Review of last week

- Expectations and conditional expectations
  - Linear
  - Iterated expectations
- Asymptotics — using large sample distribution to approximate finite sample distribution of estimators
  - LLN: sample moments converge in probability to population moments,

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n} g(x_i)}_{\text{sample moment}} \xrightarrow{p} \underbrace{\mathrm{E}[g(x)]}_{\text{population moment}}$$

  - CLT: centered and scaled sample moments converge in distribution to population moments

$$\underbrace{\sqrt{n}}_{\text{“scaling”}}\left(\frac{1}{n}\sum_{i=1}^{n} g(x_i) \underbrace{-\mathrm{E}[g(x)]}_{\text{“centering”}}\right) \xrightarrow{d} N\left(0, \mathrm{Var}(g(x))\right)$$

  - Using CLT to calculate p-values

Part I
Definition and interpretation of regression

Contents

References

- Main texts:
  - Angrist and Pischke (2014) chapter 2
  - Wooldridge (2013) chapter 2
  - Stock and Watson (2009) chapter 4-5
- More advanced:
  - Angrist and Pischke (2009) chapter 3 up to and including section 3.1.2 (pages 27-40)
  - Bierens (2012)
  - Abbring (2001) chapter 3
  - Baltagi (2002) chapter 3
  - Linton (2017) chapters 16-20, 22
- More introductory:
  - Diez, Barr, and Cetinkaya-Rundel (2012) chapter 7

The most important things to understand are contained in Angrist and Pischke (2014). The most useful references for details are Wooldridge (2013) or Stock and Watson (2009). Angrist and Pischke (2009) is also very nice, but a bit more dense. Diez, Barr, and Cetinkaya-Rundel (2012) is a simple introduction to regression with many examples and not much math. PDFs of Linton (2017) and Baltagi (2002) are available from UBC library. They are more technical and difficult than Wooldridge, but I think would still be useful. Bierens (2012) has many of the proofs that we will go through, but the typesetting is not great. Abbring (2001) moves quickly and uses matrix notation. If you have not taken linear algebra or you are not comfortable with matrix multiplication and inversion (which you do not need to know and will not be covered in this course), then Abbring (2001) may not be useful for you.

1 Motivation

General problem

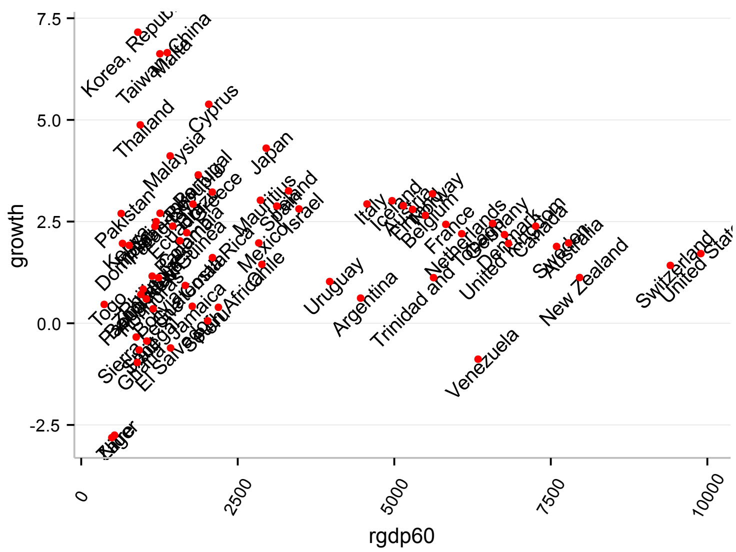
- Often interested in relationship between two (or more) variables, e.g.
  - Wages and education
  - Minimum wage and unemployment
  - Price, quantity, and product characterics
- Usually have:
  1. Variable to be explained (*dependent variable*)
  2. *Explanatory variable(s)* or *independent variables* or *covariates*

Dependent	**Independent**
Wage	Education
Unemployment	Minimum wage
Quantity	Price and product characteristics
Y	X

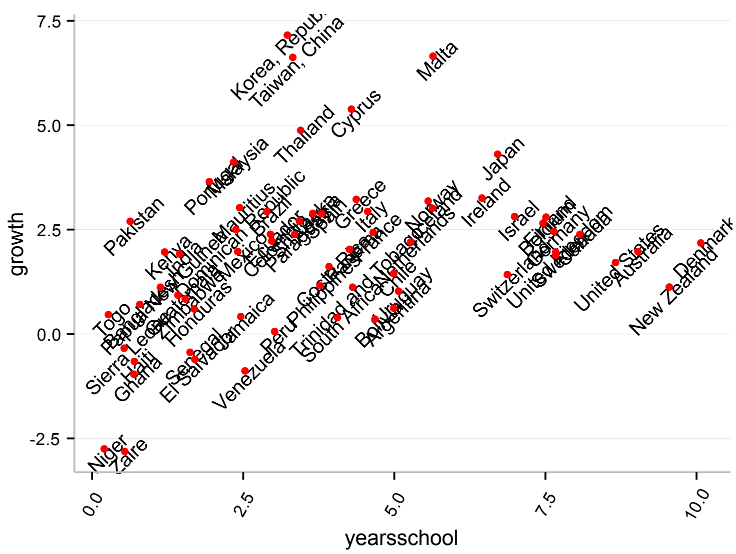
- For now agnostic about causality, but $\mathrm{E}[Y|X]$ usually is not causal

