Lecture 3:
Simple Regression II

Jan Zouhar

Introductory Econometrics

Algebraic Properties of OLS Statistics

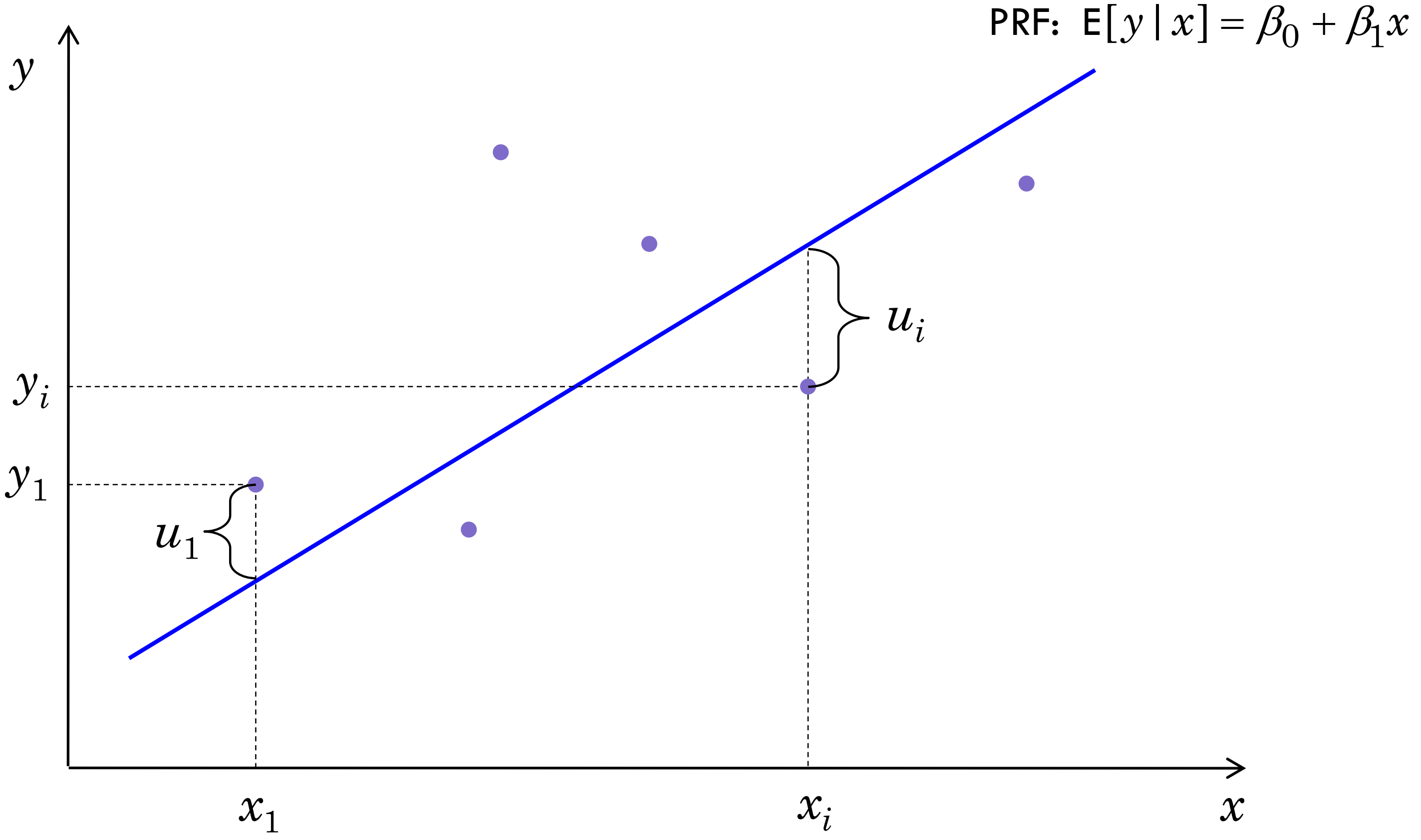
Population vs. sample regression function.

Residuals and their properties.

Goodness of fit.

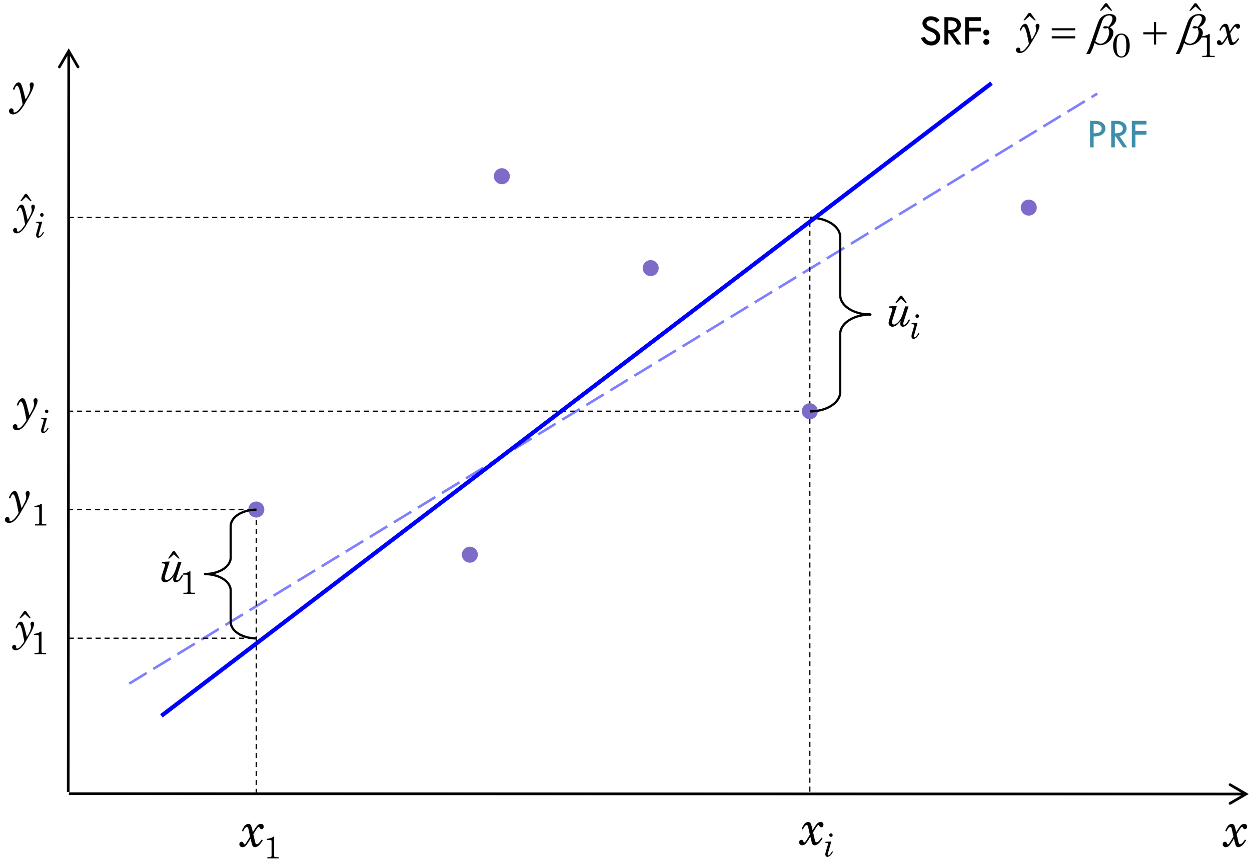
Population Vs. Sample Regression Function

- population regression function (PRF):



Population Vs. Sample Regression Function (cont'd)

- sample regression function (SRF):



Goodness of Fit

- we want to say something about how well the model fits our data (the goal is to end up with a single number, ideally expressed as a percentage)
- we will make use of the following three things:
  - **total sum of squares** (SST)

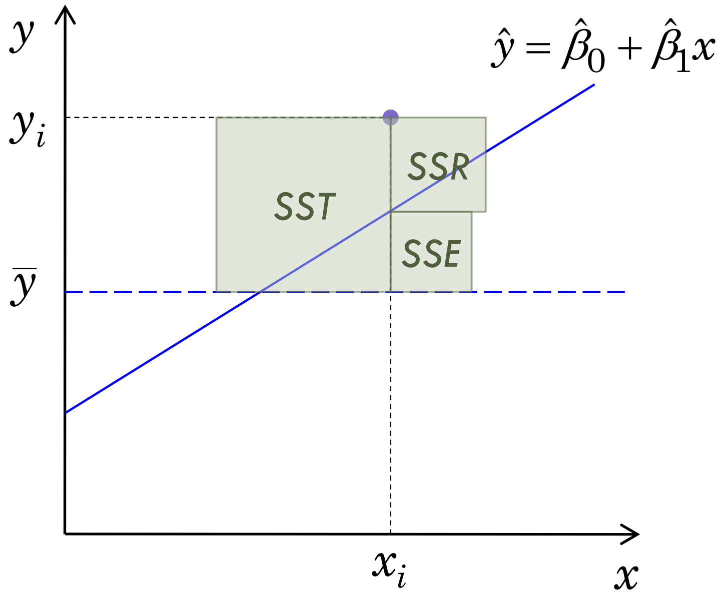
    $$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

  - **explained sum of squares** (SSE)

    $$SSE = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

  - **residual sum of squares** (SSR)

    $$SSR = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\hat{u}_i^2$$



Goodness of Fit (cont'd)

- important algebraic identity: $SST = SSR + SSE$ (we'll prove this later)
- this gives us a really nice way of describing the goodness of fit of the model
    - **R-squared** of the regression (or the **coefficient of determination**):

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

- properties of R^2:
    - $0 \leq R^2 \leq 1$
    - $R^2 = 1$ only if $SSR = 0$, which means that all residuals are zero, and all observations lie *exactly* on the regression line
    - $R^2 = 0$ only if $SSE = 0$, which implies that $\hat{\beta}_1 = 0,\ \hat{\beta}_0 = \bar{y}$

Interpretation of R-squared:

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

Goodness of Fit (cont'd)

Proof of the identity $SST = SSR + SSE$

- first remember that we know something about the residuals (see previous lecture): $\sum_{i=1}^{n} \hat{u}_i = 0$

$$\sum_{i=1}^{n} x_i \hat{u}_i = 0$$

- it follows from these properties that $\sum \hat{u}_i \hat{y}_i = 0$ and $\sum \hat{u}_i (\hat{y}_i - \bar{y}) = 0$
  - e.g., $\sum \hat{u}_i \hat{y}_i = \sum \hat{u}_i (\hat{\beta}_0 + \hat{\beta}_1 x_i) = \hat{\beta}_0 \underset{0}{\sum \hat{u}_i} + \hat{\beta}_1 \underbrace{\sum x_i \hat{u}_i}_{0} = 0$

- now we'll use this to show $SST = SSR + SSE$

$$\begin{aligned} \sum (y_i - \bar{y})^2 &= \sum (\overbrace{y_i - \hat{y}_i}^{\hat{u}_i} + \hat{y}_i - \bar{y})^2 = \\ &= \sum [\hat{u}_i + (\hat{y}_i - \bar{y})]^2 = \\ &= \underset{SSR}{\sum \hat{u}_i^2} + 2 \underbrace{\sum \hat{u}_i (\hat{y}_i - \bar{y})}_{0} + \underbrace{\sum (\hat{y}_i - \bar{y})^2}_{SSE} = \\ &= SSR + SSE \end{aligned}$$

Units and Functional Form

Changing units of measurement.

Functional form of regression models.

