BUSN 5000

Regression Fundamentals

Chris Cornwell

Terry College of Business

Spring 2022

Section 1

Wages and Schooling

The Card study

What is the payoff to completing another year of school? That's the question we'll use to launch our review of regression.

To guide us, we'll replicate key results from David Card's well-known study of the wages of young men:

Card, D., "Using Geographic Variation in College Proximity to Estimate the Return to Schooling", in *Aspects of Labour Market Behavior: Essays in Honour of John Vanderkamp*, E. Christophides, et al., eds, Toronto: University of Toronto Press (1995).

He obtained the data for his study from the National Longitudinal Survey of Young Men (NLSYM), which began with 5525 men between 14-24 years old in 1966. Card's sample is based on the 1976 wave and included 3010 observations.

The `card` data are available in the `wooldridge` package.

Learning about populations from random samples

In the broadest sense, we are interested in learning about the world through *random sampling* from some *population*. Learning involves two related activities: *estimation* and *inference*.

In the Card (1995) study, we learn about the effect of education on the wages of young men, the *population* of interest, through the *random sample* of young men who comprise the NLSYM. Using data on their wages, years of schooling and other characteristics, Card *estimates* the effect of education on wages and draws an *inference* about the rate of return to education in the population.

Recall that a random sample is a collection of independent and identically distributed (iid) random variables that take on particular values whenever we observe them. Different samples may yield different values. Random sampling is a reasonable assumption for *cross-section* analyses, which will be our focus.

Descriptive statistics

One of the first things to do in any analysis is describe the contents of your sample with a table of summary statistics, including (at least) the *means* and *standard deviations* of the key variables. And, you want the table to be pretty.

```
datasummary(wage + educ + exper + black + south + smsa
            ~ N + Mean + SD + Min + Max, data=card,
            title="Summary statistics, Card (1995) sample")
```

Table 1: Summary statistics, Card (1995) sample

	N	Mean	SD	Min	Max
wage	3010	577.28	262.96	100	2404
educ	3010	13.26	2.68	1	18
exper	3010	8.86	4.14	0	23
black	3010	0.23	0.42	0	1
south	3010	0.40	0.49	0	1
smsa	3010	0.71	0.45	0	1

Section 2

Why Regression?

What is a regression?

- Relationship between two variables.
- I think of a regression as a scatterplot showing the relationship between two variables with a line of best fit going through the graph.
- A line that portrays data points to express a correlation between two or more inputs. Usually has a Beta, an alpha, or a gamma in there somewhere.
- A statistical process by which one attempts to isolate an independent variable, and it's impact on a dependent variable, by controlling for other independent variables.
- It's a prediction model used to predict an outcome based on variable values.

“The no-code revolution”

‘No-Code’ Brings the Power of A.I. to the Masses

A growing number of new products allow anyone to apply artificial intelligence without having to write a line of computer code. Proponents believe the “no-code” movement will change the world.

Source: NYT

The conditional expectation function (CEF)

It would come as no surprise to anyone to discover the wages and education are positively correlated. In the Card data, that correlation is 0.3. Of course, Card is interested in more than that.

In serious data analysis, we want to learn more than simple correlations. Often, we want explain or predict one variable (y) in terms of another (x). In Card's study, education is an x and log wage is the y.

Either way, the place to start is the conditional expectation function (CEF), $E(y|x)$.

CEF decompostion property

Any random variable y can be decomposed into two parts: a part that is explained by x (the CEF) and a part that is uncorrelated with any function of x:

$$y = E(y|x) + u, \quad \text{where } E(u|x) = 0. \tag{1}$$

We should emphasize a distinction between unconditional and conditional expectation.

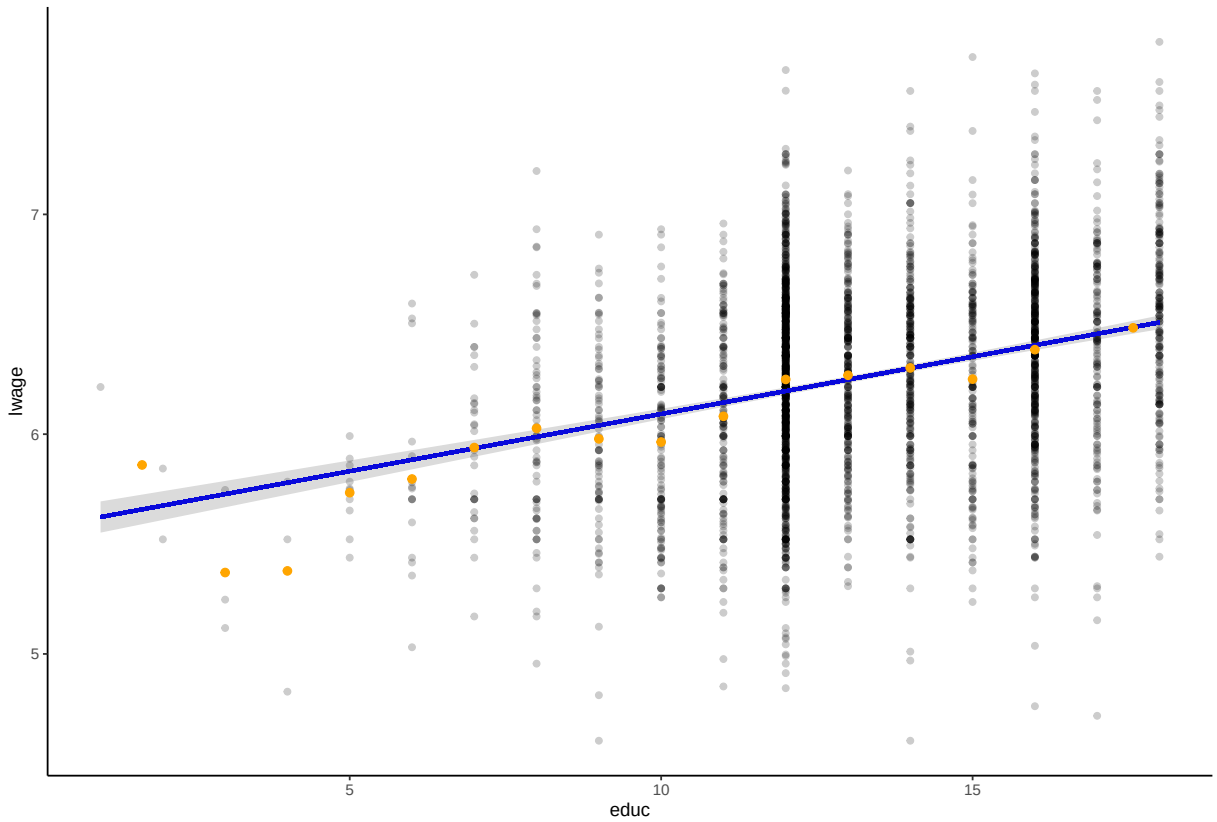
The unconditional expectation of y, $E(y)$, is not a function of y, rather it is a constant. Recall that you typically denote it by μ or μ_y.

The expectation of y given x, $E(y|x)$, is a function of x, and as such, is a random variable. We make this explicit by writing $E(y|x)$ as $\mu(x)$.

CEF calculations using binscatters

We can estimate $E(y|x)$ by calculating the sample average of y for each value of x or within small *bins* of x values.

```
source("binscatter.R")
binscatter(formula="lwage ~ educ", key_var="educ",
           data=card, bins=16, partial=FALSE)
```



The CEF and regression

Or, we can approximate the CEF for any value of x with a regression:

$$E(y|x) = \beta_0 + \beta_1 x, \tag{2}$$

where β_0 and β_1 solve the population least-squares problem:

$$\min E[(y - \beta_0 - \beta_1 x)^2]. \tag{3}$$

The *binscatter* and *regression line* overlay illustrate the contrast between computing the CEF nonparametrically and approximating it with a linear model. The **black dots** represent the raw data and the **orange dots** the bin averages, while the **blue regression line** fits both sets of points.

In reality, the CEF and its regression formulation will involve more than one explanatory variable:

$$E(y|x_1, \ldots, x_K) = \beta_0 + \beta_1 x_1 + \ldots + \beta_K x_K = \beta_0 + \sum_{k=1}^{K} \beta_k x_k. \tag{4}$$

Three justifications

We'll call (2) or (4) the *population regression function (PRF)* and its coefficients *population regression coefficients*, defined as solutions to a minimum MSE problem like (3).

Regression can be justified in three ways:

1. If the CEF is linear, the population regression function (PRF) is it.
2. The PRF provides the best linear predictor of y given the xs.
3. The PRF provides the best linear approximation to the CEF.

The binscatter of `lwage` and `educ` illustrates how regression approximates the CEF even when it is nonlinear.

The slope of the regression line captures $E[E(y|x) - E(y|x-1)]$ or the expected increase in log wages associated with an additional year of schooling.