---

Example: Growth and GDP



Years of schooling in 1960 and growth



2 Conditional expectation function

Conditional expectation function

- One way to describe relation between two variables is a function,

$$Y = h(X)$$

- Most relationships in data are not deterministic, so look at average relationship,

$$\begin{aligned} Y =& \underbrace{\mathrm{E}[Y|X]}_{\equiv h(X)} + \underbrace{(Y - \mathrm{E}[Y|X])}_{\equiv \epsilon} \\ =& \mathrm{E}[Y|X] + \epsilon \end{aligned}$$

- Note that $\mathrm{E}[\epsilon] = 0$ (by definition of ϵ and iterated expectations)
- $\mathrm{E}[Y|X]$ can be any function, in particular, it need not be linear
- Unrestricted $\mathrm{E}[Y|X]$ hard to work with
  - Hard to estimate
  - Hard to communicate if X a vector (cannot draw graphs)
- Instead use linear regression
  - Easier to estimate and communicate
  - Tight connection to $\mathrm{E}[Y|X]$

---

3 Population regression

Population regression

- The bivariate *population regression* of Y on X is

$$(\beta_0, \beta_1) = \underset{b_0, b_1}{\arg\min}\, \mathrm{E}[(Y - b_0 - b_1 X)^2]$$

  i.e. β_0 and β_1 are the slope and intercept that minimize the expected square error of $Y - (\beta_0 + \beta_1 X)$
- Calculating β_0 and β_1:
  - First order conditions:

$$\begin{aligned} [b_0] : 0 =& \frac{\partial}{\partial b_0} \mathrm{E}[(Y - b_0 - b_1 X)^2] \\ =& \mathrm{E}\left[\frac{\partial}{\partial b_0}(Y - b_0 - b_1 X)^2\right] \\ =& \mathrm{E}\left[-2(Y - \beta_0 - \beta_1 X)\right] \end{aligned} \tag{1}$$

  and

$$\begin{aligned} [b_1] : 0 =& \frac{\partial}{\partial b_1} \mathrm{E}[(Y - b_0 - b_1 X)^2] \\ =& \mathrm{E}\left[\frac{\partial}{\partial b_1}(Y - b_0 - b_1 X)^2\right] \\ =& \mathrm{E}\left[-2(Y - \beta_0 - \beta_1 X)X\right] \end{aligned} \tag{2}$$

  - (1) rearranged gives $\beta_0 = \mathrm{E}[Y] - \beta_1 \mathrm{E}[X]$
  - Substituting into (2)

$$\begin{aligned}
0 =& \mathrm{E}\left[X(-Y + \mathrm{E}[Y] - \beta_1 \mathrm{E}[X] + \beta_1 X)\right] \\
=& \mathrm{E}\left[X(-Y + \mathrm{E}[Y])\right] + \beta_1 \mathrm{E}\left[X(X - \mathrm{E}[X])\right] \\
=& -\mathrm{Cov}(X,Y) + \beta_1 \mathrm{Var}(X) \\
\beta_1 =& \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}
\end{aligned}$$

- $\beta_1 = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}$, $\beta_0 = \mathrm{E}[Y] - \beta_1 \mathrm{E}[X]$

Population regression approximates $\mathrm{E}[Y|X]$

Lemma 1. *The population regression is the minimal mean square error linear approximation to the conditional expectation function, i.e.*

$$\underbrace{\operatorname*{arg\,min}_{b_0,b_1} \mathrm{E}\left[(Y - (b_0 + b_1 X))^2\right]}_{\text{population regression}} = \operatorname*{arg\,min}_{b_0,b_1} \underbrace{\mathrm{E}_X\left[(\mathrm{E}[Y|X] - (b_0 + b_1 X))^2\right]}_{\text{MSE of linear approximation to } \mathrm{E}[Y|X]}$$

Corollary 2. *If* $\mathrm{E}[Y|X] = c + mX$, *then the population regression of* Y *on* X *equals* $\mathrm{E}[Y|X]$, *i.e.* $\beta_0 = c$ *and* $\beta_1 = m$

Proof

Proof.

- Let b_0^*, b_1^* be minimizers of MSE of approximation to $\mathrm{E}[Y|X]$
- Same steps as in population regression formula gives

$$0 = \mathrm{E}\left[-2(\mathrm{E}[Y|X] - b_0^* - b_1^* X)\right]$$

  and

$$0 = \mathrm{E}\left[-2(\mathrm{E}[Y|X] - b_0^* - b_1^* X)X\right]$$

- Rearranging and combining,

$$b_0^* = \mathrm{E}[\mathrm{E}[Y|X]] - b_1^* \mathrm{E}[X] = \mathrm{E}[Y] - b_1^* \mathrm{E}[X]$$

  and

$$\begin{aligned}
0 =& \mathrm{E}\left[X(-\mathrm{E}[Y|X] + \mathrm{E}[Y] + b_1^* \mathrm{E}[X] - b_1^* X)\right] \\
=& \mathrm{E}\left[X(-\mathrm{E}[Y|X] + \mathrm{E}[Y])\right] + b_1^* \mathrm{E}\left[X(X - \mathrm{E}[X])\right] \\
=& -\mathrm{Cov}(X,Y) + b_1^* \mathrm{Var}(X) \\
b_1^* =& \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}
\end{aligned}$$