Changing the Units of Measurement

- in the CEO example, we ended up with the following equation:

$$\widehat{salary} = 963.191 + 18.501\, roe$$

- it's crucial to know the units of measurement in order to interpret the equation
- it's good to know that if we change the units of measurement, the estimated coefficients change in a completely natural way
- if we regress $salardol = 1{,}000\, salary$ on roe (which means we express CEOs' salary in dollars), we obtain

$$\widehat{salardol} = 963{,}191 + 18{,}501\, roe$$

- if we now express roe in decimals rather than percentage points, defining $roedec = 0.01\, roe$, we get

$$\widehat{salardol} = 963{,}191 + 1{,}850{,}100\, roedec,$$

because $18{,}501\, roe = 1{,}850{,}100\, roedec$

- note that the interpretation of both slope and intercept remains the same in all cases

Functional Form

- so far, we have only dealt with a linear relationship between x and y
- this is really not as strong an assumption as you might think because we can pick x and y to be whatever we want
- as we've seen, changing the units doesn't change anything; however, we can pick a non-linear unit transform
  - **example**:

$$\mathrm{E}[\ \underbrace{\log(wage)}\ \mid\ \underbrace{educ}\ ] = \beta_0 + \beta_1 \underbrace{educ}$$

$$\mathrm{E}[\quad y \quad \mid \quad x \quad] = \beta_0 + \beta_1 \quad x$$

→ this is still considered to be a linear regression model; the word *linear* actually means *linear in parameters*

Linear in parameters	Non-linear in parameters
$y = \beta_0 + \beta_1 x + \beta_2 x^2$	$y = \beta_0 + x^{\beta_1}$
$\log y = \beta_0 + \beta_1 \log x$	$y = \dfrac{\beta_0}{\beta_1 + x}$

Functional Form (cont'd)

- which one of the following types of relationships seems more plausible:
  - with each additional year of education, a person's monthly wage increases by €50
  - with each additional year of education, a person's monthly wage increases by 5%
- "5% each year" means:
  - if we denote $\mathrm{E}[wage \mid educ = 0]$ as w, then

$$\mathrm{E}[wage \mid educ = 1] = w \times 1.05$$

$$\mathrm{E}[wage \mid educ = 2] = w \times 1.05^2$$

$$\mathrm{E}[wage \mid educ = 3] = w \times 1.05^3$$

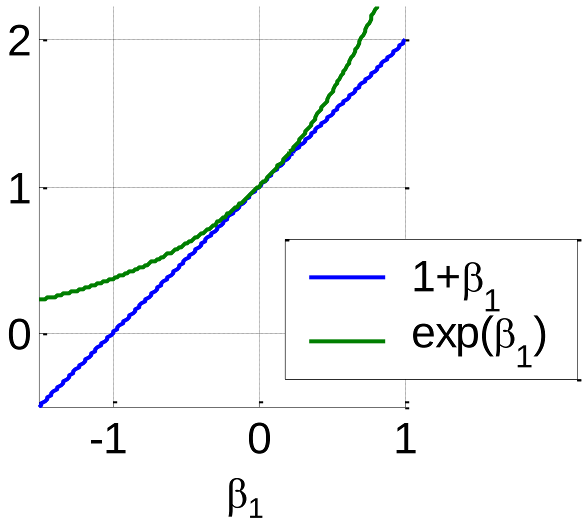
$$\ldots\ldots$$

$$\mathrm{E}[wage \mid educ] = \underbrace{w}_{e^{\beta_0}} \times \underbrace{1.05^{educ}}_{e^{\beta_1}}$$

- let's generalize this type of relationship with parameters β_0 and β_1

Functional Form (cont'd)

- this brings us to the relationship $\mathsf{E}[wage \mid educ] = \exp(\beta_0 + \beta_1\, educ)$
  - let's focus on the meaning of β_1 now
  - in the five-percent-a-year example, we had $\exp(\beta_1) = 1.05$
    - for β_1, this gives us $1.05 = e^{0.049} \approx e^{0.05}$, thus $\beta_1 \approx 0.05$
    - this can be generalized: for a small β_1, it holds $1 + \beta_1 \approx e^{\beta_1}$
    - therefore, β_1 tells us the (expected) percentage change in *wage* with an additional year of *education*



β_1	$\exp(\beta_1)$	$\%\Delta wage$
0.02	1.020	2.0%
0.05	1.051	5.1%
0.20	1.221	22.1%
0.50	1.648	64.8%

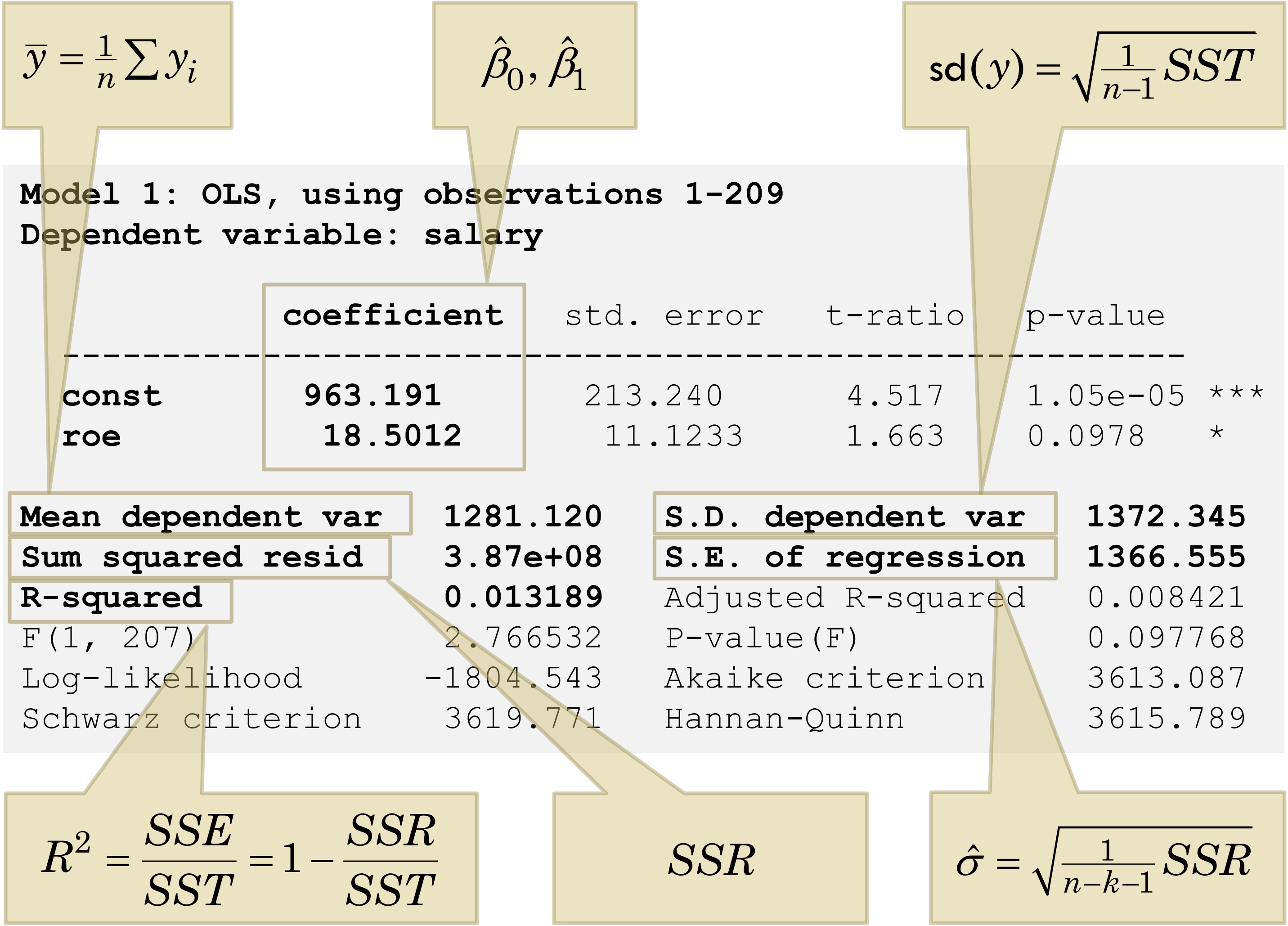
Functional Form (cont'd)

- note that $wage = \exp(\beta_0 + \beta_1 educ)$ ↔ $\log(wage) = \beta_0 + \beta_1 educ$
- logarithm transform is one of the basic econometric tools
- the rule to remember: taking the log of one of the variables means we shift from absolute changes to relative changes:

regression function	interpretation of β_1
$y = \beta_0 + \beta_1 x$	$\Delta y = \beta_1 \Delta x$
$\log y = \beta_0 + \beta_1 x$	$\%\Delta y = (100\,\beta_1)\,\Delta x$
$y = \beta_0 + \beta_1 \log x$	$\Delta y = (0.01\,\beta_1)\,\%\Delta x$
$\log y = \beta_0 + \beta_1 \log x$	$\%\Delta y = \beta_1 \%\Delta x$

- **constant elasticity model**: $\log y = \beta_0 + \beta_1 \log x + u$
  - x-elasticity of y: $\beta_1 = E_{y,x} = \dfrac{\partial \log y}{\partial \log x} = \dfrac{\partial y}{\partial x} \cdot \dfrac{x}{y} = \dfrac{\%\Delta y}{\%\Delta x}$

Gretl Output: An Overview



$\bar{y} = \frac{1}{n}\sum y_i$

$\hat{\beta}_0, \hat{\beta}_1$

$\mathrm{sd}(y) = \sqrt{\frac{1}{n-1} SST}$

```
Model 1: OLS, using observations 1-209
Dependent variable: salary

             coefficient   std. error   t-ratio    p-value
  --------------------------------------------------------
  const        963.191     213.240        4.517    1.05e-05 ***
  roe           18.5012     11.1233       1.663    0.0978   *

Mean dependent var   1281.120   S.D. dependent var   1372.345
Sum squared resid    3.87e+08   S.E. of regression   1366.555
R-squared            0.013189   Adjusted R-squared   0.008421
F(1, 207)            2.766532   P-value(F)           0.097768
Log-likelihood      -1804.543   Akaike criterion     3613.087
Schwarz criterion    3619.771   Hannan-Quinn         3615.789
```

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

SSR

$\hat{\sigma} = \sqrt{\frac{1}{n-k-1} SSR}$

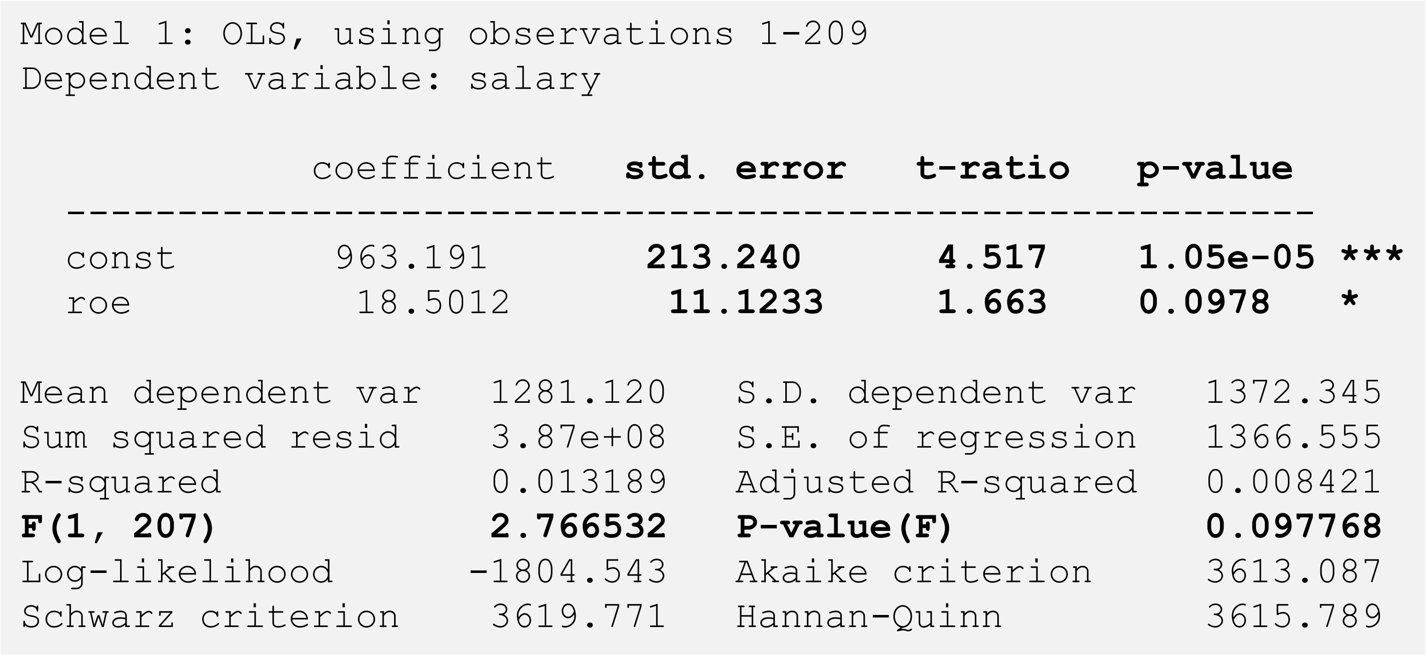
Classical Linear Regression

OLS estimates as realizations of random variables.

