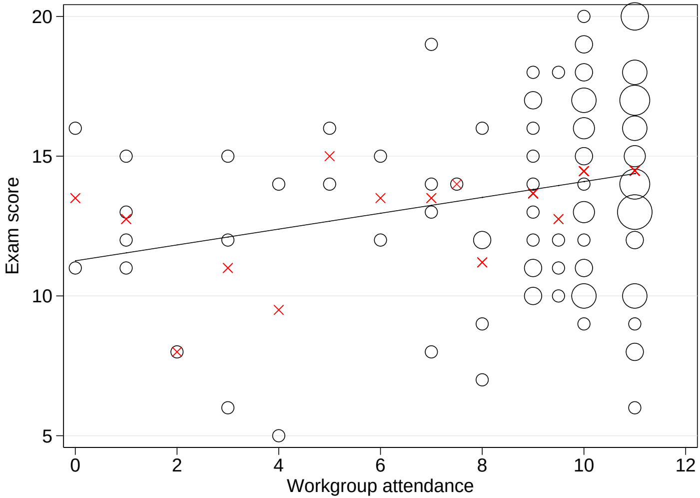
Workgroup attendance and performance - ENSAE

```
. summarize exam presence

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
   examscore |       114    13.82456    3.451934          5         20
    presence |       119    8.945378    2.974318          0         11

. tabulate presence

  Workgroup |
 attendance |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |          3        2.52        2.52
          1 |          4        3.36        5.88
          2 |          2        1.68        7.56
          3 |          3        2.52       10.08
          4 |          2        1.68       11.76
          5 |          2        1.68       13.45
          6 |          2        1.68       15.13
          7 |          4        3.36       18.49
        7.5 |          1        0.84       19.33
          8 |          5        4.20       23.53
          9 |         13       10.92       34.45
        9.5 |          4        3.36       37.82
         10 |         27       22.69       60.50
         11 |         47       39.50      100.00
------------+-----------------------------------
      Total |        119      100.00
```

Attendance & performance - ENSAE (contd)



Attendance & performance - ENSAE (contd)



Attendance & performance - ENSAE (contd)

```
. reg exam presence

      Source |       SS       df       MS              Number of obs =     114
-------------+------------------------------           F(  1,   112) =    6.42
       Model |  73.0333985     1  73.0333985           Prob > F      =  0.0126
    Residual |  1273.45783   112  11.3701592           R-squared     =  0.0542
-------------+------------------------------           Adj R-squared =  0.0458
       Total |  1346.49123   113  11.9158516           Root MSE      =   3.372

------------------------------------------------------------------------------
   examscore |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    presence |   .2834703   .1118485     2.53   0.013     .0618569    .5050838
       _cons |   11.25717   1.061101    10.61   0.000      9.15473     13.3596
------------------------------------------------------------------------------
```

I understand/can apply..

- Simple linear regression
  - calculate $\hat{\beta}_0, \hat{\beta}_1$ and their s.e.'s, $\hat{\sigma}^2$, R^2, $\hat{u}_i$
  - interpret estimated parameters
- Specification and fit
- Unbiasedness (key assumption: $E[u|x] = 0$)
- Finite sample properties (i.e. true for any n) under homoscedasticity
- Understand and use Stata's regression output