□

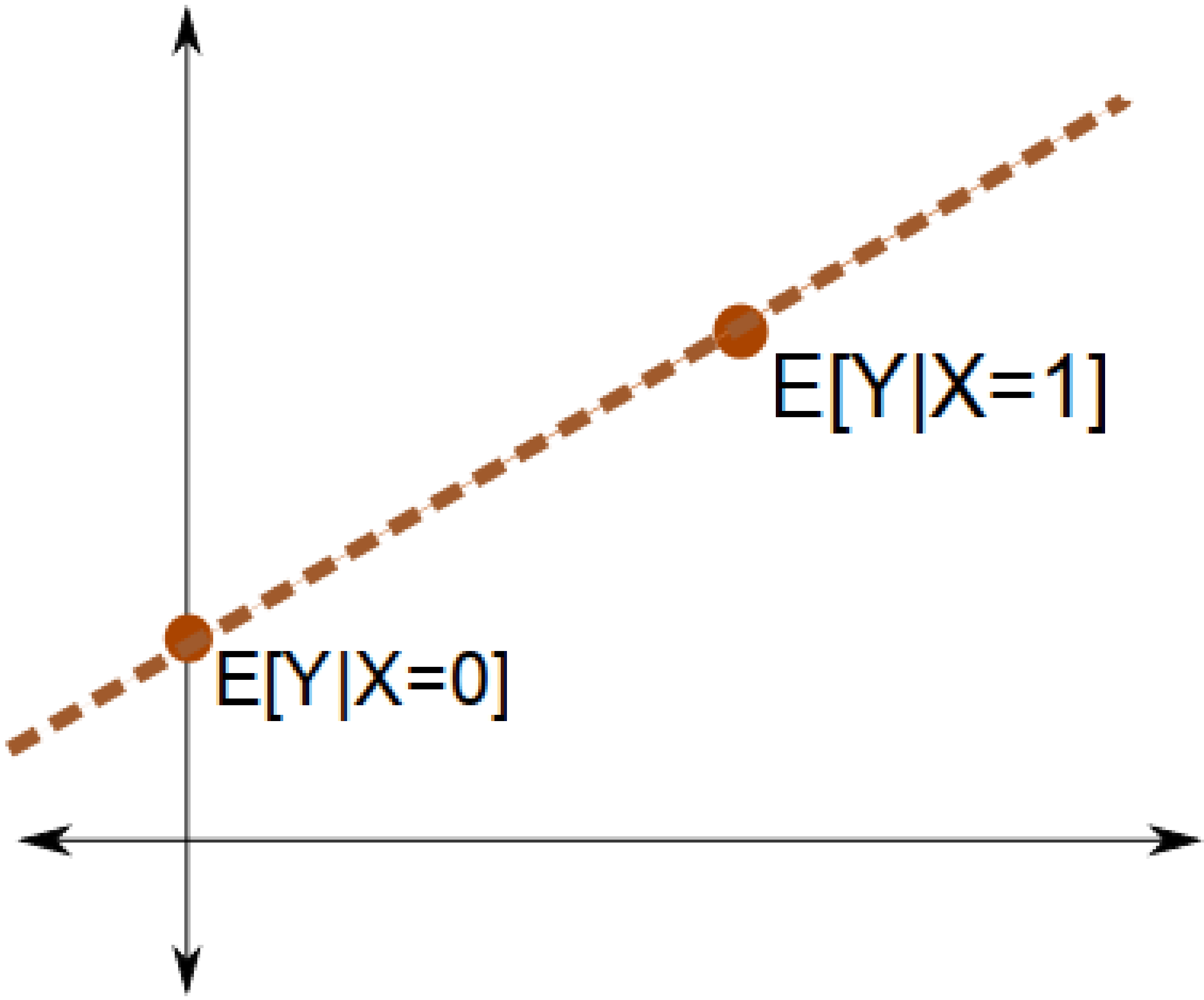
3.1 Interpretation

Regression interpretation

- Regression = best linear approximation to $\mathrm{E}[Y|X]$
- $\beta_0 \approx \mathrm{E}[Y|X=0]$
- $\beta_1 \approx \frac{d}{dx}\mathrm{E}[Y|X] \approx$ change in average Y per unit change in X
- Not necessarily a causal relationship (usually not)
- Always can be viewed as description of data

Regression with binary X

- Suppose X is binary (i.e. can only be 0 or 1)
- We know $\beta_0 + \beta_1 X =$ best linear approximation to $\mathrm{E}[Y|X]$
- X only takes two values, so can draw line connecting $\mathrm{E}[Y|X=0]$ and $\mathrm{E}[Y|X=1]$, so $\beta_0 + \beta_1 X = \mathrm{E}[Y|X]$
  - $\beta_0 = \mathrm{E}[Y|X=0]$
  - $\beta_0 + \beta_1 = \mathrm{E}[Y|X=1]$



4 Sample regression

Sample regression

- Have sample of observations: $\{(y_i, x_i)\}_{i=1}^n$
- The *sample regression* (or when unambiguous just "regression") of Y on X is

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{b_0, b_1} \frac{1}{n} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

  i.e. $\hat{\beta}_0$ and $\hat{\beta}_1$ are the slope and intercept that minimize the sum of squared errors, $(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$

  - Same as population regression but with sample average instead of expectation

- Same calculation as for population regression would show

$$\hat{\beta}_1 = \frac{\widehat{\text{Cov}}(X,Y)}{\widehat{\text{Var}}(X)} = \frac{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2}$$

and

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

Since $\hat{\beta}_1$ and $\hat{\beta}_0$ come from minimizing a sum of squares, they are called the ordinary least squares estimates, or OLS for short.