Mean and variance of the OLS estimator.

A Note on Where We're Heading…

- as you've seen, we've only covered a small part of the *Gretl* output yet
- gradually, we'll build up the theory behind the following parts:



```
Model 1: OLS, using observations 1-209
Dependent variable: salary

             coefficient   std. error   t-ratio   p-value
  --------------------------------------------------------
  const       963.191       213.240      4.517    1.05e-05 ***
  roe          18.5012       11.1233     1.663    0.0978   *

Mean dependent var   1281.120   S.D. dependent var   1372.345
Sum squared resid    3.87e+08   S.E. of regression   1366.555
R-squared            0.013189   Adjusted R-squared   0.008421
F(1, 207)            2.766532   P-value(F)           0.097768
Log-likelihood      -1804.543   Akaike criterion     3613.087
Schwarz criterion    3619.771   Hannan-Quinn         3615.789
```

- all of this tells us something about *hypotheses tests* about the β's (this is important for empirical verification of *economic* theories)

OLS Estimator as a Random Variable

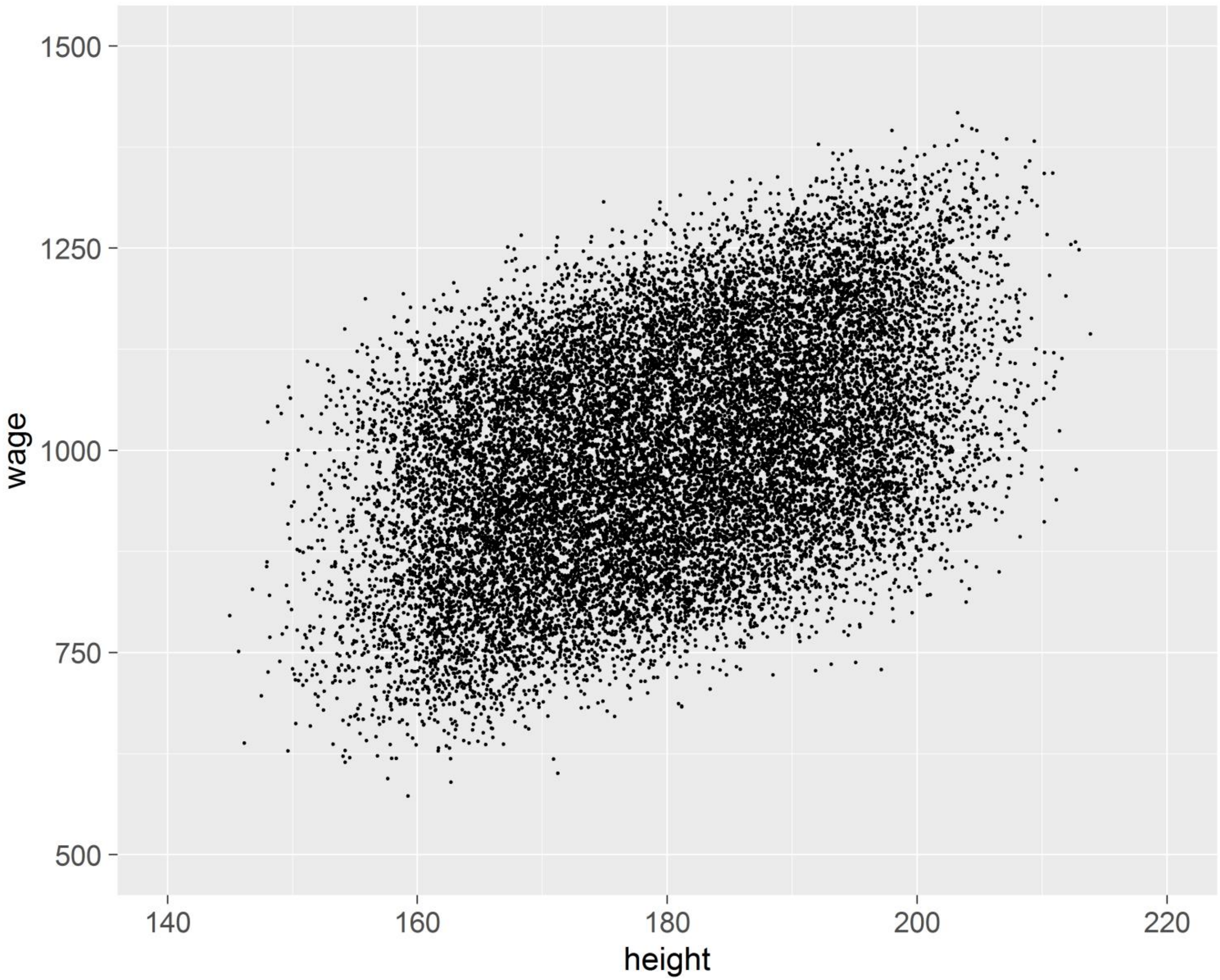
- in our previous discussion, we always tried to estimate a population regression function based on a (random) sample of the population
  - we believe there are real (population) values of β_0 and β_1 out there
  - however, we always end up with only their estimates $\hat{\beta}_0$ and $\hat{\beta}_1$
  - the value of these estimates depends on the specific sample we get the data for → if we go and collect another sample, we'll have different estimates

→ because of random sampling, $\hat{\beta}_0$ and $\hat{\beta}_1$ can be treated as random variables; the eventual values that we obtain are their realizations

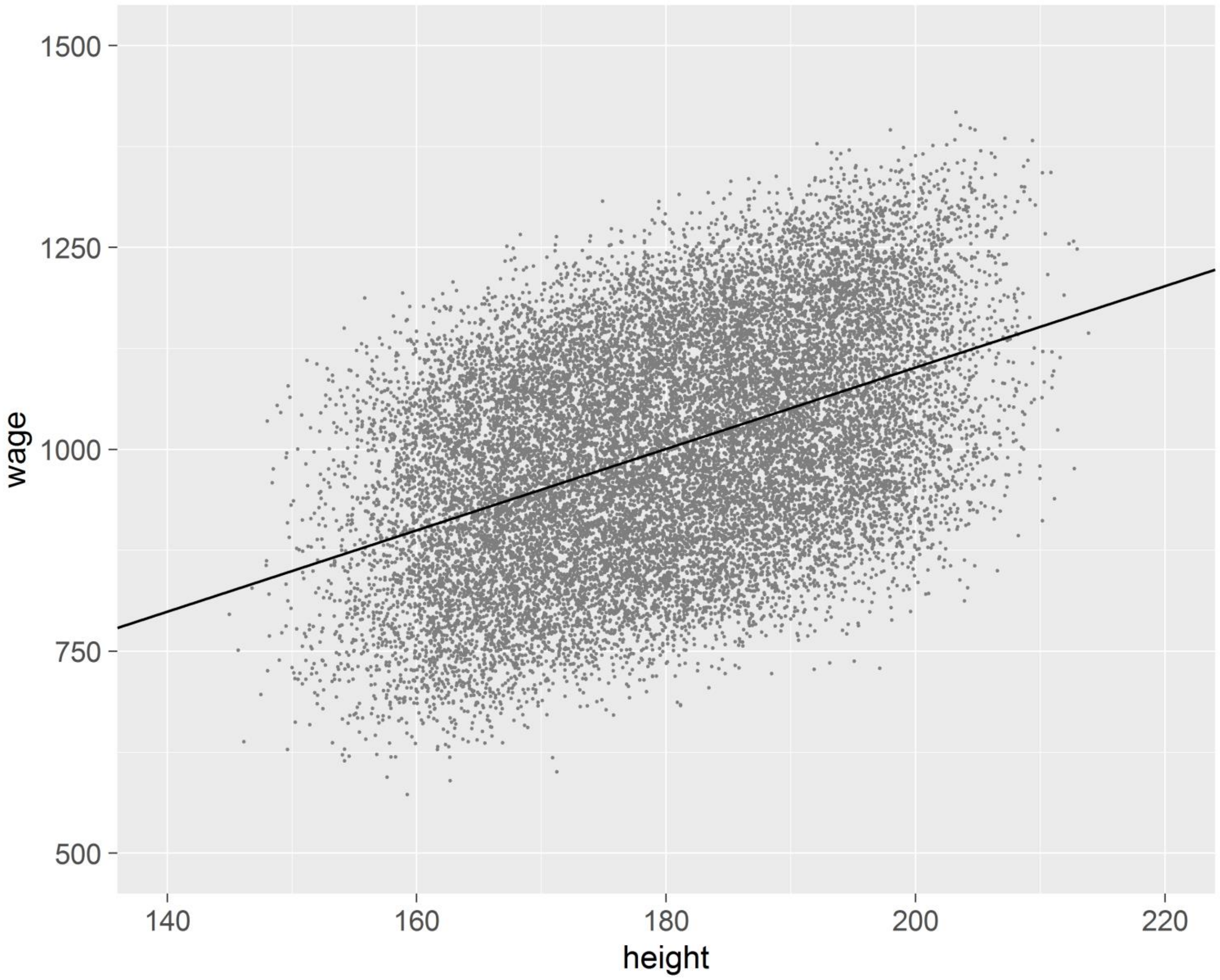
  - note the difference between *estimators* (the RVs) and *estimates* (eventual values)

- it's quite natural to ask questions like:
  - are my estimates accurate enough? What level of imprecision should I count with?
  - is the OLS estimator *unbiased*? Or is it possible that, *on average*, the estimates tend to overrate/underrate the intercept/slope?