3.3: Statistical Inference in Regression Analysis

- In the previous section, we built a regression model to predict a response variable y by fitting a straight line model based on *sample* data. This is called the sample regression model, denoted as:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots b_k x_k$$

- Typically however we actually want to make generalizations beyond our sample, which means making conclusions about the wider population from which our sample was taken.

- The least-squares regression model for an entire population can be expressed as:

$$\mu_{y|x} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

where:

$\mu_{y\|x}$	=	the conditional mean of y given the values of $x_1, x_2, \ldots, x_k$
β_0	=	the y-intercept of the population regression equation
$\beta_1, \ldots, \beta_k$	=	the slope(s) of the population regression equation
$x_1, x_2, \ldots, x_k$	=	the independent (explanatory) variable(s)

The standard linear regression model

Using the linear approximation to the CEF, we have the standard expression of the population regression model in terms of y:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_K x_{iK} + u_i, \tag{5}$$

where y is the *dependent variable* or *regressand*, x_k is an *explanatory variable* or *regressor*, u is an *error term* capturing unobserved factors and the index i indicates the unit of observation.

There is nothing special about the notation y, x and β in expressing a regression model. However, we will generally reserve the Latin alphabet for variables and the Greek alphabet for coefficients or parameters. Estimated quantities will be distinguished by a ˆ ("hat") or maybe a ˜("tilde"). From here on, we will include the unit index in our model expressions.

Section 3

Regression Anatomy

The population least-squares problem

Let's start with just one x.

The population regression coefficients, β_0 and β_1, are defined as the solutions to (3). The first-order conditions are:

$$\begin{aligned} E(u_i) &= E(y_i - \beta_0 - \beta_1 x_i) = 0 \\ E(x_i u_i) &= E[x_i(y_i - \beta_0 - \beta_1 x_i)] = 0. \end{aligned} \tag{6}$$

And, the solutions can be expressed as

$$\begin{aligned} \beta_0 &= E(y_i) - \beta_1 E(x_i) \\ \beta_1 &= \frac{\text{cov}(x_i, y_i)}{\text{var}(x_i)}. \end{aligned} \tag{7}$$

Ordinary least-squares estimators

If $\{y_i, x_i\}$ are iid draws in a sample of size N, then β_0 and β_1 can be estimated by simply replacing the population *moments* in (7) with their sample counterparts:

$$\begin{aligned} \hat{\beta}_0 &= \hat{E}(y_i) - \beta_1 \hat{E}(x_i) \\ \hat{\beta}_1 &= \frac{\widehat{\text{cov}}(x_i, y_i)}{\widehat{\text{var}}(x_i)}. \end{aligned} \tag{8}$$

and let the *law of large numbers (LLN)* do its work.

STAT flashback: What are these "sample counterparts" (terms with "hats")? What is the LLN?

While this derivation uses the *method of moments*, $\hat{\beta}_0$ and $\hat{\beta}_1$ are called the ordinary least-squares (OLS) estimators because they solve the sample analog to the least-squares problem in (3).

"Letting the LLN do its work"

We should pause for a second to think about what this means.

Basically, invoking the LLN is claiming that $\widehat{\text{cov}}(x_i, y_i)$ and $\widehat{\text{var}}(x_i)$ converge to their population counterparts (underlying estimands) for increasingly large samples.

An estimator with such a property is *consistent*.

Suffice it to say, if we want to do causal inference with regression, we'll want to claim that OLS consistently estimates the regression coefficients.

We'll have more to say about consistency later.

Interpretation

In a model with just one x, β_1 is measures the change in y *associated* with a unit change in x, holding the unobserved factors constant:

$$\Delta y = \beta_1 \Delta x, \quad \Delta u = 0.$$

Understand there are *no causal claims* here.

If y is the wage rate and x is years of schooling, then β_1 measures the average change in the wage associated with an additional year of schooling.

If y is the log wage, then β_1 measures the average percentage change in the wage associated with an additional year of schooling, or the rate of return.

Wages or log wages?

Table 2: Simple wage and log-wage regressions, Card (1995) sample

	Dependent variable:	
	wage	lwage
	(1)	(2)
educ	29.655***	0.052***
	(1.708)	(0.003)
Constant	183.949***	5.571***
	(23.104)	(0.039)
Observations	3,010	3,010
R^2	0.091	0.099
Note:	*p<0.1; **p<0.05; ***p<0.01	

What changes when we add xs?

Consider the multiple regression model in (5). Now the population least-squares problem is

$$\min E[(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_K x_{iK})^2], \tag{9}$$

where the solutions, $\beta_0, \ldots, \beta_K$, satisfy the first-order conditions,

$$\begin{aligned} E(u_i) &= 0 \\ E(x_{ik} u_i) &= 0. \end{aligned} \tag{10}$$

It turns out that solutions are conceptually the same. Take β_1, for example:

$$\beta_1 = \frac{\text{cov}(\hat{\tilde{x}}_{i1}, y_i)}{\text{var}(\hat{\tilde{x}}_{i1})}, \tag{11}$$

where $\tilde{x}_{i1}$ is a *residualized* version of x_{i1}.

Frisch-Waugh-Lovell theorem

The concept of a "residualized" x is the heart of the Frisch-Waugh-Lovell (FWL) theorem. This result makes explicit what "hold constant" or "control for" other factors means.

A residualized x is that part of the variable left over after you take into account the other xs. More formally, it is the residual in a regression of x_{i1} on all of the other explanatory variables:

$$x_{i1} = \alpha_0 + \alpha_1 x_{i2} + \cdots + \alpha_{K-1} x_{iK} + \tilde{x}_{i1} \tag{12}$$

where $\tilde{x}_{i1}$ denotes the residualized x_{i1}.

The FWL theorem says that the coefficient of $\tilde{x}_{i1}$ in a bivariate regression of y_i on $\tilde{x}_{i1}$ is the same as the coefficient of x_{i1} in a multiple regression of y_i on $(x_{i1}, \ldots, x_{iK})$.

OLS in general

Again, if $\{y_i, x_{i1}, \ldots, x_{iK}\}$ are iid in a sample of size N, then the OLS estimators can be obtained by simply replacing the population *moments* in with their sample counterparts. For β_1 in (11), this gives:

$$\hat{\beta}_1 = \frac{\widehat{\text{cov}}(\hat{\tilde{x}}_{i1}, y_i)}{\widehat{\text{var}}(\hat{\tilde{x}}_{i1})} = \frac{\sum_i \hat{\tilde{x}}_{i1} y_i}{\sum_i \hat{\tilde{x}}_{i1}^2}. \tag{13}$$

Of course the expression in (13) generalizes to any slope coefficient. Just replace "1" with "k" and think about residualizing in terms of all of the other xs.

Interpretation with controls

In a model with controls (other xs), β_1 is measures the change in y associated with a unit change in unit change in x_1 holding $x_2, \ldots, x_K$ and u constant:

$$\Delta y = \beta_1 \Delta x_1, \quad \Delta x_2 = \ldots = \Delta x_K = \Delta u = 0.$$

If y is the log wage, then β_1 measures the average rate of return to another year of schooling, controlling for experience, race etc.

You might say that β_1 is represents the *partial effect* of x_1, but care should be taken with such language so as not to unjustifiably ascribe causality. We'll have more to say about this later.

Demonstrating the FWL theorem

```
# Regress lwage on educ and control variables.
card_tab2 <- lm(lwage ~ educ + exper + expersq + black + south + smsa
                      + reg661 + reg662 + reg663 + reg664 + reg665
                      + reg666 + reg667 + reg668 + smsa66, card)

# Regress educ on the control variables.
educ_reg <- lm(educ ~ exper + expersq + black + south + smsa
                    + reg661 + reg662 + reg663 + reg664 + reg665
                    + reg666 + reg667 + reg668 + smsa66, card)

# Compute residualized educ and lwage variables.
educ_tilde <- resid(educ_reg)

# Regress lwage on residualized educ.
card_tab2_fwl <- lm(lwage ~ educ_tilde, card)
```

Table 2, Column (2) two ways

Table 3: Card Table 2, Column (2) replication

	(1)	(2)
educ	0.075***	
	(0.003)	
educ_tilde		0.075***
		(0.004)
Constant	4.739***	6.262***
	(0.072)	(0.008)
Controls	Yes	Yes
Observations	3,010	3,010
R^2	0.300	0.107
Note:	*p<0.1; **p<0.05; ***p<0.01	

Estimator = Estimand + Sampling Error

We can write any *estimator* in terms of the underlying *estimand* and *sampling error*. For the OLS estimator of any β_k, this decomposition is

$$\hat{\beta}_k = \frac{\sum_i \hat{\hat{x}}_{ik} y_i}{\sum_i \hat{\hat{x}}_{ik}^2} = \underbrace{\beta_k}_{\text{estimand}} + \underbrace{\frac{\sum_i \hat{\hat{x}}_{ik} u_i}{\sum_i \hat{\hat{x}}_{ik}^2}}_{\text{sampling error}} . \tag{14}$$

Of course the estimand is the unobserved thing we are trying to learn about, say the partial of effect of x_k on y.