The formulas for $\hat{\beta}_1$ and $\hat{\beta}_0$ come from the first order conditions. These estimators minimize the sum of squared differences between the regression line and the observed y_i,

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{b_0, b_1}{\arg\min} \frac{1}{n} \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2.$$

The first order condition for $\hat{\beta}_0$ is:

$$0 = \frac{1}{n}\sum_{i=1}^n 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$$

which can be rearranged to get

$$\begin{aligned} \hat{\beta}_0 &= \frac{1}{n}\left(\sum_{i=1}^n y_i - \hat{\beta}_1 x_i\right) \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}. \end{aligned}$$

The first order condition for $\hat{\beta}_1$ is:

$$0 = \frac{1}{n}\sum_{i=1}^n 2x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i).$$

Substituting in the previous expression for $\hat{\beta}_0$ gives

$$0 = \frac{1}{n}\sum_{i=1}^n 2x_i(y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i).$$

Rearranging to solve for $\hat{\beta}_1$:

$$\begin{aligned} \hat{\beta}_1 \sum_{i=1}^n x_i(x_i - \bar{x}) &= \sum_{i=1}^n x_i(y_i - \bar{y}) \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^n x_i(y_i - \bar{y})}{x_i(x_i - \bar{x})} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(x_i - \bar{x})^2}. \end{aligned}$$

Sample regression

- Sample regression is an estimator for the population regression
- Given an estimator we should ask:
  - Unbiased?
  - Variance?
  - Consistent?
  - Asymptotically normal?
- We will address these questions in the next week or two

---

5 Regression in R

Regression in R

```r
require(datasets) ## some datasets included with R
stateDF <- data.frame(state.x77)
summary(stateDF) ## summary statistics of data

## Sample regression function
regress <- function(y, x) {
  beta <- vector(length=2)
  beta[2] <- cov(x,y)/var(x)
  beta[1] <- mean(y) - beta[2]*mean(x)
  return(beta)
}

## Regress life expectancy on income
beta <- regress(stateDF[,"Life.Exp"],stateDF$Income)
beta

## builtin regression
lm(Life.Exp ~ Income, data=stateDF)
## more detailed output
summary(lm(Life.Exp ~ Income, data=stateDF))
```

https://bitbucket.org/paulschrimpf/econ326/src/master/notes/03/regress.R?at=master

---

Part II
Properties of regression

6 Fitted value and residuals

These algebraic identities about fitted values and residuals are things that we will use repeatedly later. I would not recommend spending time trying to memorize these. The important ones will come up repeatedly and you will remember them without any special effort. The first time we use these identities, we will go through how to get them again. We may even go through them yet again the second and third time we use them. Eventually we will use some of these identities so often that you will either be able to quickly derive them or just remember them.

Fitted values and residuals

- *Fitted values*:
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$
- *Residuals*:
$$\hat{\epsilon}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i = y_i - \hat{y}_i$$
$$y_i = \hat{y}_i + \hat{\epsilon}_i$$
- Sample mean of residuals $= 0$
  - First order condition for $\hat{\beta}_0$,
$$0 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$$
$$0 = \frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i$$
- Sample covariance of x and $\hat{\epsilon} = 0$
  - First order condition for $\hat{\beta}_1$,
$$0 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i$$
$$0 = \frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i x_i$$

- Sample mean of $\hat{y}_i = \bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} y_i =& \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i + \hat{\epsilon}_i \\
=& \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i \\
=& \frac{1}{n}\sum_{i=1}^{n} \hat{\beta}_0 + \hat{\beta}_1 x_i \\
=& \hat{\beta}_0 + \hat{\beta}_1 \bar{x}
\end{aligned}$$

- Sample covariance of y and $\hat{\epsilon}$ = sample variance of $\hat{\epsilon}$:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} y_i(\hat{\epsilon}_i - \bar{\hat{\epsilon}}) =& \frac{1}{n}\sum_{i=1}^{n} y_i \hat{\epsilon}_i \\
=& \frac{1}{n}\sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\epsilon}_i)\hat{\epsilon}_i \\
=& \hat{\beta}_0 \frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i + \beta_1 \frac{1}{n}\sum_{i=1}^{n} x_i \hat{\epsilon}_i + \frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i^2 \\
=& \frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i^2
\end{aligned}$$

R^2

- Decompose y_i

$$y_i = \hat{y}_i + \hat{\epsilon}_i$$

- Total sum of squares $=$ explained sum of squares $+$ sum of squared residuals

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n} (y_i - \bar{y})^2}_{SST} = \underbrace{\frac{1}{n}\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}_{SSE} + \underbrace{\frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i^2}_{SSR}$$

- *R-squared*: fraction of sample variation in y that is explained by x

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST} = o\widehat{\text{Corr}}(y, \hat{y})$$