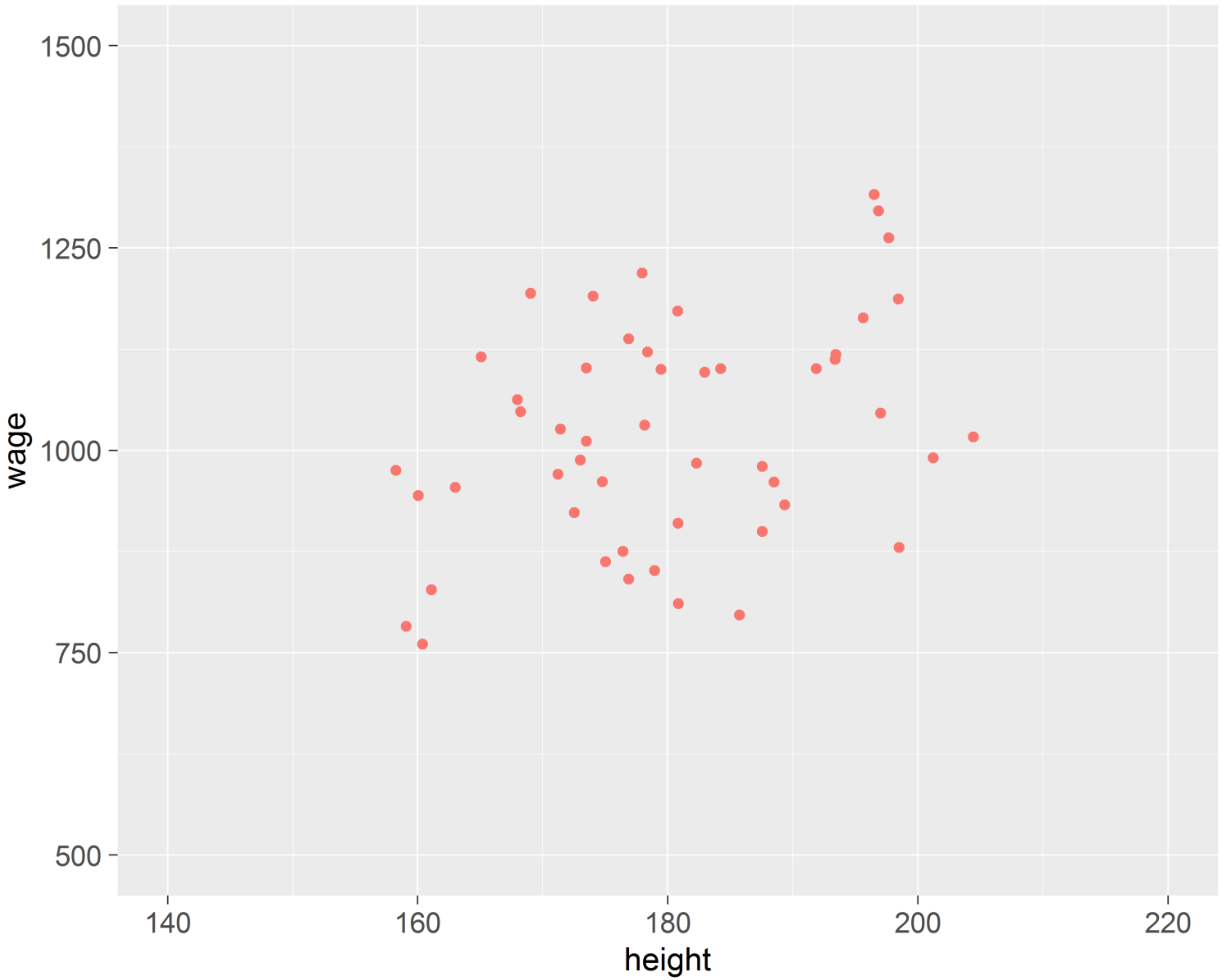
Wages vs. height in a (fictitious) population – complete data