The sampling error is the direct result of observing a sample and not the population. In repeated sampling, the *estimate* produced by the estimator will vary from the estimand, reflecting the natural variation that distinguishes one sample from another. Estimators whose estimates are closer together in repeated sampling we say are more *precise*.

Unbiasedness

We say an estimator is *unbiased* if its sampling error is zero on average.

Formally, unbiasedness means that the expected value of the estimator equals the estimand:

$$E(\hat{\beta}_k \mid x_{i1}, \ldots x_{iK}) = \beta_k + \frac{\sum_i \hat{\tilde{x}}_{ik} E(u_i | x_{i1}, \ldots x_{iK})}{\sum_i \tilde{x}_{ik}^2} = \beta_k, \tag{15}$$

which will be true only if the errors are *mean independent* of the xs,

$$E(u_i | x_{i1}, \ldots x_{iK}) = 0. \tag{16}$$

We will see that something like the condition in (16) will be required for causal claims. We will find that such a claim is often hard to justify.

Omitted variable bias (OVB)

Let's say to estimate the effect of x_1 on y we should control for x_2,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i,$$

but we don't. If x_2 is omitted, the OLS estimator of β_1 will be

$$\begin{aligned} \tilde{\beta}_1 &= \frac{\sum_i (x_{i1} - \bar{x}_1) y_i}{\sum_i (x_{i1} - \bar{x}_1)^2} \\ &= \underbrace{\beta_1}_{\text{estimand}} + \underbrace{\beta_2 \frac{\sum_i (x_{i1} - \bar{x}_1) x_{i2}}{\sum_i (x_{i1} - \bar{x}_1)^2}}_{\text{bias}} + \underbrace{\frac{\sum_i (x_{i1} - \bar{x}_1) u_i}{\sum_i (x_{i1} - \bar{x}_1)^2}}_{\text{sampling error}}. \end{aligned} \tag{17}$$

Because $\text{cov}(x_{i1}, u_i) = 0$ by (10), the OVB formula is:

$$E(\tilde{\beta}_1) = \beta_1 + \beta_2 \frac{\sum_i (x_{i1} - \bar{x}_1) x_{i2}}{\sum_i (x_{i1} - \bar{x}_1)^2} = \beta_1 + \beta_2 \tilde{\delta}. \tag{18}$$

Signing the bias

Using the OVB formula, $\tilde{\beta}_1$ is biased unless

1. $\beta_2 = 0$ (x_2 did not belong in the model).
2. $\tilde{\delta} = 0$ (x_1 and x_2 are uncorrelated).

It is difficult to sign the bias in models with more explanatory variables because multiple correlations are involved. In general, OVB will affect all of the coefficient estimates, even those not directly correlated with the omitted variable.

The simple OVB formula is nevertheless helpful in sorting out the consequences of unobservables on regression estimates.

Effect of omitting experience

Table 4: Log wage regressions, without and with experience

	(1)	(2)
educ	0.052***	0.093***
	(0.003)	(0.004)
exper		0.041***
		(0.002)
Constant	5.571***	4.666***
	(0.039)	(0.064)
Observations	3,010	3,010
R^2	0.099	0.181
Note:	*p<0.1; **p<0.05; ***p<0.01	

Regression fit

Regression splits the dependent variable into two parts:

$$y_i = \hat{y}_i + \hat{u}_i,$$

where the $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_K x_{iK}$ are the *fitted values* and $\hat{u}_i = y_i - \hat{y}_i$ are the *residuals*. Fit is about how well $\hat{y}_i$ accounts for or "explains" the variation in y_i. Define

$$\underbrace{SST = \sum_i (y_i - \overline{y})^2}_{\text{Total}}, \quad \underbrace{SSE = \sum_i (\hat{y}_i - \overline{y})^2}_{\text{Explained}}, \quad \underbrace{SSR = \sum_i \hat{u}_i^2}_{\text{Residual}}.$$

Because $SST = SSE + SSR$, we can define the goodness-of-fit measure:

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}. \tag{19}$$

For causal inference, R^2 will not be of primary importance.

Section 4

Regression with Logs, Quadratics and Dummies

Functional forms involving logs

A *log-level* model reflects an exponential relationship between y and x:

$$y_i = \exp(\beta_0 + \beta_1 x_i + u_i) \quad \Rightarrow \quad \log y_i = \beta_0 + \beta_1 x_i + u_i$$

In this case, β_1 is interpreted as a *rate of return*:

$$\beta_1 = \frac{d \log y_i}{dx_i} = \frac{dy_i / y_i}{dx_i}.$$

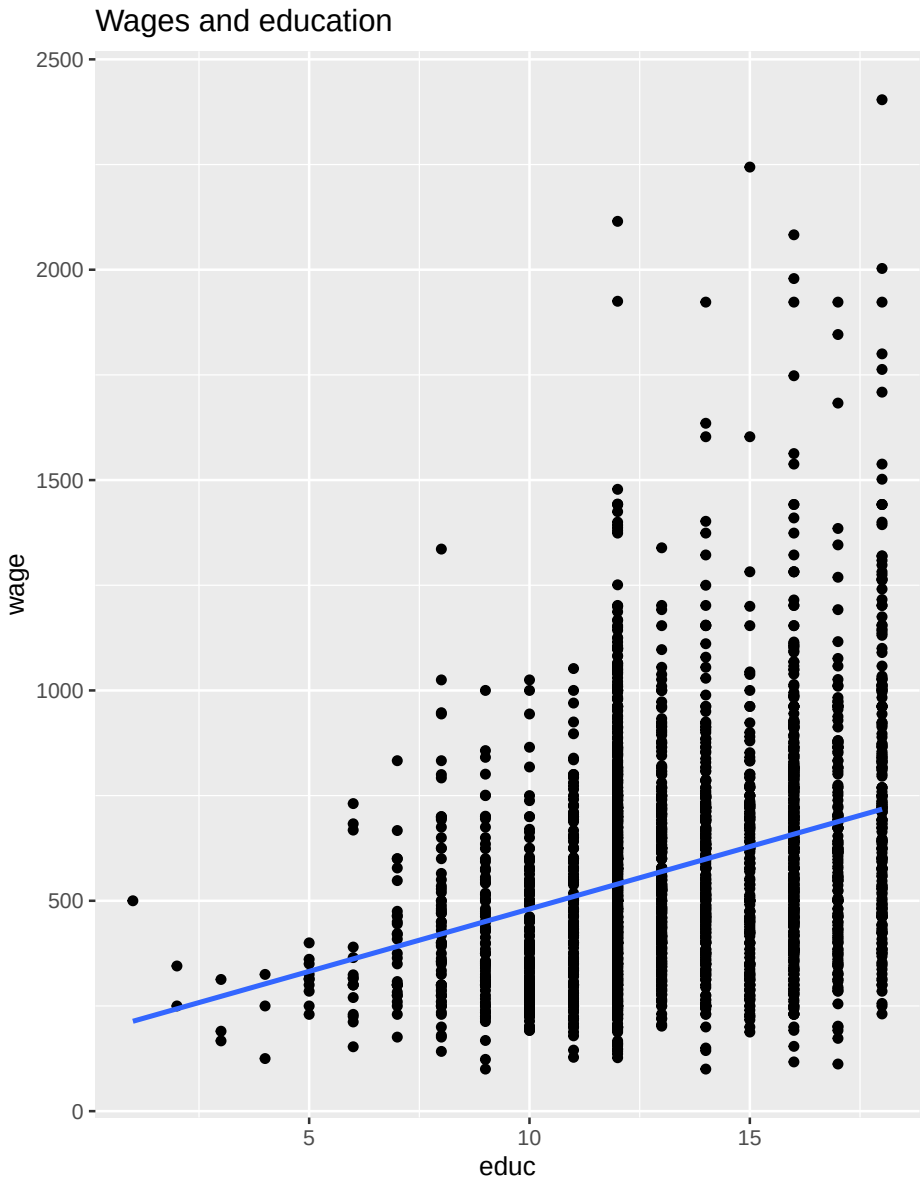
A *log-log* model is implied by the following specification for y:

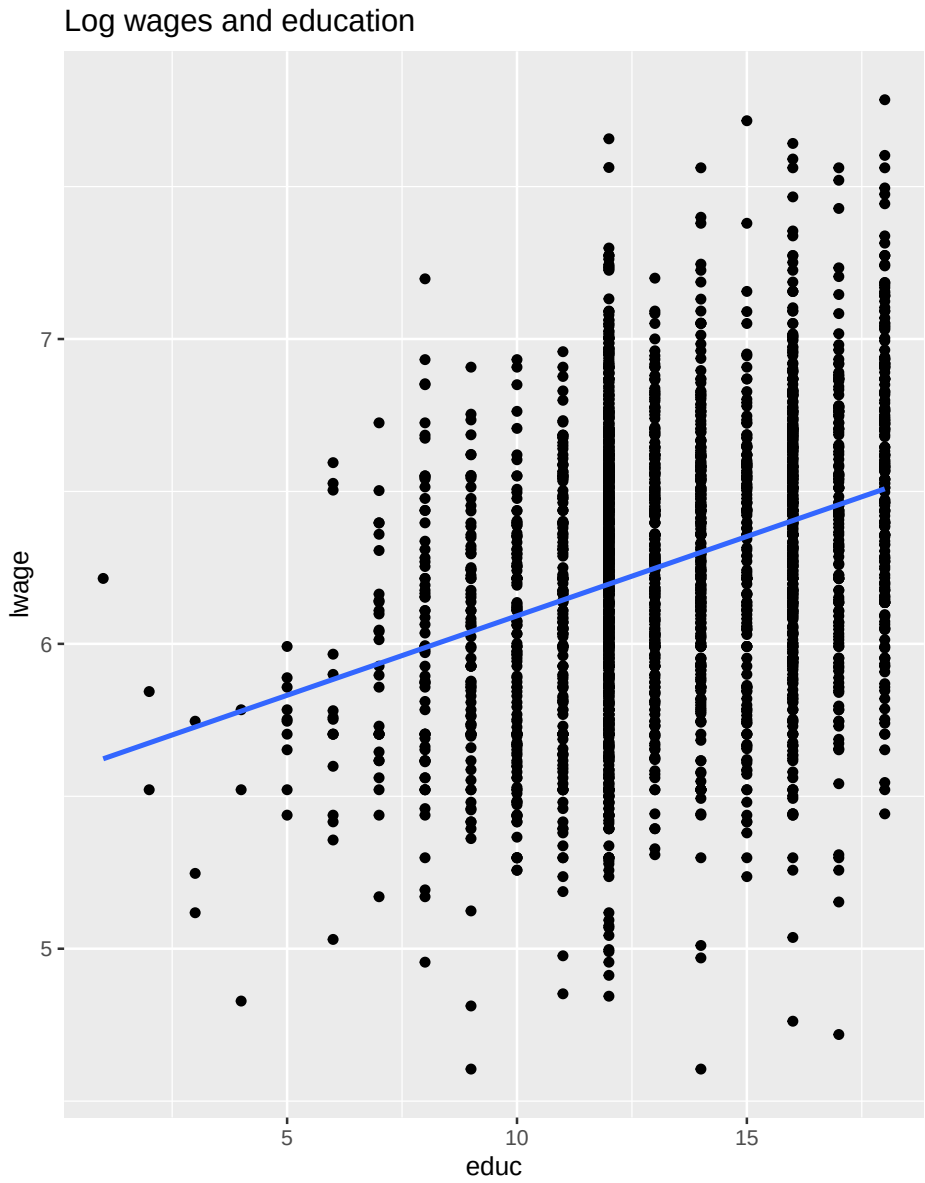
$$y_i = \beta_0 \, x_i^{\beta_1} \exp(u_i) \quad \Rightarrow \quad \log y_i = \log \beta_0 + \beta_1 \log x_i + u_i$$

In this case, β_1 is interpreted as an *elasticity*:

$$\beta_1 = \frac{d \log y_i}{d \log x_i} = \frac{dy_i / y_i}{dx_i / x_i} = \frac{dy_i}{dx_i} \frac{x_i}{y_i}.$$

Wages and log wages again





Interpretation summary for models with logs

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	$\log(y)$	$\log(x)$	$\%\Delta y = \beta_1 \%\Delta x$

Source: Wooldridge

Rules of thumb for specifying variables in log form

Consider using the log form when the variable is a positive monetary value (wages, salaries, sales, market value, etc) or typically a large integer value (population, employment, enrollment, etc).

Variables measured in years (education, experience, tenure, etc) are generally used in their level form.

Variables that are measured in shares, fractions and percentages (enrollment rate, arrest rates, unemployment rate, etc) can appear in either form. Note, however, that marginal effects in the levels involve *percentage point* changes, while marginal effects in the logs involve *percentage* changes.

Log transformations should not be used for variables that can take on values less than or equal to zero.

Quadratics, interactions and average partial effects

Consider a regression model with linear and quadratic terms:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u.$$

The effect of a unit change in x is

$$\frac{dy}{dx} = \beta_1 + 2\beta_2 x.$$

Interactions are evaluated similarly. Consider

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + u.$$

The partial effect of x_1 is

$$\frac{dy}{dx} = \beta_1 + \beta_3 x_2$$

In both cases, it is common to report the *average partial effect (APE)*. For example, in the interaction case:

$$APE_{x_1} = \hat{\beta}_1 + \hat{\beta}_3 \overline{x}_2.$$

Returns to experience

Table 5: Returns to experience, Card (1995) sample

	(1)	(2)
Education	0.052	0.075
	(0.003)	(0.003)
Experience		0.085
		(0.007)
Experience^2		-0.002
		(0.000)
Black		-0.199
		(0.018)
South		-0.148
		(0.026)
SMSA		0.136
		(0.020)
Constant	5.571	4.739
	(0.039)	(0.072)
N	3010	3010
R^2	0.10	0.30

The return to the 1st, 5th and 10th year of experience

```
bhat <- coef(card_tab2)
bhat["exper"] + 2*bhat["expersq"]*1
```

```
  exper
```

```
0.08025795
```

```
bhat["exper"] + 2*bhat["expersq"]*5
```

```
  exper
```

```
0.06196163
```

```
bhat["exper"] + 2*bhat["expersq"]*10
```

```
  exper
```

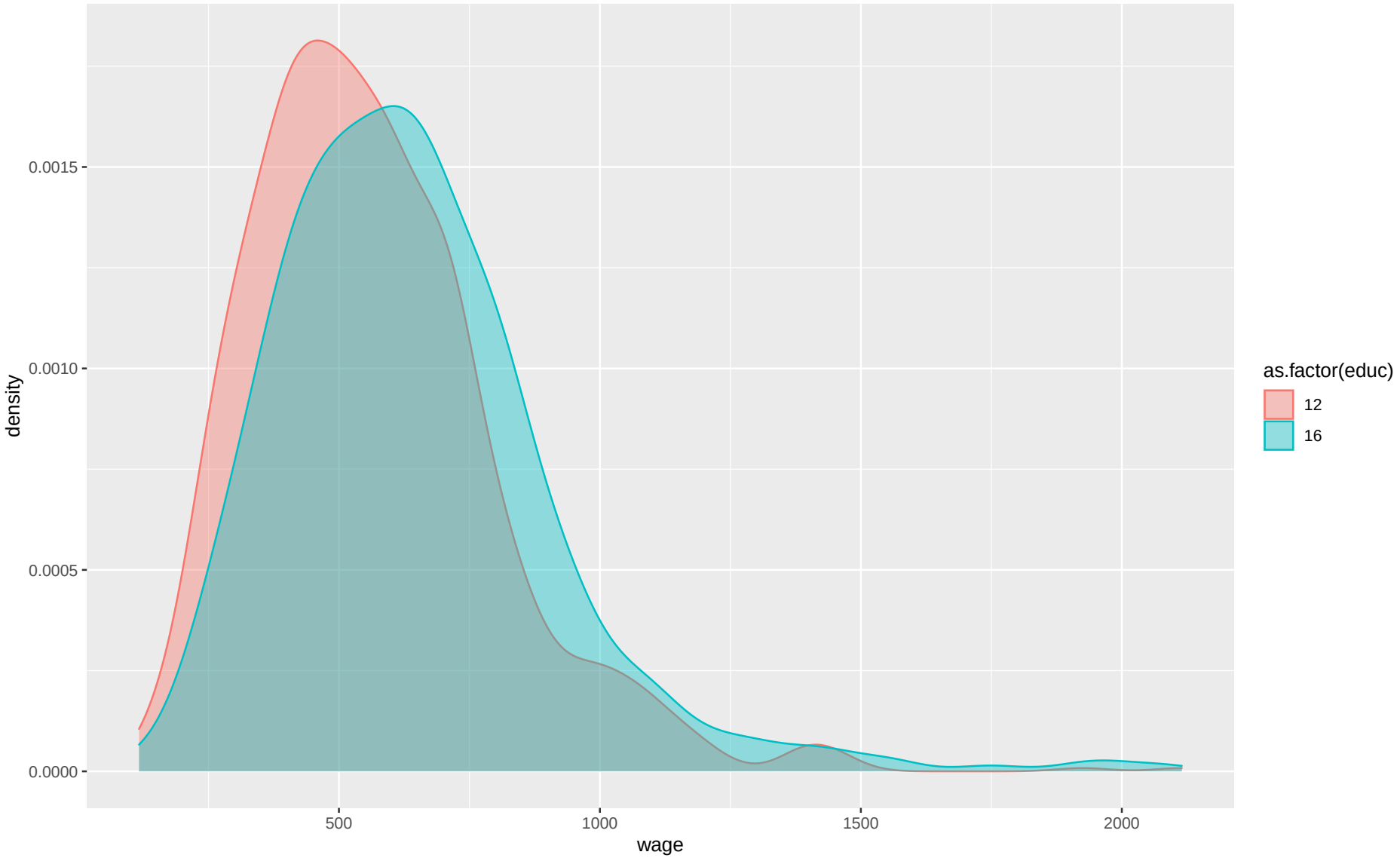
```
0.03909122
```

Categorical *x*s

Type	Examples
Individual characteristics	race, gender, education
Firm characteristics	industry, ownership
Program evaluation	policy indicator, program participant
Panel/multilevel data	"fixed effects"