  - $0 \leq R^2 \leq 1$
  - If all data on regression line, then $R^2 = 1$
  - Magnitude of R^2 does not have direct bearing on economic importance of a regression

---

7 Statistical properties

7.1 Unbiased

Unbiased

- $\mathrm{E}[\hat{\beta}] = ?$
- Assume:

SLR.1 (linear model) $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

SLR.2 (independence) $\{(x_i, y_i)\}_{i=1}^n$ is independent random sample

SLR.3 (rank condition) $\widehat{\mathrm{Var}}(X) > 0$

SLR.4 (exogeneity) $\mathrm{E}[\epsilon|X] = 0$

- Then, $\mathrm{E}[\hat{\beta}_1] = \beta_1$ and $\mathrm{E}[\hat{\beta}_0] = \beta_0$

It is more important to understand the meaning of these four assumptions (discussed below) than the proof that regression is unbiased.

Regression is unbiased. We need to calculate $\mathrm{E}[\hat{\beta}]$. First, substitute in the formula for $\hat{\beta}$.

$$\mathrm{E}[\hat{\beta}_1] = \mathrm{E}\left[\frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x}) x}\right]$$

Next, substitute in the model for y_i, $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$,

$$
\begin{aligned}
\mathrm{E}[\hat{\beta}_1] =& \mathrm{E}\left[\frac{\sum_{i=1}^n (x_i - \bar{x})(\beta_0 + \beta_1 x_i + \epsilon_i)}{\sum_{i=1}^n (x_i - \bar{x})x}\right] \\
& \text{rearrange} \\
=& \mathrm{E}\left[\frac{\overbrace{\sum_{i=1}^n x_i - \bar{x}}^{=0}}{\sum_{i=1}^n (x_i - \bar{x})x}\beta_0 + \overbrace{\left(\frac{\sum_{i=1}^n (x_i - \bar{x})x_i}{\sum_{i=1}^n (x_i - \bar{x})x}\right)}^{=1}\beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x})\epsilon_i}{\sum_{i=1}^n (x_i - \bar{x})x}\right] \\
& \text{use linearity of expectation} \\
=& \beta_1 + \mathrm{E}\left[\frac{\sum_{i=1}^n (x_i - \bar{x})\epsilon_i}{\sum_{i=1}^n (x_i - \bar{x})x}\right] \\
& \text{use iterated expectations} \\
=& \beta_1 + \mathrm{E}_X\left[\mathrm{E}_{\epsilon|X}\left[\left.\frac{\sum_{i=1}^n (x_i - \bar{x})\epsilon_i}{\sum_{i=1}^n (x_i - \bar{x})x}\right| x_1, x_2, ..., x_n\right]\right] \\
& \text{conditional on } x_1, ..., x_n, x_i \text{ is constant} \\
=& \beta_1 + \mathrm{E}_X\left[\frac{\sum_{i=1}^n (x_i - \bar{x})\mathrm{E}[\epsilon_i | x_1, ..., x_n]}{\sum_{i=1}^n (x_i - \bar{x})x}\right] \\
& \text{independent observations implies } \mathrm{E}[\epsilon_i | x_1, ..., x_n] = \mathrm{E}[\epsilon_i | x_i] \\
=& \beta_1 + \mathrm{E}_X\left[\frac{\sum_{i=1}^n (x_i - \bar{x})\mathrm{E}[\epsilon_i | x_i]}{\sum_{i=1}^n (x_i - \bar{x})x}\right] \\
& \text{exogeneity assumption says that } \mathrm{E}[\epsilon_i | x_i] = 0. \\
=& \beta_1
\end{aligned}
$$

□

Note that the first few steps of the above proof showed that

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

This is a very useful expression that can be used as a starting point for calculating the variance of $\hat{\beta}_1$, and thinking about what happens if exogeneity fails and $\mathrm{E}[\epsilon | x] \neq 0$.

7.2 Variance

Variance

- $\mathrm{Var}(\hat{\beta})$?
- Assume SLR.1-4 and

SLR.5 (homoskedasticity) $\mathrm{Var}(\epsilon|X) = \sigma^2$

- Then,

$$\mathrm{Var}(\hat{\beta}_1|\{x_i\}_{i=1}^n) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

and

$$\mathrm{Var}(\hat{\beta}_0|\{x_i\}_{i=1}^n) = \frac{\sigma^2 \frac{1}{n} \sum_{i=1}^n x_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

As in the proof that regression is unbiased,

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

We now want to take the variance of this expression. Before doing so, it will be useful to review some properties of the variance of a sum of random variables.