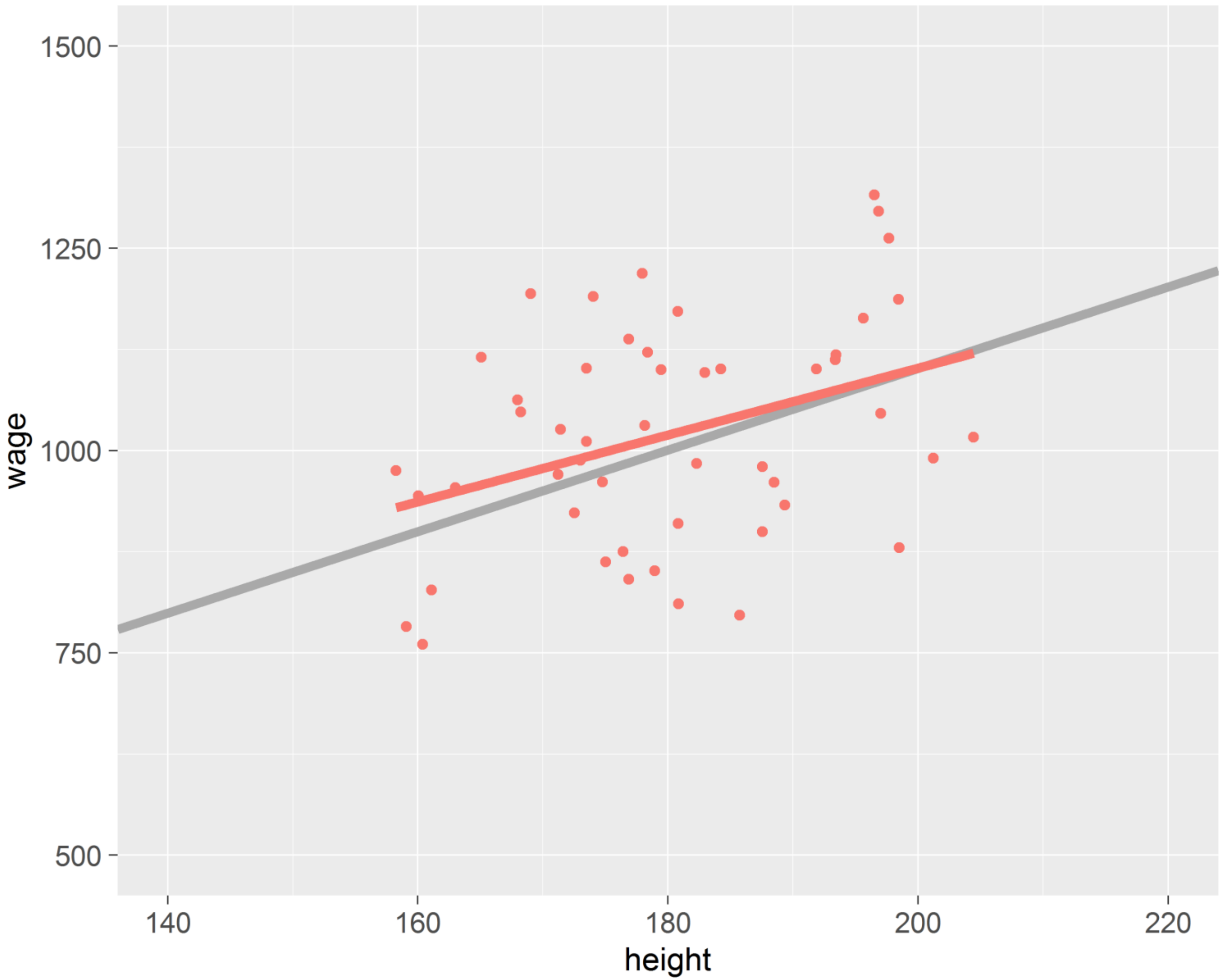
Population regression function

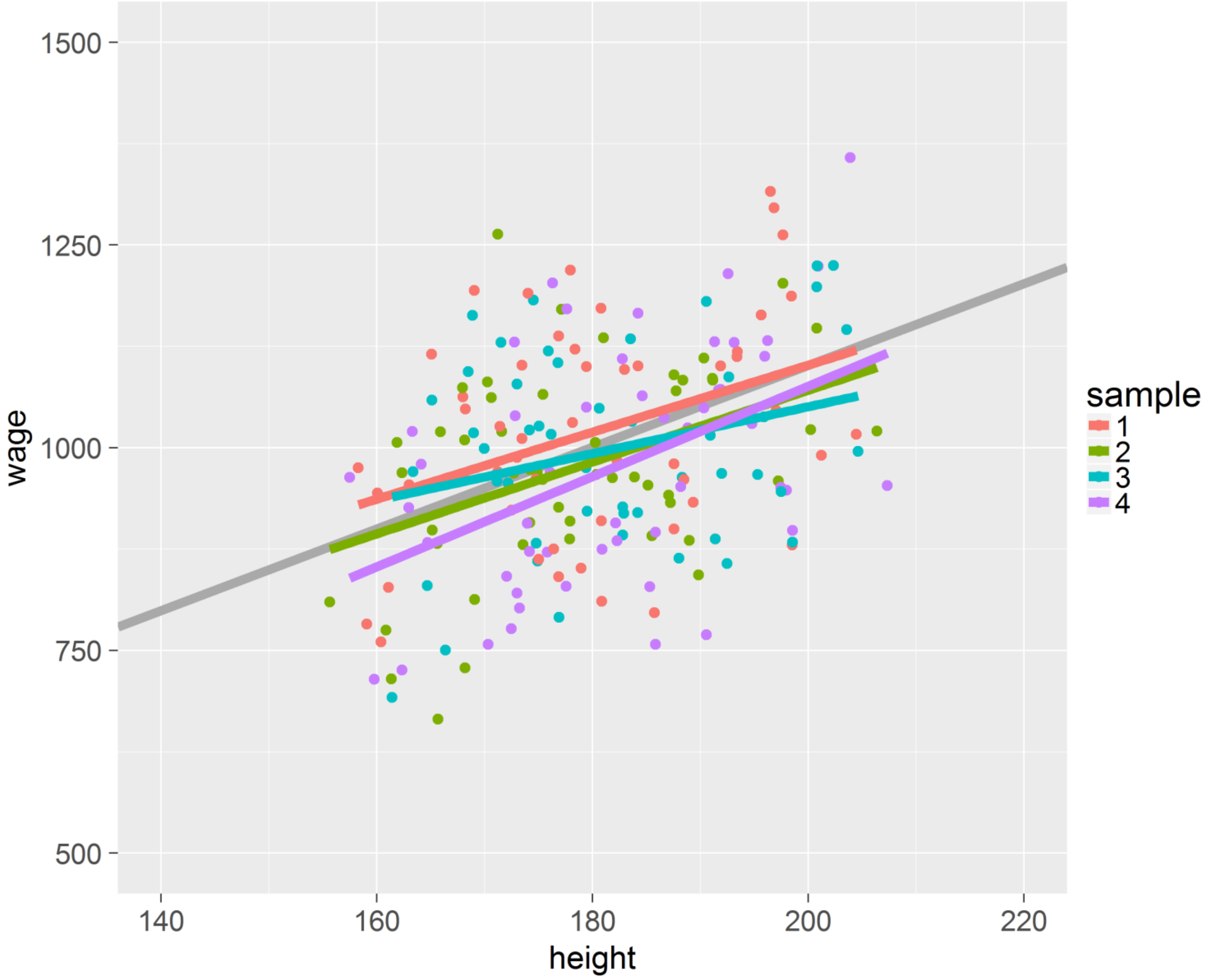


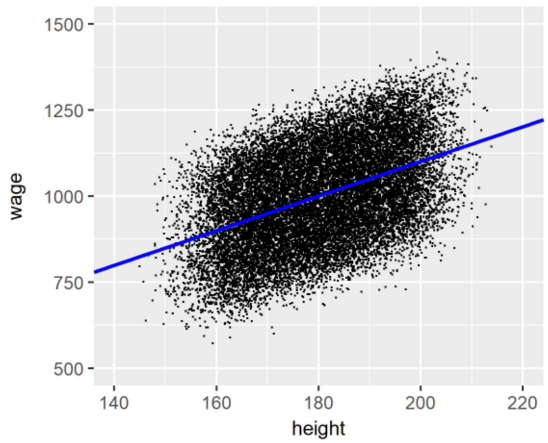
Typically, we only know one sample



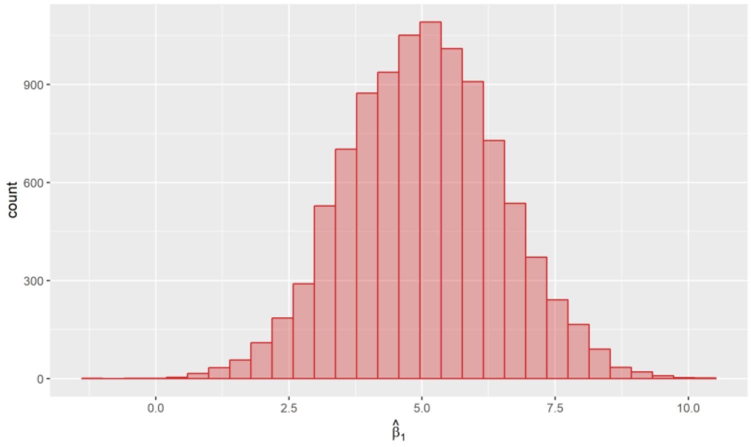
SRF vs PRF




1500
1250
1000
750
500
wage
140
160
180
200
220
height
sample
1
2
3
4



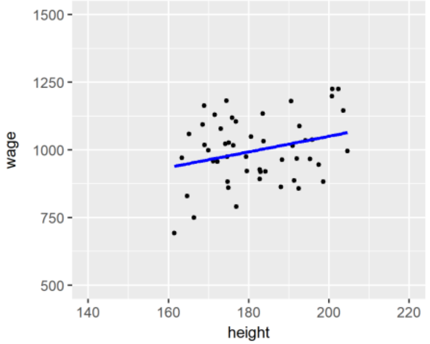
Sampling distribution of $\hat{\beta}_1$



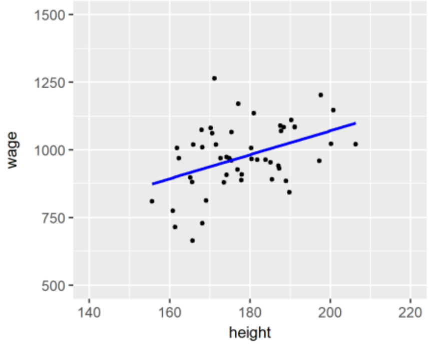
Sample	$\hat{\beta}_1$
1	3.53
2	5.76
⋮	⋮
10^6	4.71
Mean	5.040
SD	1.438

sample 1   sample 2   . . .   sample 10^6

$\hat{\beta}_1 = 3.53$

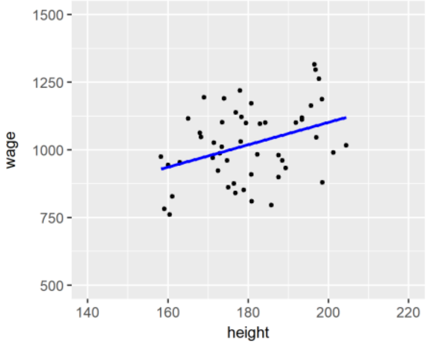


$\hat{\beta}_1 = 5.76$



. . .

$\hat{\beta}_1 = 4.71$



OLS Estimator as a Random Variable (cont'd)

- if we translate these questions into the RV framework, we'll be asking about the *variance* and *mean* of $\hat{\beta}_0$ and $\hat{\beta}_1$
- so far, it hasn't really made a difference whether we took the descriptive, causal or predictive approach
  - the estimates were the same, and so were their algebraic properties
  - the discussion about units and functional form were not related to all of this
  - the goodness of fit wasn't either
- in order to say something about the properties of RVs $\hat{\beta}_0$ and $\hat{\beta}_1$, we need to make some assumptions about the population and the sample
  - these will be mostly in line with the causal model (note that the causal model was the one with the most assumptions)
  - e.g., the simple descriptive approach doesn't really work with the respective part of the *Gretl* output (!)
- the set of assumptions (SLR.1 through SLR.6) we'll introduce is often referred to as the **classical linear regression model** (CLRM)

Assumptions of CLRM

- we'll introduce assumptions SLR.1 to SLR.4 ("SLR" stands for ***simple linear regression***)

Assumption **SLR.1** (linear population model) :

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

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

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

- notice that in making this assumption we have really moved to the "structural world"
- we are really saying that this is the actual **data-generating process** and our goal is to uncover the true parameters

Assumptions of CLRM (cont'd)

Assumption **SLR.2** (random sampling):

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

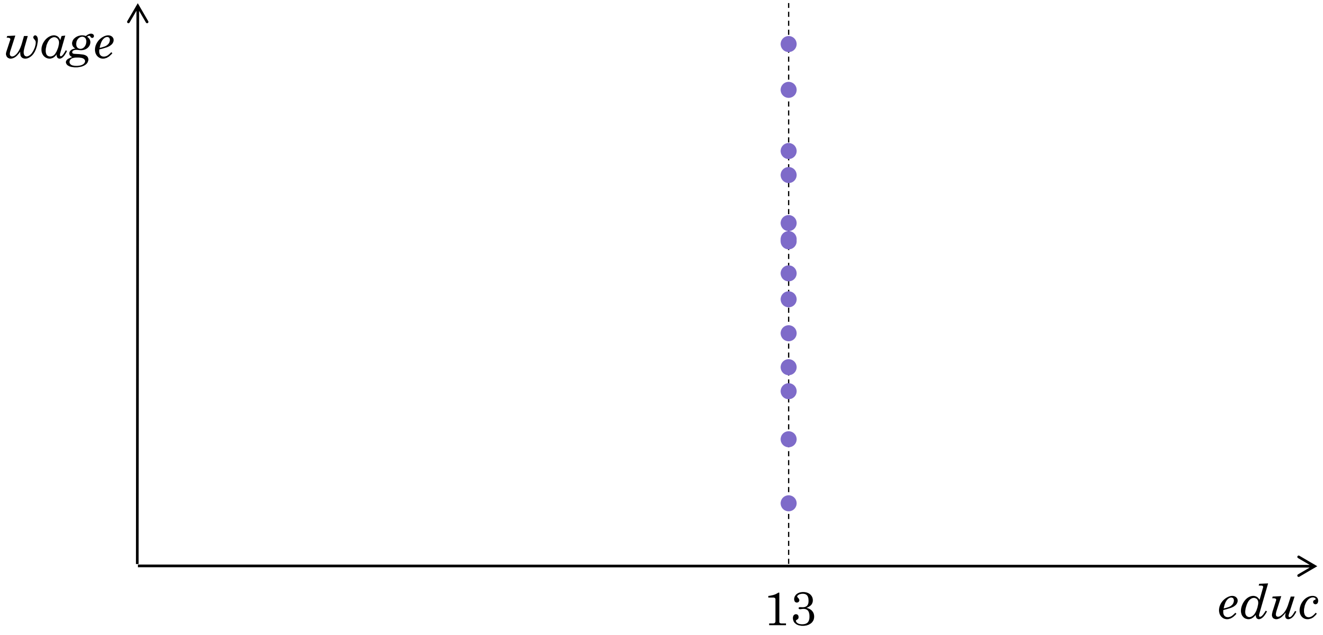
- not all cross-sectional samples can be viewed as outcomes of random samples, but many can be
  - with time series, we'll have to put things differently
- the next assumption effectively allows us to estimate the model

Assumption **SLR.3** (sample variation in the explanatory variable):

The sample outcomes on x, namely $\{x_i , i = 1,\ldots, n\}$, are not all the same value.

Assumptions of CLRM (cont'd)

- technically, the denominator for $\hat{\beta}_1$ is $\sum_{i=1}^{n}(x_i - \overline{x})^2$, which would be zero if SLR.3 didn't hold
- in other words, how would you estimate the slope here:



- *note*: in practical applications, SLR.3 always holds

Assumptions of CLRM (cont'd)

Assumption **SLR.4** (zero conditional mean of *u*):

The error *u* has an expected value of zero given any value of the explanatory variable. In other words, $\mathsf{E}[u \mid x] = 0$.

- as you know, this assumption is the crucial one for causal interpretation; at the same time, we need it in order to derive the theoretical properties of the OLS estimator
- as I've already noted, we make this assumption without being able to check it by statistical means
- therefore, in applications, its validity has to be argued from outside (economic theories, common sense)
  - in practice, this means we have to rule out the $y \rightarrow x$ and $y \leftarrow z \rightarrow x$ causation schemes (see lecture 2 for more details)
- note that for our random sample, SLR.4 implies $\mathsf{E}[u_i \mid x_1,\ldots,x_n] = 0$
  - we'll use the shorthand notation $\mathbf{x}$ for $x_1,\ldots,x_n$ (e.g., $\mathsf{E}[u_i \mid \mathbf{x}] = 0$)

Mean of the OLS Estimator

- you already know that under the assumption of random sampling (SLR.2), $\hat{\beta}_0$ and $\hat{\beta}_1$ can be treated as RVs
- our goal now is to find $\mathsf{E}\hat{\beta}_0$ and $\mathsf{E}\hat{\beta}_1$
- a short preview:
  - somehow, we want to use the assumption that $\mathsf{E}[u \mid x] = 0$
  - this, however, can apply only when speaking about *conditional expectations* of the estimates
  - therefore, we'll first learn something about $\mathsf{E}[\hat{\beta}_0 \mid \mathbf{x}]$ and $\mathsf{E}[\hat{\beta}_1 \mid \mathbf{x}]$
  - then we'll use the *law of iterated expectations* (see our Exercise 1.13*b* or Wooldridge, page 687) which tells us

$$\mathsf{E}\hat{\beta}_0 = \mathsf{E}\left(\mathsf{E}[\hat{\beta}_0 \mid \mathbf{x}]\right)$$

$$\mathsf{E}\hat{\beta}_1 = \mathsf{E}\left(\mathsf{E}[\hat{\beta}_1 \mid \mathbf{x}]\right)$$

- we'll start with $\hat{\beta}_1$
- in order to use the assumption above, we need to express $\hat{\beta}_1$ using u

A Note on the Law of Iterated Expectations

$$\mathsf{E}(wage) = \mathsf{E}\big(\mathsf{E}[wage \mid educ]\big)$$

- an analogy to the following population problem
- for simplicity, education classified into three categories

education	low	medium	high
average wage	500	700	800
% of the population	20	50	30

- the average wage in the population:

$$500 \times .20 \; + \; 700 \times .50 \; + \; 800 \times .30$$

- or, in words, the weighted average, $\mathsf{E}(\cdot)$, of the average wage in individual categories, $\mathsf{E}[wage \mid educ]$

Mean of the OLS Estimator (cont'd)

- I won't show all the algebra behind it here (see Wooldridge, pages 49–50 for details, or try to derive it yourselves), but the idea is:

We substitute SLR.1 into the OLS formula…

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

SLR.1: $$y_i = \beta_0 + \beta_1 x_i + u_i$$

…to finish with this:

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

estimate

true (population) value

note: only *x* and *u*, → we got rid of *y*

- now we're ready to take the conditional expectation of $\hat{\beta}_1$ and use SLR.4

given **x**, all of this is constant

$$\mathrm{E}[\hat{\beta}_1 \mid \mathbf{x}] = \mathrm{E}\left(\beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x})u_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \,\middle|\, \mathbf{x} \right) = \beta_1 + \underbrace{\frac{\sum_{i=1}^n (x_i - \bar{x})\overbrace{\mathrm{E}[u_i \mid \mathbf{x}]}^{0}}{\sum_{i=1}^n (x_i - \bar{x})^2}}_{0} = \beta_1$$