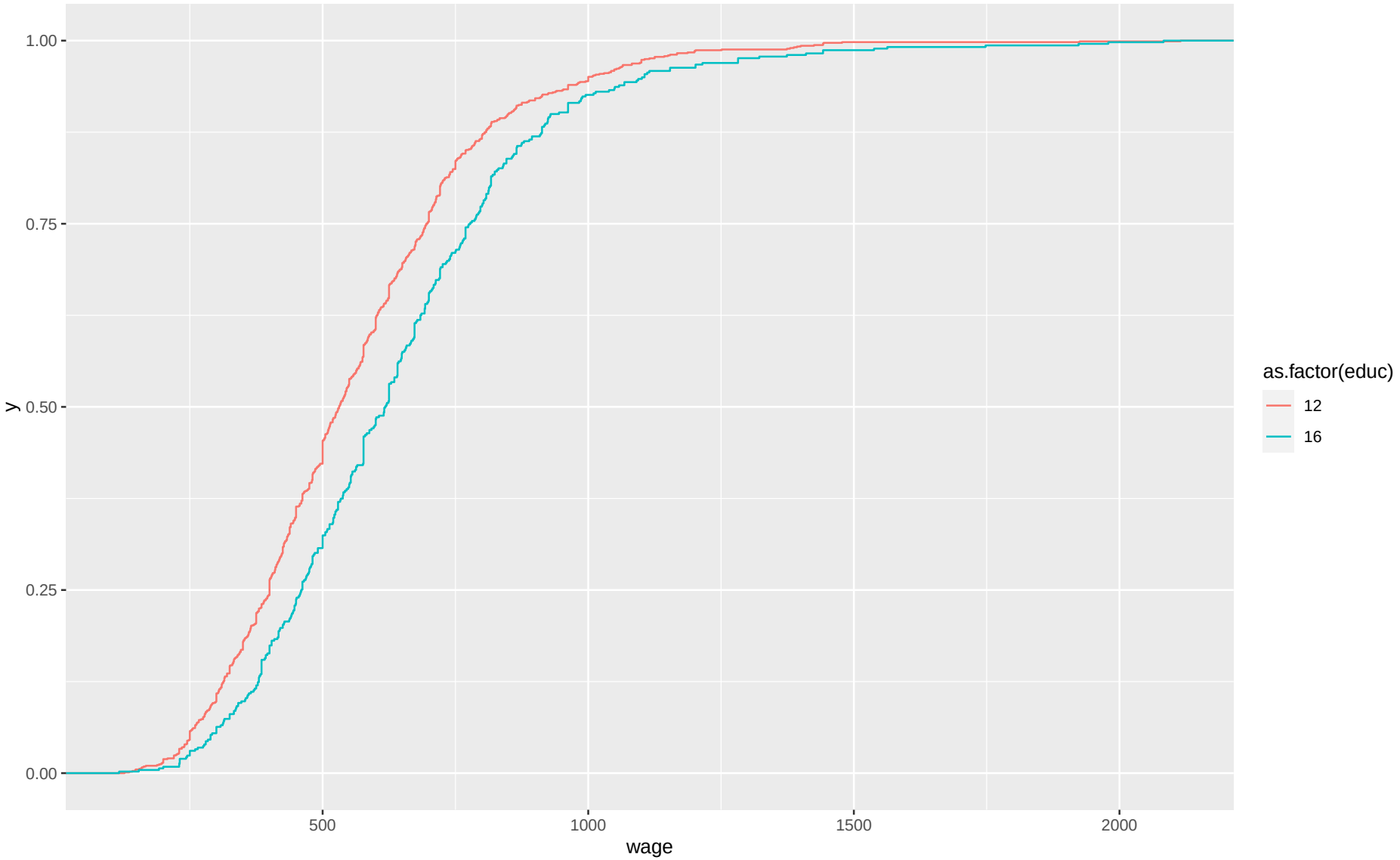
Empirical wage distributions conditional on education

Figure 1. Empirical distribution of wages by educational attainment



Empirical wage cdfs conditional on education

Figure 2. Empirical CDF of wages by educational attainment



Interpreting the coefficient of a single dummy

The single dummy variable has the effect of an *intercept shifter*.

To see this, consider the log wage regression

$$lwage = \beta_0 + \beta_1 \, exper + \delta_0 \, BA + u,$$

where $BA = 1$ if an individual has at least a college degree and 0 if they are only a high-school graduate. Then, δ_0 is the difference in conditional means:

$$\delta_0 = E(lwage \,|\, exper, BA = 1) - E(lwage \,|\, exper, BA = 0)$$

Because wages are expressed in logs, δ_0 is the approximate average percentage difference in wages. The exact percentage difference is obtained by exponentiating δ_0:

$$\frac{wage(1) - wage(0)}{wage(0)} = \exp(\delta_0) - 1.$$

Interactions

By *interacting* experience with the BA *indicator*, we can allow the returns to experience to differ between college and high-school graduates:

$$\begin{aligned} lwage &= \beta_0 + \beta_1\, exper + \delta_0\, BA + \delta_1\, exper \cdot BA + u \\ &= \beta_0 + \delta_0\, BA + (\beta_1 + \delta_1\, BA) exper + u. \end{aligned}$$

Table 7: Returns to experience, educ=12 & 16

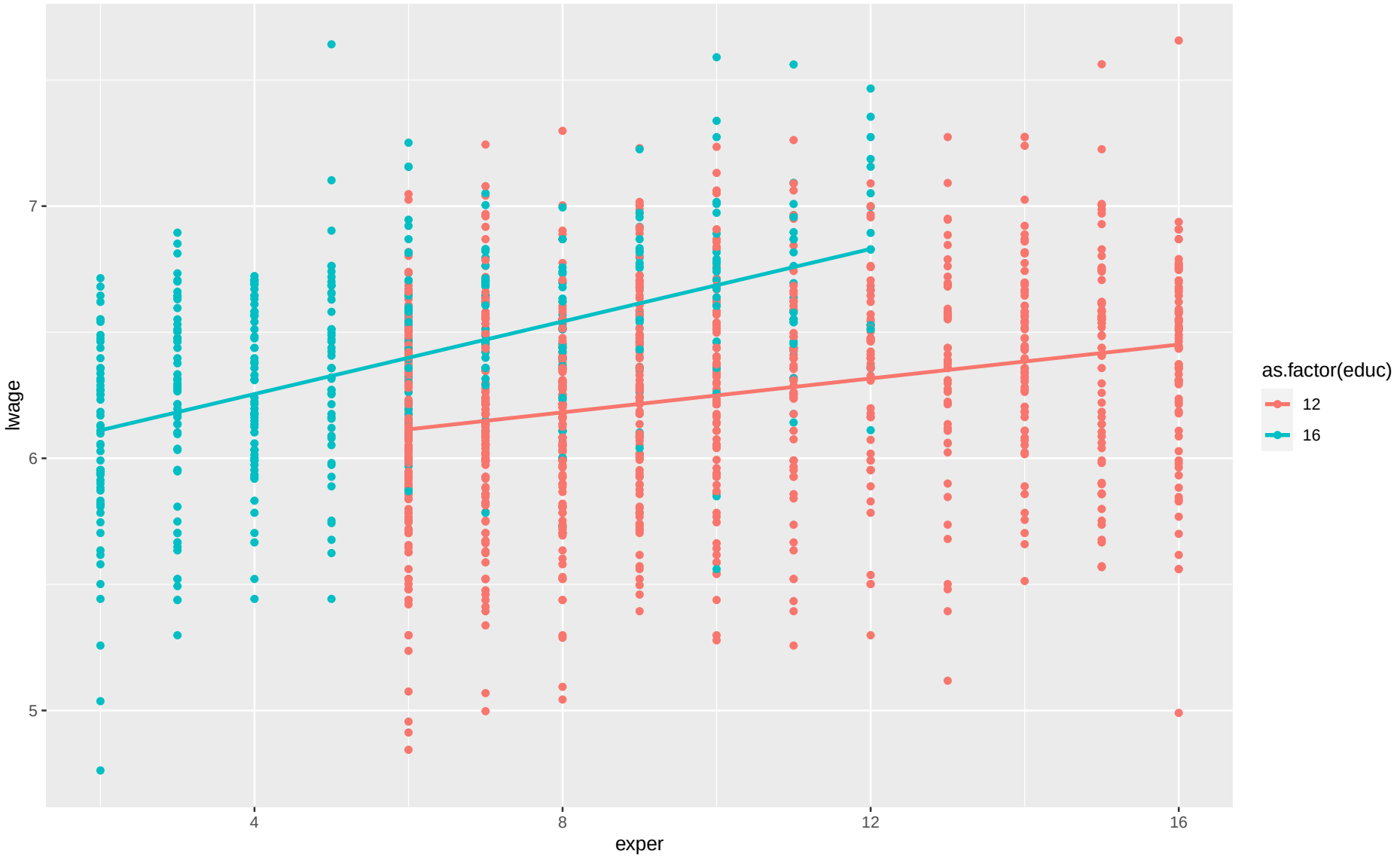
	(1)	(2)	(3)
exper	0.034***	0.072***	0.034***
	(0.004)	(0.006)	(0.004)
factor(educ)16			0.054
			(0.057)
exper:factor(educ)16			0.038***
			(0.007)
Constant	5.913***	5.967***	5.913***
	(0.042)	(0.037)	(0.040)
Observations	992	459	1,451
R^2	0.066	0.264	0.146
Note:			*p<0.1; **p<0.05; ***p<0.01