Lemma 3. *Let a, b, and c be constants, and Z and W be random variables. Then,*

$$\mathrm{Var}(a + bZ + cW) = b^2\mathrm{Var}(Z) + c^2\mathrm{Var}(W) + 2bc\mathrm{Cov}(Z, W).$$

Proof. We can prove this using the definition of variance.

$$\begin{aligned}
\mathrm{Var}(a + bZ + cW) =& \mathrm{E}\left[(a + bZ + cW - \mathrm{E}[a + bZ + cW])^2\right] \\
=& \mathrm{E}\left[(a + bZ + cW - a - b\mathrm{E}[Z] - c\mathrm{E}[W])^2\right] \\
=& \mathrm{E}\left[(b(Z - \mathrm{E}[Z]) + c(W - \mathrm{E}[W]))^2\right] \\
=& \mathrm{E}\left[b^2(Z - \mathrm{E}[Z])^2 + 2bc(Z - \mathrm{E}[Z])(W - \mathrm{E}[W]) + c^2(W - \mathrm{E}[W])^2\right] \\
=& b^2\mathrm{E}\left[(Z - \mathrm{E}[Z])^2\right] + 2bc\mathrm{E}\left[(Z - \mathrm{E}[Z])(W - \mathrm{E}[W])\right] + c^2\mathrm{E}\left[(W - \mathrm{E}[W])^2\right] \\
=& b^2\mathrm{Var}(Z) + c^2\mathrm{Var}(W) + 2bc\mathrm{Cov}(Z, W)
\end{aligned}$$

□

Generalizing the above to the sum of more than two random variables, we have

Corollary 4. *Let $a_1, \ldots, a_n$ be constants, and $Z_1, \ldots, Z_n$ be random variables, then,*

$$\mathrm{Var}\left(\sum_{i=1}^n a_i Z_i\right) = \sum_{i=1}^n \sum_{j=1}^n a_i a_j \mathrm{Cov}(Z_i, Z_j)$$

Futhermore, if Z_i and Z_j are independent (or just uncorrelated) for $i \neq j$, then

$$\mathrm{Var}\left(\sum_{i=1}^n a_i Z_i\right) = \sum_{i=1}^n a_i^2 \mathrm{Var}(Z_i)$$

We can apply this corollary to

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}_1|x) =& \mathrm{Var}\left(\beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x})\epsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \middle| x\right) \\
=& \mathrm{Var}\left(\sum_{i=1}^n \underbrace{\frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}}_{a_i} \underbrace{\epsilon_i}_{Z_i} \middle| x\right) && \text{using the corollary} \\
=& \sum_{i=1}^n \sum_{j=1}^n \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \frac{x_j - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \mathrm{Cov}(\epsilon_i, \epsilon_j | x) && \text{independence} \\
=& \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{\sum_{i=1}^n x_i - \bar{x}}\right)^2 \mathrm{Var}(\epsilon_i|x) && \text{homoskedasticity} \\
=& \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{\left(\sum_{i=1}^n (x_i - \bar{x})^2\right)^2} \sigma_\epsilon^2 \\
=& \frac{\sigma_\epsilon^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.
\end{aligned}$$

7.3 Distribution

Distribution with normal errors

- Assume SLR.1-SLR.5 and

  SLR.6 (normality) $\epsilon_i | x_i \sim N(0, \sigma^2)$

- Then $Y|X \sim N(\beta_0 + \beta_1 X, \sigma^2)$, and

$$\hat{\beta}_1 | \{x_i\}_{i=1}^n \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)$$

- Even without assuming normality, the central limit theorem implies $\hat{\beta}$ is asymptotically normal (details in a later lecture)

An important property of normal random variables is that if Z and W are independent, and $Z \sim N(\mu_z, \sigma_z^2)$ and $W \sim N(\mu_w, \sigma_w^2)$, then

$$a + bZ + cW \sim N(a + b\mu_z + c\mu_w, b^2\sigma_z^2 + c^2\sigma_w^2).$$

If we assume that ϵ_i is normally distributed conditional on x, then since $\hat{\beta}_1$ is just a sum of the ϵ_i,[1] $\hat{\beta}_1$ will also be normally distributed.

---

[1]Specifically, $\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x})\epsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$.

Summary

- Simple linear regression model assumptions:

  SLR.1 (linear model) $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
  SLR.2 (independence) $\{(x_i, y_i)\}_{i=1}^n$ is independent random sample
  SLR.3 (rank condition) $\widehat{\text{Var}}(X) > 0$
  SLR.4 (exogeneity) $\text{E}[\epsilon|X] = 0$
  SLR.5 (homoskedasticity) $\text{Var}(\epsilon|X) = \sigma^2$
  SLR.6 (normality) $\epsilon_i|x_i \sim N(0, \sigma^2)$

- $\hat{\beta}$ unbiased if SLR.1-SLR.4
- If also SLR.5, then $\text{Var}(\hat{\beta}_1|\{x_i\}_{i=1}^n) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$
- If also SLR.6, then $\hat{\beta}_1|\{x_i\}_{i=1}^n \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)$

7.4 Discussion of assumptions

Discussion of assumptions

SLR.1 Having a linear model makes it easier to state the other assumptions, but we could instead start by saying let $\beta_1 = \frac{\text{Cov}(X,Y)}{\text{Var}(X)}$ and $\beta_0 = \text{E}[Y] - \beta_1 \text{E}[X]$ be the population regression coefficients and define $\epsilon_i = y_i - \beta_0 - \beta_1 x_i$

To say whether an estimator is unbiased, we first have to define what parameter we want to estimate. Assuming that there is a linear model defines the parameter we want to estimate. This linear model could be population regression, in which case, $\beta_1 = \frac{\text{Cov}(X,Y)}{\text{Var}(X)}$, and by construction we must have $\text{E}[X\epsilon] = 0$.