Mean of the OLS Estimator (cont'd)

- we have $\mathsf{E}[\hat{\beta}_1 \mid \mathbf{x}] = \beta_1$, and the law of iterated expectations tells us

$$\mathsf{E}[\hat{\beta}_1] = \mathsf{E}\big(\mathsf{E}[\hat{\beta}_1 \mid \mathbf{x}]\big) = \mathsf{E}(\beta_1) = \beta_1$$

- this tells us that the **OLS estimator is unbiased =** it doesn't systematically overestimate/underestimate the true parameters
  - obviously, unbiasedness is a nice property
  - however, it is only a feature of the *sampling distributions* of $\hat{\beta}_0$ and $\hat{\beta}_1$ which says nothing about the *estimate* that we obtain for a given sample
  - we hope that, if the sample we obtain is somehow "typical," then our estimate should be "near" the population value
- from here, it's easy to show the unbiasedness of $\hat{\beta}_0$:
  - first, note that $\bar{y} = \beta_0 + \beta_1 \bar{x} + \bar{u}$ (just averaging across the sample)
  - therefore, $\hat{\beta}_0 \overset{\text{OLS}}{=} \bar{y} - \hat{\beta}_1 \bar{x} = \beta_0 + (\beta_1 - \hat{\beta}_1)\bar{x} + \bar{u}$
  - and finally $\mathsf{E}\hat{\beta}_0 = \mathsf{E}[\beta_0 + (\beta_1 - \hat{\beta}_1)\bar{x} + \bar{u}] = \mathsf{E}\beta_0 + \underbrace{\mathsf{E}(\beta_1 - \hat{\beta}_1)\bar{x}}_{0} + \underset{0}{\mathsf{E}\bar{u}} = \beta_0$

Mean of the OLS Estimator (cont'd)

- revision: what did we need to show unbiasedness?
  - we started with SLR.1 and the OLS formula to get

OLS + SLR.1 $\Longrightarrow$ $\hat{\beta}_1 = \beta_1 + \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})u_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

  - note that in here, SLR.3 was implicitly used (no SLR.3, no slope)
  - then we needed SLR.2 and SLR.4:

SLR.4 + SLR.2 $\Longrightarrow$ $\mathsf{E}[u_i \mid \mathbf{x}] = 0$ $\Longrightarrow$ $\mathsf{E}[\hat{\beta}_1 \mid \mathbf{x}] = \beta_1$

$\mathsf{E}[u \mid x] = 0$ random sampling

  - …and finally we used the law of iterated expectations

→ to sum up, we needed *all four SLR assumptions*

- even though one can sometimes doubt the validity of SLR.1 (*linear* population relationship) or SLR.2 (true *random sampling*), SLR.4 is typically the most the problematic one

Example: Math Performance Vs. Lunch Program

- suppose we wish to estimate the effect of the federally funded school lunch program on student performance. If anything, we expect the lunch program to have a positive ceteris paribus effect on performance: all other factors being equal, if a student who is too poor to eat regular meals becomes eligible for the school lunch program, his or her performance should improve.
- *math10* the percentage of tenth graders at a high school receiving a passing score on a standardized mathematics exam
- *lnchprg* the percentage of students who are eligible for the lunch program

1. Open the `lunch.gdt` data file and regress *math10* on *lnchprg*.
2. Do you think the estimated effect if *lunch program* is causal?
3. Or, do you think that the estimate is *biased*? Why? Explain why one of the SLR assumptions is violated.
4. Suppose an estimator exhibits a downward bias. Is it possible that our eventual estimate will be higher than the population parameter?

Accuracy of OLS Estimates, Efficiency

- so far, we have only dealt with the mean value of our estimates
- we know that with OLS there's no bias, which means that on average, OLS doesn't overestimate/underestimate the true parameters
- it's good to know what happens *on average*, but normally we're only given one shot
- unbiasedness actually tells us nothing about the accuracy of the estimates
- a good measure of accuracy (actually, the most widely-used one) is the *variance* of the estimates
  - if two estimates (A and B) are both unbiased, and $\operatorname{var} A < \operatorname{var} B$, then A is taken as the better of the two (more accurate)
  - we can also say that A is *more efficient* (we'll have a more detailed discussion on the efficiency of estimates later on)
- in order to be able to derive a nice formula for the variance of the OLS estimator, we need to adopt one more assumption about the variance of u

Homoskedasticity

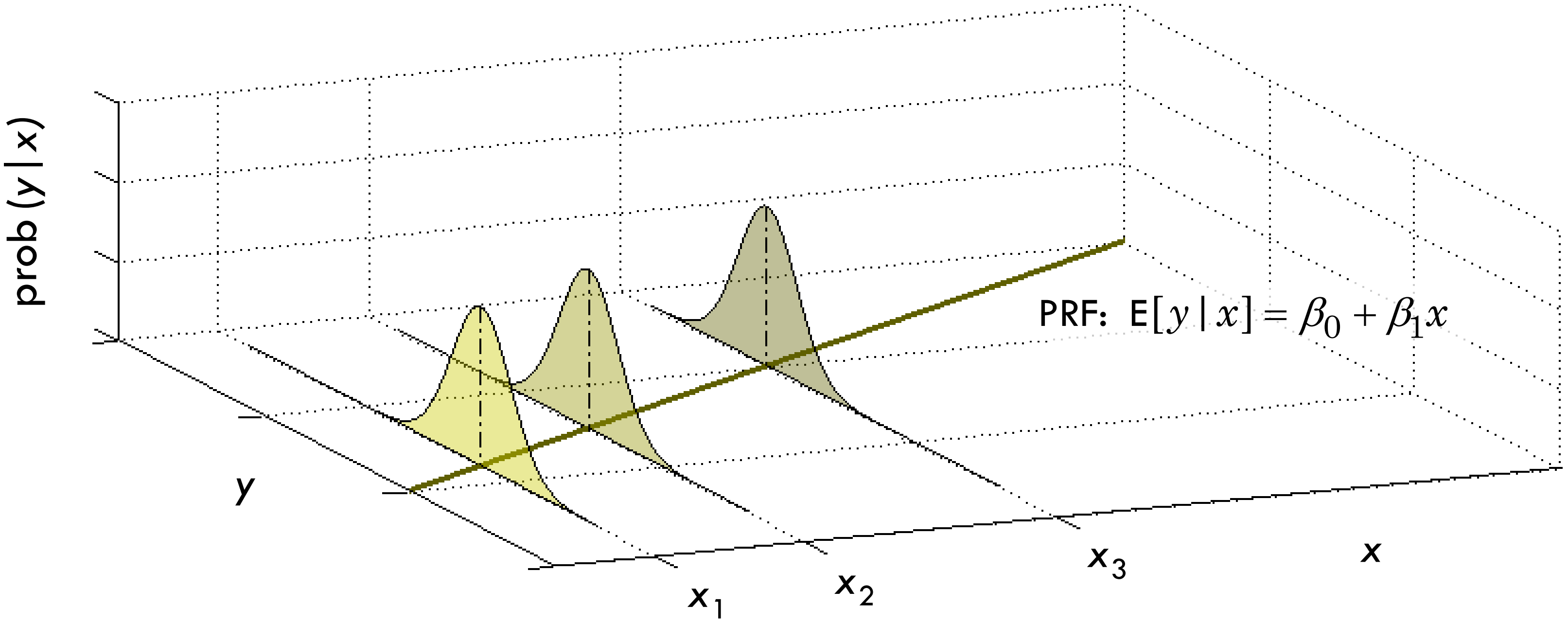
Assumption **SLR.5** (homoskedasticity):

Variance of *u* does not vary with *x*. More precisely, $\text{var}[u \mid x] = \sigma^2$.

- as with the conditional expectation of u (SLR.4), SLR.5 implies two things:
  1. $\text{var}[u \mid x]$ is constant (not varying with x)
  2. $\text{var}\, u = \sigma^2$, i.e. the *unconditional* variance of u is σ^2
- note that once we know x, the only thing that can make y change is u (our model is $y = \beta_0 + \beta_1 x + u$, so u is the only non-constant term on the right-hand side once x is known)
- therefore, we can also re-write SLR.5 as $\text{var}[y \mid x] = \sigma^2$
  - this is typically easier to interpret

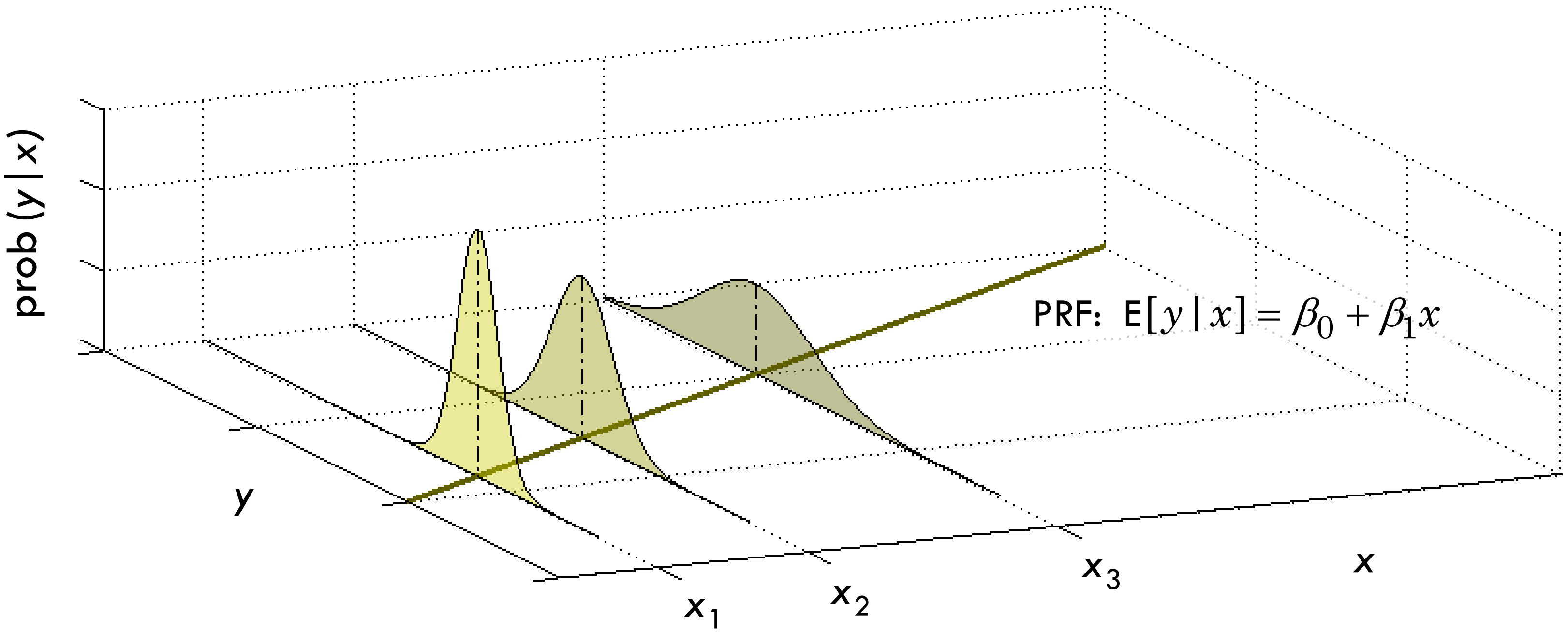
Homoskedasticity (cont'd)

- a model satisfying our assumptions might look as follows
  - the conditional distributions of y have the same "width" (SLR.5) and are centered about the PRF (SLR.4), which is linear (SLR.1)