Plotting education-specific, log wage-experience regressions

```
ggplot(data = subset(card, educ==12 | educ==16),
       aes(x=exper, y=lwage, col=as.factor(educ))) +
  geom_point() +
  geom_smooth(data =  subset(card, educ==12),
              method ='lm',
              formula = y ~ x,
              se=FALSE) +
  geom_smooth(data =  subset(card, educ==16),
              method ='lm',
              formula = y ~ x,
              se=FALSE)
  labs(title="Figure 2. Returns to experience, educ=12 & 16")
```

Returns to experience by education visualized

Figure 2. Returns to experience, educ=12 & 16



Multiple categories

What if there are more than 2 education categories? Let

$$educ = \begin{cases} 2, \text{if BA} \\ 1, \text{if HSG} \\ 0, \text{if less than HSG} \end{cases}$$

How would you interpret δ_0 is this case? What's the problem here?

$$\begin{aligned} E(lwage \,|\, educ = 2) &= \beta_0 + \beta_1 \, exper + 2\delta_0 \\ E(lwage \,|\, educ = 1) &= \beta_0 + \beta_1 \, exper + \delta_0 \\ E(lwage \,|\, educ = 0) &= \beta_0 + \beta_1 \, exper \end{aligned}$$

Always allow for separate category effects (intercepts):

$$lwage = \beta_0 + \beta_1 \, exper + \delta_0 \, BA + \delta_1 \, HSG + u.$$

And, don't fall into the *dummy variable trap*.

What about dummy outcomes?

Suppose y is binary. Does that change anything? It depends. If you are interested in causal inference the answer is **no**.

The CEF is still

$$E(y|x_1, \ldots, x_K) \equiv E(y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K,$$

and regression provides the best linear approximation to the CEF. In addition, it is always the case that the *response probability* is the same as the CEF,

$$P(y = 1|\mathbf{x}) = E(y|\mathbf{x}) \equiv p(\mathbf{x}),$$

and

$$\beta_k = \frac{\partial P(y = 1|\mathbf{x})}{\partial x_k}.$$

Issues with LPMs

However, there are two well-known issues with linear probability models (LPMs):

1. LPMs are *heteroscedastic* by construction.
   - Because y is Bernoulli, $\text{var}(y|\mathbf{x}) = p(\mathbf{x})[1 - p(\mathbf{x})]$.
   - So always report robust se's. More about this later.
2. Predicted response probabilities may fall outside the unit interval.
   - You can still report the *percentage correctly predicted*, defining

$$\tilde{y}_i = \begin{cases} 1, \text{if } \hat{y}_i \geq .5 \\ 0, \text{if } \hat{y}_i < .5 \end{cases}$$

   - Or, estimate a true probability model.

If you are primarily interested in prediction, the second issue should probably move you toward a logit or probit model.

Section 5

Basic Inference

Sampling distribution

To do inference, we need the *sampling distribution* of the OLS estimator.

The sampling distribution of $\hat{\beta}_k$ is the distribution we would get if we repeatedly computed the estimator an infinite number of times.

If we knew the distribution of the underlying population we were sampling from, we could say for sure what it is, but we never know that.

So, we rely on the central limit theorem (CLT) to *approximate* it.

The variance of the sampling distribution measures how close $\hat{\beta}_k$ is to its average value, or, the underlying estimand, if $\hat{\beta}_k$ is unbiased.

The *standard error* of $\hat{\beta}_k$ is the standard deviation of its sampling distribution.

"Classical" assumptions from BUSN 4000

6.2: Assumptions of the Multiple Linear Regression Model

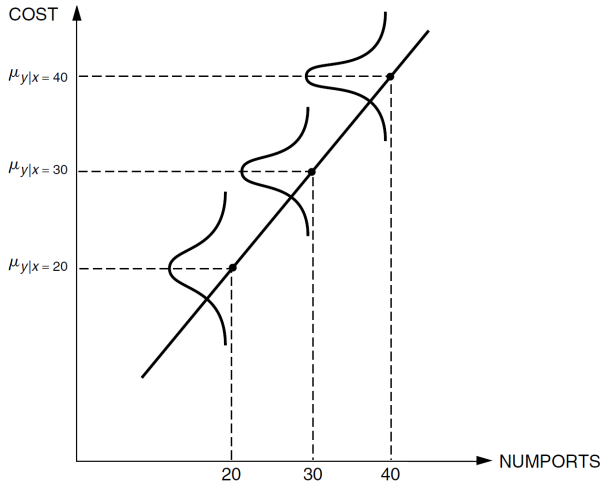
- To allow statistical inference from a sample to the population, some assumptions about the population regression line are necessary.

1. The expected value of the disturbances is zero: $E(e_i) = 0$. This implies that the regression line passes through the conditional means of the x variable.

2. The variance of each e_i is equal to σ_e^2. This means that each of the distributions along the regression line has the same variance regardless of the value of x.

3. The e_i are normally distributed.

4. The e_i are independent. This is an assumption that is most important when data are gathered over time. When the data are cross-sectional (that is, gathered at the same point in time for different individual units), this is typically not an assumption of concern.



COVERED VIA VIDEO

Deriving the sampling variance of the OLS estimator

To derive the sampling variance, we start with (14), where we decomposed $\hat{\beta}_k$ in terms β_k and the sampling error:

$$\hat{\beta}_k = \frac{\sum_i \hat{\hat{x}}_{ik} y_i}{\sum_i \hat{\hat{x}}_{ik}^2} = \beta_k + \frac{\sum_i \hat{\hat{x}}_{ik} u_i}{\sum_i \hat{\hat{x}}_{ik}^2}. \tag{20}$$

Because β_k is a constant and the sampling error has a zero mean,

$$\text{var}(\hat{\beta}_k) = \text{var}\left[\frac{\sum_i \hat{\hat{x}}_{ik} u_i}{\sum_i \hat{\hat{x}}_{ik}^2}\right] = E\left[\left(\frac{\sum_i \hat{\hat{x}}_{ik} u_i}{\sum_i \hat{\hat{x}}_{ik}^2}\right)^2\right]. \tag{21}$$

From this point, there are two ways to go: (a) the way toward the default/classical inference provided `lm`, which is *wrong*, and (b) the modern way, which is *right*.

Assume homoscedasticity: the default (wrong) way

To assume *homoscedasticity* is to claim that the variance of regression error, u, is unrelated to the xs:

$$\text{var}(u_i|x_{i1},\ldots,x_{ik}) = E(u_i^2|x_{i1},\ldots,x_{ik}) = \sigma^2. \tag{22}$$

Given (22), the sampling variance of the OLS estimator is

$$\text{var}(\hat{\beta}_k) = \frac{\sigma^2}{\sum_i \hat{\tilde{x}}_{ik}^2} = \frac{\sigma^2}{SST_k(1-R_k^2)}, \tag{23}$$

where R_k^2 is the R^2 from the regression of x_k on the other xs. The square root of the sampling variance is the estimator's *standard error*:

$$\text{se}(\hat{\beta}_k) = \frac{\sigma}{\sqrt{\sum_i \hat{\tilde{x}}_{ik}^2}}. \tag{24}$$

Estimating σ^2

To operationalize the formula in (26), we need an estimator of σ^2. The right choice is

$$\hat{\sigma}^2 = \frac{\sum_i \hat{u}_i^2}{N - K - 1} = \frac{SSR}{N - K - 1}. \tag{25}$$

Simply averaging the *SSR* might be more intuitive, but the *degrees of freedom* correction is required for unbiasedness because $E(SSR) = (N - K - 1)\,\sigma^2$.

What does "degrees of freedom" refer to anyway?