However, the linear model may also be motivated by economic theory. For example, consider a Cobb-Douglass production function with only one input, labor,

$$Y = AL^\alpha,$$

where Y is output, L is labor, and A is productivity. If we take logs, then

$$\log Y = \log A + \alpha \log L.$$

If we rearrange slightly, we get something that looks just like a linear regression model,

$$\underbrace{\log Y_i}_{y_i} = \underbrace{\text{E}[\log A]}_{\beta_0} + \underbrace{\alpha}_{\beta_1} \underbrace{\log L_i}_{x_i} + \underbrace{(\log A_i - \text{E}[\log A])}_{\epsilon_i}.$$

If this is the model we want to estimate, then ϵ_i is not the error term in the population regression. Instead ϵ_i is the difference between the log productivity of firm i and average log productivity. It is unlikely that this ϵ_i would be uncorrelated with $\log L_i$. More productive firms generally choose to use more inputs, so we should suspect that ϵ_i and $\log L_i$ are positively correlated.

Discussion of assumptions

SLR.2 Independent observations is a good assumption for data from a simple random sample

- Common situations where it fails in economics are when we have a time series of observations, e.g. $\{(x_t, y_t)\}_{t=1}^n$ could be unemployment and GDP of Canada for many different years; and clustering, e.g. the data could be students test scores and hours studying and our sample consists of randomly chosen courses or schools—students in the same course would not be independent, but across different courses they might be.
- Still have $E[\hat{\beta}_1] = \beta_1$ with non-independent observations as long as $E[\epsilon_i | x_1, ..., x_n] = 0$
- The variance of $\hat{\beta}_1$ will change with non-independent observations
- Simulation code

Independence says that knowing the values of x_1 and y_1 tells you nothing about the distribution of x_2 and y_2 (or any other observation). When we have cross-sectional data, this assumption usually makes sense. In economics, we sometimes deal with time-series data, where (x_1, y_1) would be the observation of something at time 1 and (x_2, y_2) is the observation that same thing at time 2. In this case, independence is unlikely to hold. Another common situation is panel data, where we observe a sample of individuals over time, so (x_{it}, y_{it}) would be what we observe from individual i at time t. Again, it is unlikely that these observations would be independent over time. Later in the course, we will talk about how to deal with non-independent observations.

Discussion of assumptions

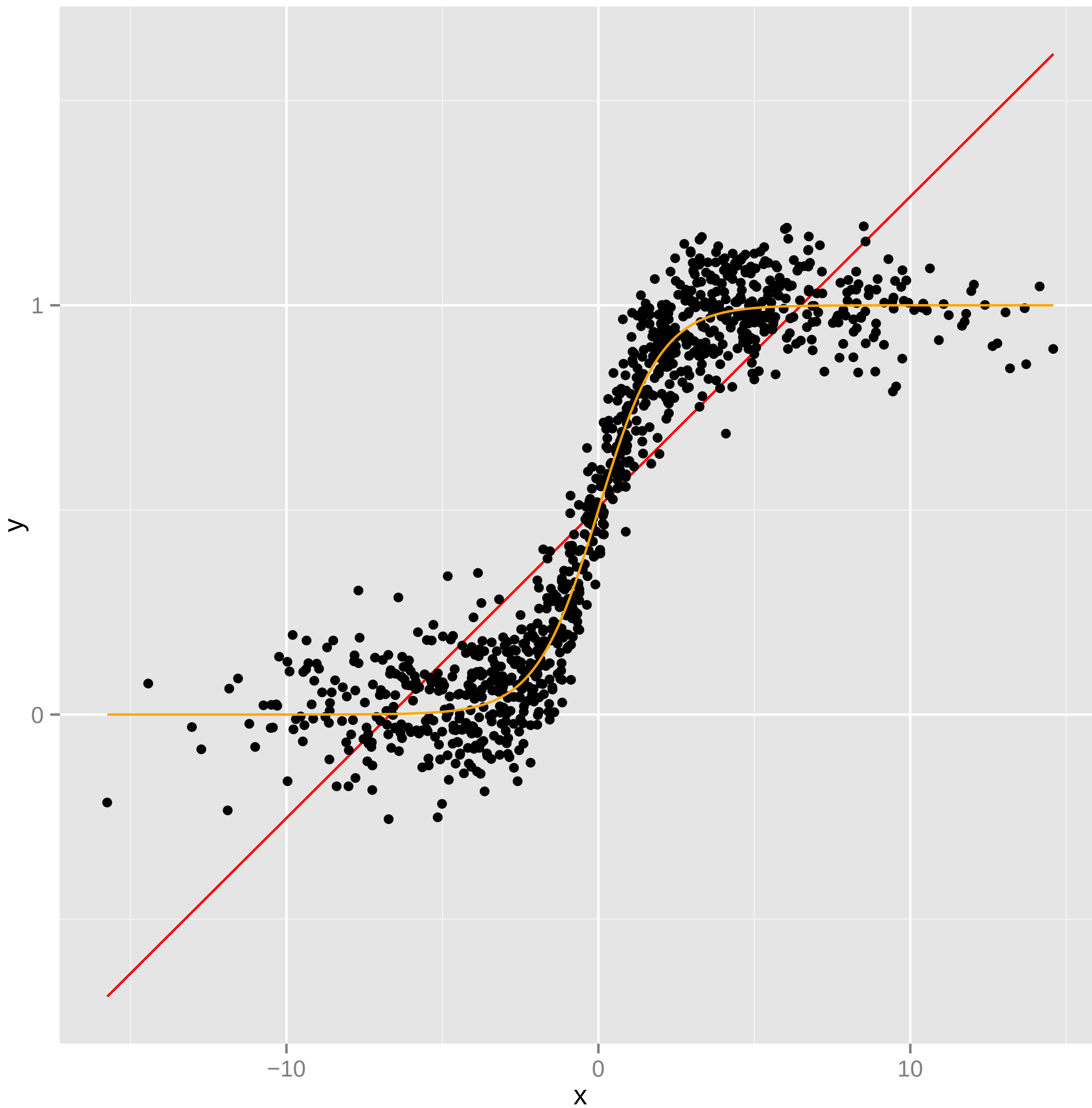
SLR.3 If $\widehat{\text{Var}}(X) = 0$, then $\hat{\beta}_1$ involves dividing by 0