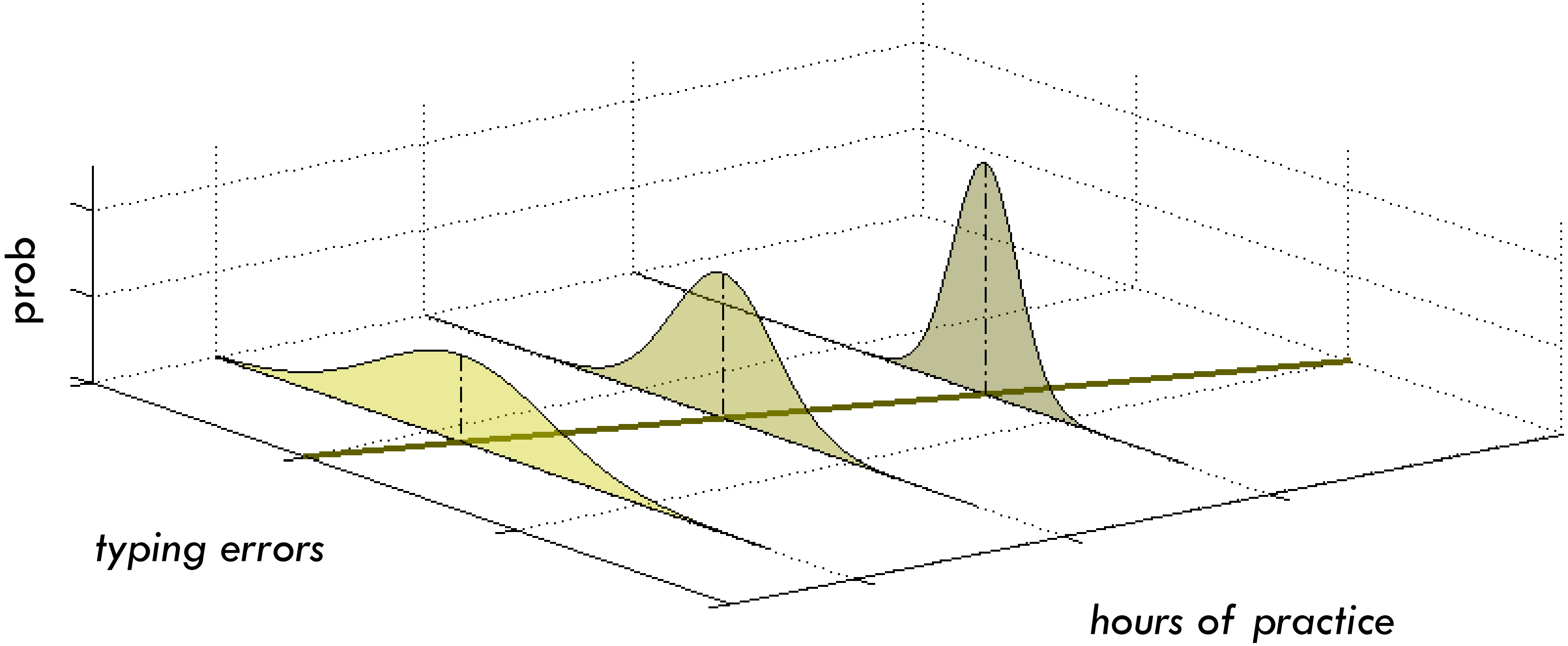
Homoskedasticity (cont'd)

- here, SLR.5 is violated: $\mathsf{var}[y\,|\,x]$ changes with x
  - we call this **heteroskedasticity**
- note: the remaining assumptions are still fulfilled here



Homoskedasticity (cont'd)

- sometimes, we can easily argue that SLR.5 doesn't hold, as in the example with *typing errors* vs. *hours of practice*:
  - with more practice, people cut down on mistakes, and their natural prerequisites gradually cease to play an important role (thus reducing the variance of results)



Variance of the OLS Estimator

- revision of the rules for variance calculations:
  - $\text{var}(3u + \cancel{4}) = 3^2 \text{ var } u$
  - $\text{var}[\Sigma u_i] = \Sigma \text{ var } u_i$ if u_i are independent (for us, this is true because of random sampling – SLR.2)
  - these rules apply to *conditional variance* as well
- when we derived the mean of the OLS estimator, we used the following:

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

- in order to simplify notation, we define $s_x^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$, thus

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

- note that SLR.5 and random sampling give us $\text{var}[u_i \,|\, \mathbf{x}] = \sigma^2$
- we can also write $\text{var}[(x_i - \bar{x})u_i \,|\, \mathbf{x}] = (x_i - \bar{x})^2\sigma^2$, because conditional on $\mathbf{x}$, $(x_i - \bar{x})$ can be treated as a constant

Variance of the OLS Estimator (cont'd)

$$\text{var}[\hat{\beta}_1 \mid \mathbf{x}] = \text{var}\left(\beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})u_i}{s_x^2} \middle| \mathbf{x} \right) =$$

$$= \frac{\text{var}\left[\sum_{i=1}^{n}(x_i - \bar{x})u_i \middle| \mathbf{x} \right]}{(s_x^2)^2} =$$

$$= \frac{\sum_{i=1}^{n} \text{var}[(x_i - \bar{x})u_i \mid \mathbf{x}]}{(s_x^2)^2} =$$

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

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

$$= \frac{\sigma^2}{s_x^2}$$

Variance of the OLS Estimator (cont'd)

$$\text{var}[\hat{\beta}_1 \mid \mathbf{x}] = \text{var}\left(\cancel{\beta_1} + \frac{\sum_{i=1}^{n}(x_i - \bar{x})u_i}{s_x^2} \,\middle|\, \mathbf{x}\right) =$$

$$= \frac{\text{var}\left[\sum_{i=1}^{n}(x_i - \bar{x})\,u_i \,\middle|\, \mathbf{x}\right]}{(s_x^2)^2} =$$

$$= \frac{\sum_{i=1}^{n} \text{var}[(x_i - \bar{x})\,u_i \mid \mathbf{x}]}{(s_x^2)^2} =$$

$= (x_i - \bar{x})^2\sigma^2$

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

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

$= s_x{}^2$

$$= \frac{\sigma^2}{s_x^2}$$

Variance of the OLS Estimator (cont'd)

- put together, we have:

$$\text{var}[\hat{\beta}_1 \mid \mathbf{x}] = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

the variance of *u*

the sample variance of *x* (times *n* – 1)

→ as far as the accuracy of $\hat{\beta}_1$ is concerned…

- …the *less* variance in the disturbances, the better
- …the *more* variance in the explanatory variable, the better

- on the meaning of *conditional on* **x**:
  - it's the same as treating the x_i as *fixed in repeated samples*
  - this is easily done in a computer simulation study
    - imagine we keep the *x*-values constant instead of generating them at random each time, and for new samples, we generate *u* only
    - running the trials this way tells us something about the conditional distribution of $\hat{\beta}_1$

Estimating the Error Variance (σ^2)

- first note that as $\mathsf{E}\,u = 0$, it holds $\mathsf{var}\,u = \mathsf{E}\,u^2$
- therefore, in our sample, $\frac{1}{n}\sum_{i=1}^{n} u_i^2$ is an unbiased estimator of $\mathsf{var}\,u = \sigma^2$
- unfortunately, in practical applications this is useless, as we don't know the u_i's
- instead of random errors, we'll use the residuals (which we do know)
- however, $\frac{1}{n}\sum_{i=1}^{n} \hat{u}_i^2 = \frac{1}{n} SSR$ is not an unbiased estimator of σ^2
  - the reason is that the residuals are not independent: we know that

$$\sum_{i=1}^{n} \hat{u}_i = 0$$

$$\sum_{i=1}^{n} x_i \hat{u}_i = 0$$

  - therefore, if I tell you the first $n-2$ residuals, you can tell me the values of the remaining two (by solving the equations above)
- it can be shown (see the Wooldridge book) that an unbiased estimator is

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} \hat{u}_i^2 = \frac{SSR}{n-2}$$

Standard Errors of OLS Estimates

- in the formula for $\text{var}[\hat{\beta}_1|\,\mathbf{x}]$, we needed σ^2 in order to calculate the conditional variance
- once we have estimated the error variance, we can use it to estimate the variance of the OLS estimator based on our sample
- we'll work with standard deviations rather than variances
- the standard deviation of $\hat{\beta}_1$ is the square root of its variance:

$$\text{sd}(\hat{\beta}_1) = \sqrt{\frac{\sigma^2}{\sum (x_i - \overline{x})^2}}$$

- if we replace σ^2 with estimate $\hat{\beta}_1$, we'll obtain an estimate of $\text{sd}(\hat{\beta}_1)$, which is called the *standard error of* $\hat{\beta}_1$

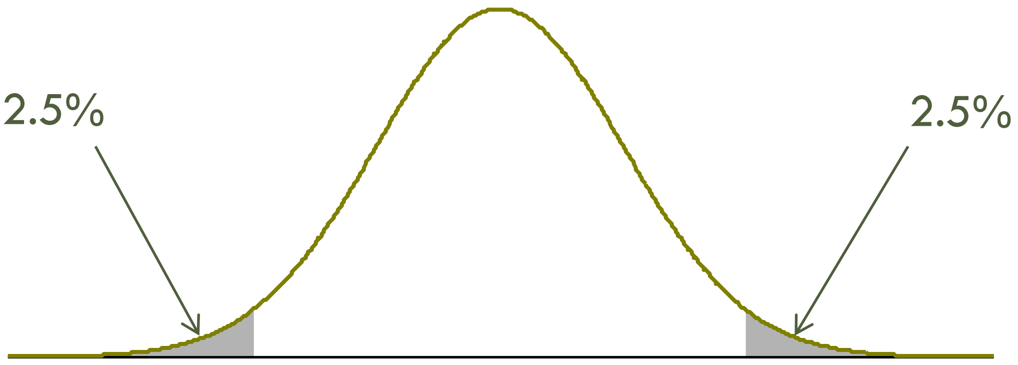
$$\text{se}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{\sum (x_i - \overline{x})^2}}$$

Sampling Distribution of the OLS Estimator

□ so far, we've discussed the basic characteristics of the OLS estimator

□ if we need to test hypotheses about the parameter values, we need to know more than this: we need to know the *sample distribution* of the OLS estimator