With (25), we get the default standard errors calculated by `lm`:

$$\text{se}(\hat{\beta}_k) = \frac{\hat{\sigma}}{\sqrt{\sum_i \hat{\hat{x}}_{ik}^2}}. \tag{26}$$

Don't assume homoscedasticity: the modern (right) way

Homoscedasticity is rarely justified. If you are wrong about it, your standard errors will be biased and the inference they are based on invalid.

The good news is that it is easy to compute standard errors that are *robust* to *heteroscedasticity*.

The basic strategy is to replace the population quantities in (21) with their sample counterparts like this:

$$\widehat{\text{var}}(\hat{\beta}_k) = \frac{\sum_i \hat{\tilde{x}}_{ik}^2 \hat{u}_i^2}{(\sum_i \hat{\tilde{x}}_{ik}^2)^2}. \tag{27}$$

The standard errors based on this variance estimator are robust in the sense they are consistent whether or not the errors are homoscedastic.

These standard errors, along with a number of tweaked alternatives, are offered as options in most regression packages. To learn more about the R implementation, see the `sandwich` vignette.

Asymptotically valid inference

For causal inference, we only need OLS to have two properties: *consistency* and *asymptotic normality*.

As we hinted at earlier, consistency is linked to the LLN.

Asymptotic normality follows from an application of the CLT. To say that $\hat{\beta}_k$ is asymptotically normal roughly means that its sampling distribution can be *approximated* by the normal with a large enough sample.

Under random sampling, its not hard to establish both properties for $\hat{\beta}_k$.

Consistency

We say $\hat{\beta}_k$ is *consistent* if the OLS estimator converges to β_k when computed with increasingly large samples.

This would be true if its sampling error vanished as sample size grew. The argument goes like this: by the LLN,

$$\frac{1}{N}\sum_i \hat{\tilde{x}}_{ik} u_i \to E(\hat{\tilde{x}}_{ik} u_i)$$

$$\frac{1}{N}\sum_i \hat{\tilde{x}}_{ik}^2 \to \text{var}(\hat{\tilde{x}}_{ik}),$$

as $N \to \infty$, and by (10), $E(\hat{\tilde{x}}_{ik} u_i) = 0$.

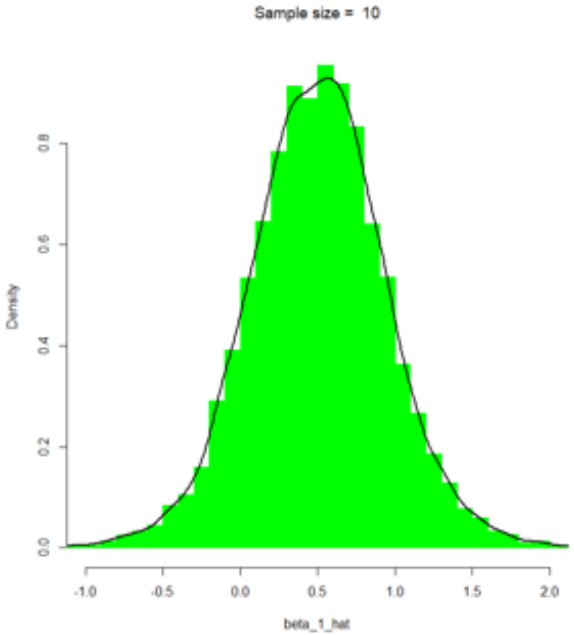
How does this differ from unbiasedness?

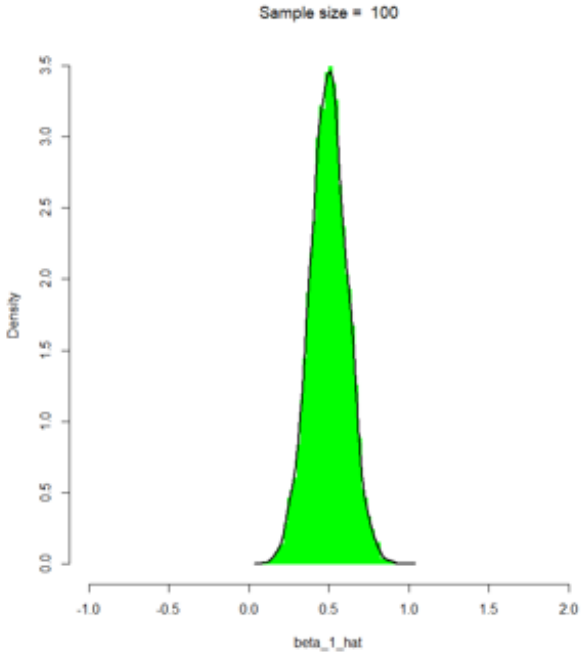
Monte carlo demonstration of consistency and asymptotic normality

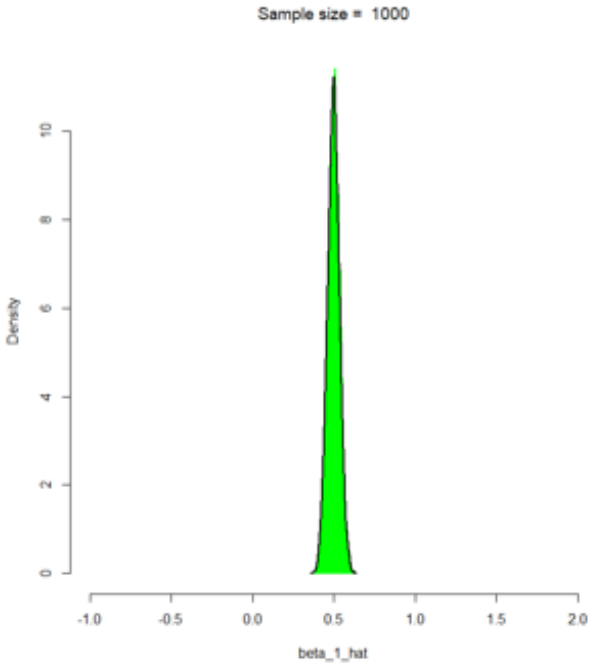
Below we simulate the sampling distribution of the OLS estimator for three cases: $N = (10, 100, 1000)$. Each case involves drawing N xs from a $N(4, 1)$ distribution and us from a *uniform*$(0, 4)$ distribution, constructing y, and estimating a simple regression model. For each N, we conduct $J = 10,000$ trials and construct histograms of the estimates generated from those trials. The true coefficient values in the simulations are $\beta_0 = 1$ and $\beta_1 = .5$.

```
set.seed(1234567)
beta_1_hat <- numeric(10000)
beta_0 <- 1
beta_1 <- .5
for(n in c(10,100,1000)) {
  for(j in 1:10000) {
    x <- rnorm(n,4,1)
    u <- runif(n,0,4)
    y <- beta_0 + beta_1*x + u
    beta_hat <- coef(lm(y~x))
    beta_1_hat[j] <- beta_hat["x"]
  }
  hist(beta_1_hat, freq=FALSE, col="green", border="green",
       xlim = c(-1,2), breaks = 40,
        main = bquote("Sample size = "~.(n)))
  lines(density(beta_1_hat, kernel=c("gaussian")), lwd=2)
  rootN_beta_1_hat_err <- sqrt(n)*(beta_1_hat - .5)
  hist(rootN_beta_1_hat_err, freq=FALSE, col="red", border="red",
       xlim = c(-4,4), breaks = 40,
       main = bquote("Sample size = "~.(n)))
  lines(density(rootN_beta_1_hat_err, kernel=c("gaussian")), lwd=2)
}
```

Consistency simulated







Asymptotic normality

What does it mean to say β_k's sampling distribution can be *approximated* by the normal with a large enough sample?

We won't get much insight from looking at the distribution of $\hat{\beta}_k$ as $N \to \infty$ because the sampling distribution collapses to β_k.

Instead, we focus on a *standardized* version of $\hat{\beta}_k$, $\sqrt{N}(\hat{\beta}_k - \beta_k)$. Multiplying by $\sqrt{N}$ guarantees that the variance of $\hat{\beta}_K$ does not vanish.