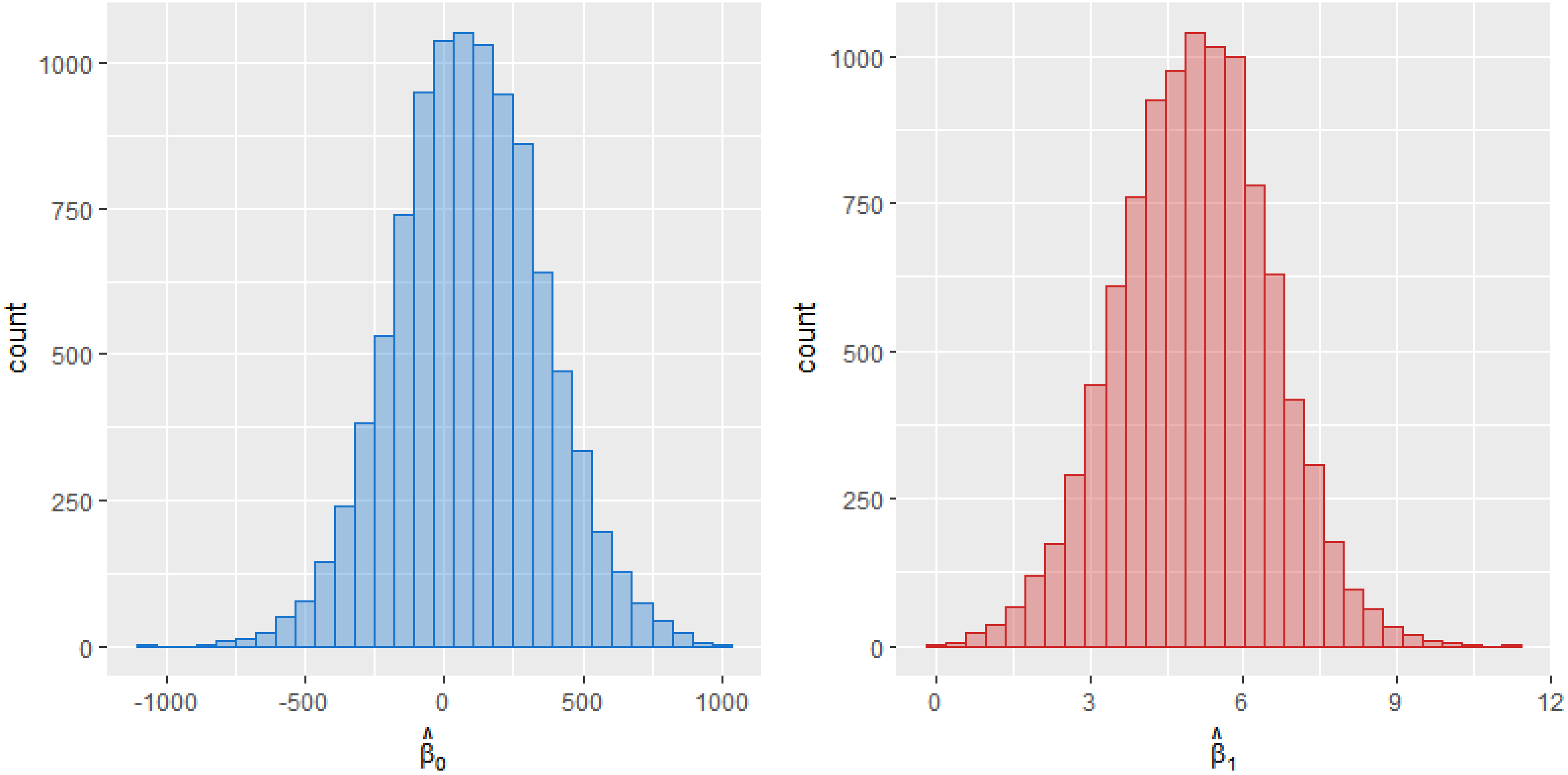
□ recall that in hypothesis testing, we use pictures like this



□ as you've seen in the simulation exercises, the OLS estimates have a distribution that "looks somewhat like the normal distribution"

Sampling Distribution of the OLS Estimator (cont’d)

- the frequency plot for the „wage vs. height“ example was:



Sampling Distribution of the OLS Estimator (cont'd)

- there is a clear tendency towards normality: this obviously has something to do with the *central limit theorem* (CLT)
- the word "tendency" is related to the size of our sample here
  - for the CLT to take effect, we need many observations; the more observations, the closer we are to normality
  - unfortunately, econometricians do not agree on a "safe" number of observations (recommendations vary from 30 to hundreds)
  - in our exercise, 15 was already pretty good, but this depends on many things
- we'll state a theorem about *asymptotic normality* of the OLS estimator
- this theorem can put in many different versions (see Wooldridge, page 168)
- the version I'll show you is the easiest one to write down, and the most useful in calculations
- it works with *standardized* (or "*Studentized*") *estimates*: $\dfrac{\hat{\beta}_j - \beta_j}{\mathrm{se}(\hat{\beta}_j)}$

Sampling Distribution of the OLS Estimator (cont'd)

Theorem: Asymptotic normality of the OLS estimator

Under the assumptions SLR.1 through SLR.5, as the sample size increases, the distributions of standardized estimates converge towards the standard normal distribution *Normal*(0,1).

- we can use this theorem to carry out hypothesis tests about β's in case our sample is large enough (but, what does "large enough" mean, eh?)
- with a small sample, the theorem is rather useless; however, we can give precise results here if we introduce another assumption:

Assumption **SLR.6** (normality):

The population error u is *independent* of the explanatory variable and is normally distributed with zero mean and variance σ^2 :

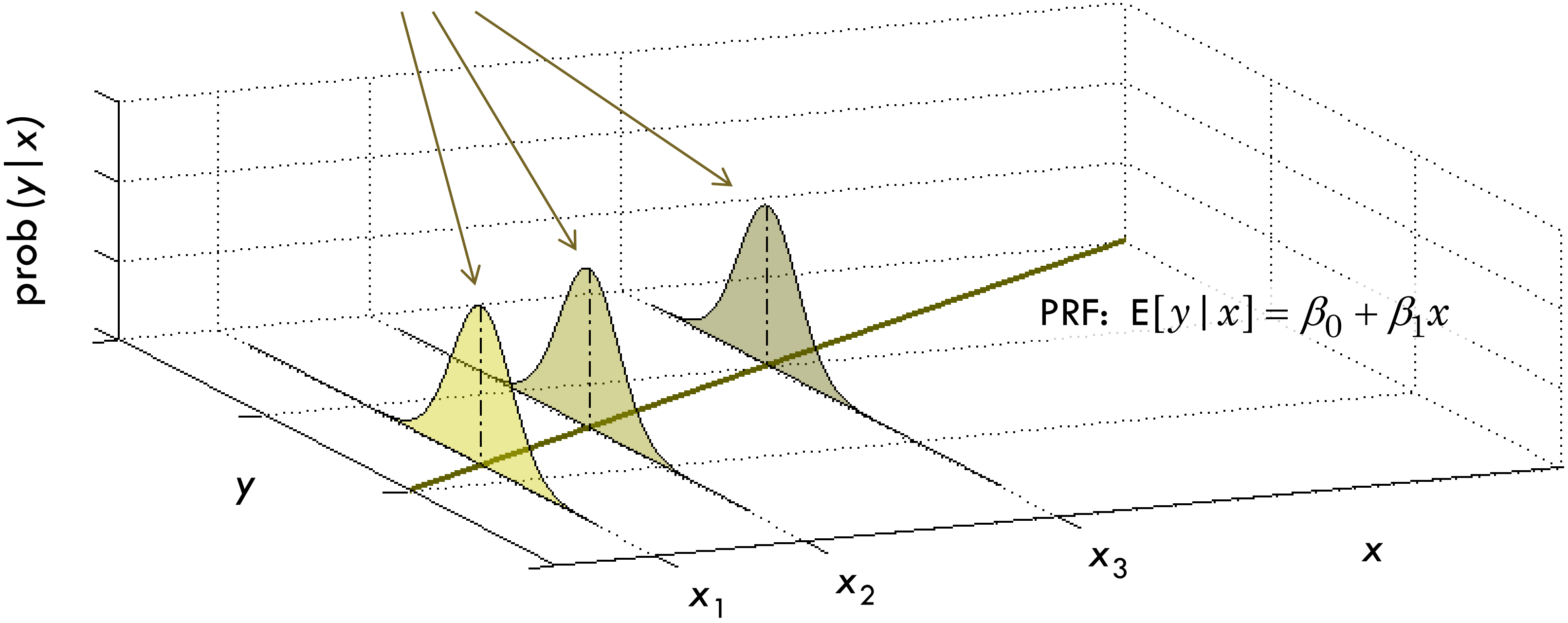
$$u \sim Normal(0, \sigma^2).$$

Sampling Distribution of the OLS Estimator (cont'd)

- SLR.6 is much stronger than any of our previous assumptions
  - it actually implies both SLR.4 and SLR.5 (why?)
- a succinct way to put the population assumptions (all but SLR.2) is:

$$y \mid x \sim \text{Normal}(\beta_0 + \beta_1 x, \sigma^2)$$

identical normal distributions



PRF: $\mathrm{E}[y \mid x] = \beta_0 + \beta_1 x$

Sampling Distribution of the OLS Estimator (cont'd)

- even though some arguments can be given that justify this assumption in real applications, many examples where SLR.6 cannot hold can be found; we'll talk about this later on in more detail

Theorem: Sampling distributions under normality.

Under the assumptions SLR.1 to SLR.6, conditional on the sample values of the explanatory variable,

$$\hat{\beta}_1 \sim \mathrm{Normal}(\beta_1, \mathrm{var}\,\hat{\beta}_1),$$

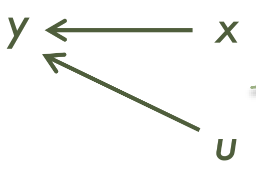
which implies that $\left(\hat{\beta}_1 - \beta_1\right) / \mathrm{sd}(\hat{\beta}_1) \sim \mathrm{Normal}(0,1)$.

Moreover, it holds $\left(\hat{\beta}_1 - \beta_1\right) / \mathrm{se}(\hat{\beta}_1) \sim t_{n-2}$ (Student's *t* distribution).

- the same holds for β_0 estimates, but we haven't talked about the formulas for standard errors in this case

Omitted Variable Bias: A Case for Multiple Regression

- imagine we're regressing y on x, even though there's a substantial role of the $y \leftarrow z \rightarrow x$ relationship
- in ignoring z, we basically omitted an important variable from our considerations
- for the reasons we discussed earlier, SLR assumptions of model $y = \beta_0 + \beta_1 x + u$ result in the following causal picture:

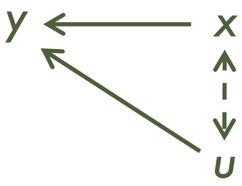


no correlation between x and u (SLR.4)

- however, if there's the $y \leftarrow z \rightarrow x$ influence, then necessarily u contains z, and is therefore correlated with x
- therefore, in the picture above

Omitted Variable Bias (cont’d)

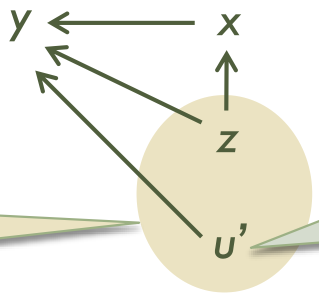
□ therefore, the correct version of our picture is



x and *u* are correlated

which already is a problem

□ a more precise picture should contain z



all of this was *u* before

u’ represents whatever is left in *u* after we pull *z* out of it. We assume here that *u*’ is correlated with neither *x* nor *z*.

□ here, the connection between x and y leads through two paths: $x \rightarrow y$ (direct influence) and $x \leftarrow z \rightarrow y$ (indirect influence)

Omitted Variable Bias (cont'd)

- if we estimate the CLRM model $y = \beta_0 + \beta_1 x + u$ (despite knowing that the SLR assumptions are not satisfied), the estimate of β_1 captures both the direct and indirect influence
- therefore, $\hat{\beta}_1$ is *not unbiased* anymore!
- in fact, one can show that…

$$\mathrm{E}\hat{\beta}_1 = \underbrace{\beta_1}_{\text{direct influence } x \to y} + \overbrace{\underbrace{\mathrm{corr}(x,z)\cdot\mathrm{corr}(z,y)}_{\text{indirect influence } x \leftarrow z \to y}\cdot\underbrace{\frac{\sigma_y}{\sigma_x}}_{\text{scaling factor}}}^{\text{omitted variable bias}}$$

- fortunately, there's an easy way out of this problem: multiple regression
- it suffices to estimate $y = \beta_0 + \beta_1 x + \beta_2 z + u$ instead (next lecture)

Omitted Variable Bias (cont'd)

corr(x,z)	corr(z,y)	OVB
+	+	+
+	–	–
–	+	–
–	–	+

Lecture 3:
Simple Regression II

Jan Zouhar

Introductory Econometrics