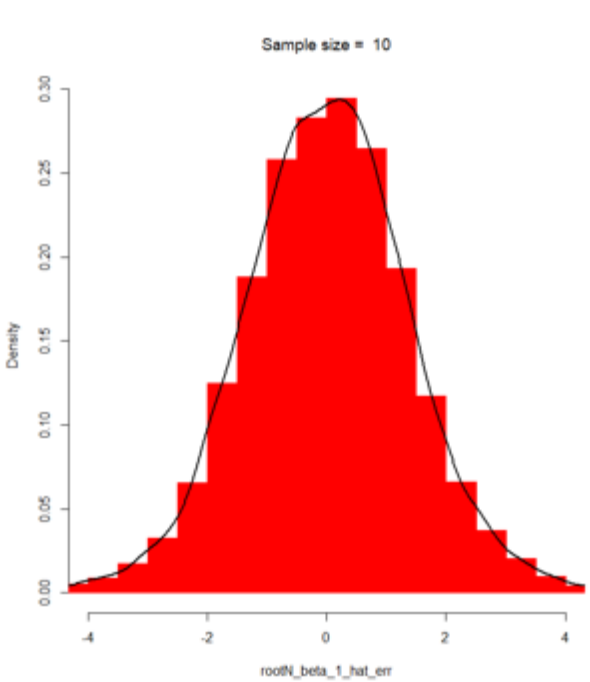
The CLT says that under random sampling,

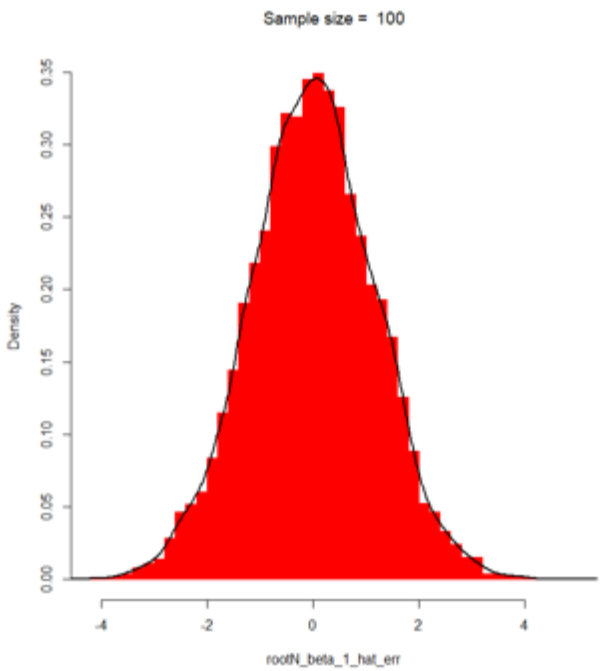
$$\sqrt{N}(\hat{\beta}_k - \beta_k) \xrightarrow{d} N\Big(0,\, \text{var}(\hat{\beta}_k)\Big),$$

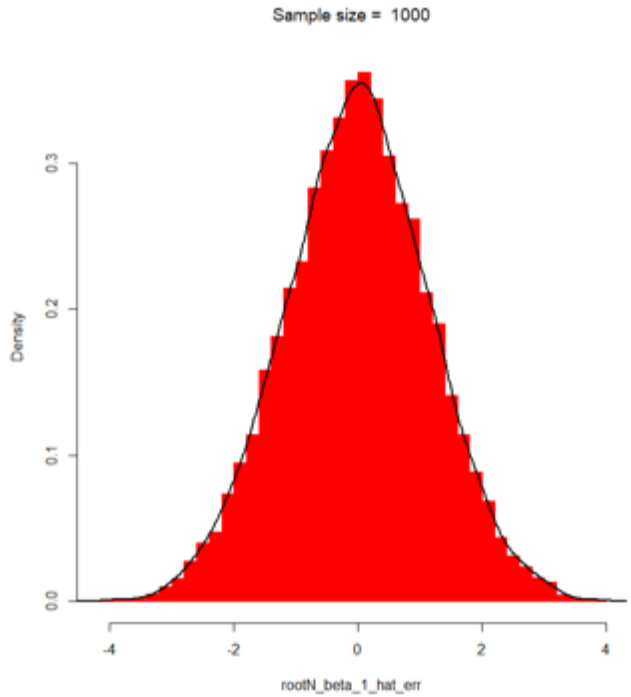
which implies that we can treat $\hat{\beta}_k$ as asymptotically normal,

$$\hat{\beta}_k \overset{A}{\sim} N\Big(\beta_k,\, \frac{1}{N} \cdot \text{var}(\hat{\beta}_k)\Big). \tag{28}$$

Asymptotic normality simulated







Asymptotically valid CIs

Given the asymptotic distribution of $\hat{\beta}_k$ and a consistent estimator of $\text{var}(\hat{\beta}_k)$, we can state

$$P\left(-z_{(1-\alpha/2)} < \frac{\hat{\beta}_k - \beta_k}{\text{se}(\hat{\beta}_k)} < z_{(1-\alpha/2)} \right) = 1 - \alpha$$

where z is a critical value from the standard normal distribution.

Thus, $(1-\alpha)$ CI for β_k is

$$\hat{\beta}_k \pm z_{(1-\alpha/2)} \, \text{se}(\hat{\beta}_k). \tag{29}$$

Hypothesis testing

Statistical hypothesis tests are basically just CIs viewed from another angle.

Statistical tests involve two hypotheses: the null (H_0) and the alternative (H_1). The null is the hypothesis of interest, as it specifies the value of the parameter being tested.

When we perform hypothesis tests, we can go wrong in two ways:

Error Type	Explanation
Type I (false positive)	Reject a true null
Type II (false negative)	Failure to reject a false null

We call $P(\text{Type I})$ the *significance level*, which is commonly denoted by α. So, to be precise

$$\alpha = P(\text{reject } H_0 | H_0).$$

If you set $\alpha = .05$ you are accepting a 5% probability of falsely rejecting the null to detect true deviations from it.

Asymptotically valid t tests

Because the standardized version of the OLS estimator is asymptotically standard normal, the usual t tests are asymptotically valid:

$$t = \frac{\hat{\beta}_k - \beta_k}{\text{se}(\hat{\beta}_k)} \xrightarrow{d} N(0,1).$$

The simplest version of the t test in a regression context is the *test of significance*:

$$H_0 : \beta_k = 0$$
$$H_1 : \beta_k \neq 0.$$

The t statistic for this null the ratio of the estimator to its standard error:

$$t = \frac{\hat{\beta}_k - 0}{\text{se}(\hat{\beta}_k)} = \frac{\hat{\beta}_k}{\text{se}(\hat{\beta}_k)}. \tag{30}$$

Asymptotically valid F tests

The same argument can be made for F statistics, which we use to test the joint significance of multiple coefficients.

Suppose H_0 is that some subset q of the β_k are equal to zero. An F test of this null compares the "long" model, which uses all of the xs, to the "short" model, which omits the q of them (because their coefficients are zero under the null).

The non-robust version of the F statistic is

$$F = \frac{(SSR_s - SSR_l)/q}{SSR_l/(N-K-1)} \xrightarrow{d} \chi^2_q/q. \tag{31}$$

The robust version is a little more complicated to write down, but is easily computed with R packages that deliver robust inference.

When $q = 1$, there is a direct relationship between the t and F statistics: $F = t^2$, which implies $t = \sqrt{F}$.

p values

The p value, or *marginal significance level* is the probability of drawing a test statistic at least as adverse to the null as the one you actually calculated, conditional on the null being true

$$p = P(\text{test-stat} \geq \text{test-stat value} | H_0).$$

Put differently, it is the largest significance level at which could conduct the test and still fail to reject the null.

Just as larger t-statistic values provide greater evidence against the null, so do smaller p values.

Economic vs statistical significance

While a large test statistic value speaks to statistical significance, economic significance is tied up in the magnitude of the coefficient estimate.

A t statistic can be large either because the value of $\hat{\beta}_k$ is large or its estimated standard error is small.

Thus, it will be wise to take both in consideration when interpreting the results of statistical inference.

Depending on sample size, a p value larger than .05 may not lead to accepting the null.

Section 6

Assignment 1

B.3. OLS estimation of the return to schooling

Table 9: Estimated returns to schooling, Card (1995) sample

	(1)	(2)	(3)	(4)	(5)
Education	0.052	0.075	0.075	0.076	0.070
	(0.003)	(0.003)	(0.004)	(0.005)	(0.005)
Experience		0.085	0.085	0.092	0.095
		(0.007)	(0.007)	(0.009)	(0.009)
Experience^2		-0.002	-0.002	-0.003	-0.003
		(0.000)	(0.000)	(0.000)	(0.000)
Black		-0.199	-0.199	-0.183	-0.148
		(0.018)	(0.018)	(0.025)	(0.028)
South		-0.148	-0.148	-0.100	-0.100
		(0.026)	(0.028)	(0.036)	(0.036)
SMSA		0.136	0.136	0.125	0.123
		(0.020)	(0.019)	(0.023)	(0.023)
IQ					0.002
					(0.001)
Constant	5.571	4.739	4.739	4.675	4.501
	(0.039)	(0.072)	(0.075)	(0.100)	(0.114)
N	3010	3010	3010	2061	2061
R^2	0.10	0.30	0.30	0.23	0.24

Columns (2)-(4) also include 1966 region and SMSA dummies.
Columns (3)-(5) report robust standard errors.

B.4. Demonstration of the FWL theorem

```
##
## Call:
## lm(formula = lwage_tilde ~ educ_tilde, data = card)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.62326 -0.22141  0.02001  0.23932  1.33340
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.082e-17  6.770e-03     0.0        1
## educ_tilde  7.469e-02  3.490e-03    21.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3714 on 3008 degrees of freedom
## Multiple R-squared:  0.1321, Adjusted R-squared:  0.1319
## F-statistic:   458 on 1 and 3008 DF,  p-value: < 2.2e-16
```