- If there is no variation in X, then we cannot see how Y is related to X

Discussion of assumptions

SLR.4 To think about mean independence of ϵ from x we should have a model motivating the regression

- If the model we want is just a population regression, then automatically $E[\epsilon X] = 0$, and $E[\epsilon | X] = 0$ if the conditional expectation function is linear; if conditional expectation nonlinear maybe still a useful approximation



Code

Discussion of assumptions

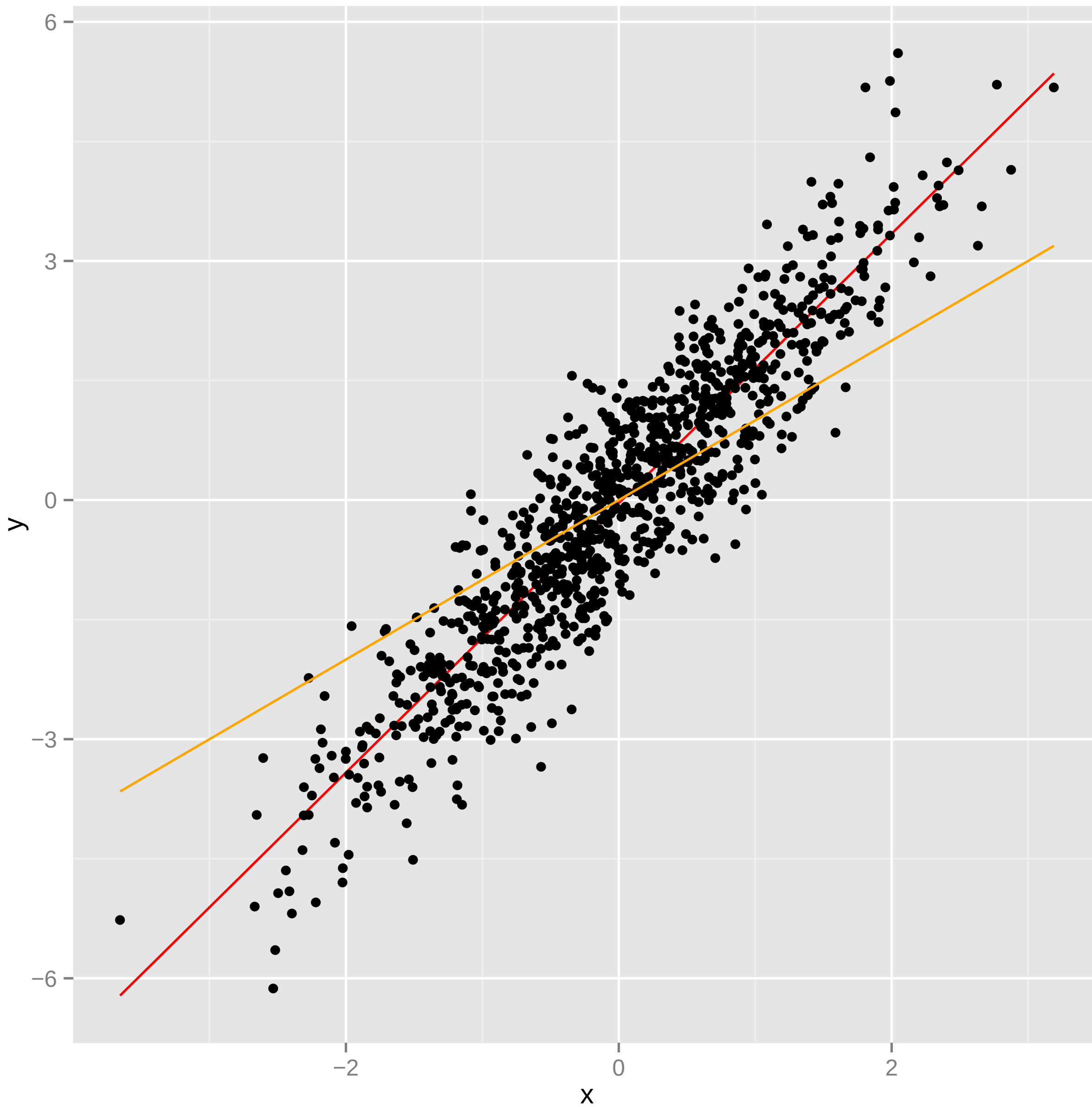
SLR.4 To think about mean independence of ϵ from x we should have a model motivating the regression

- If the model we want is anything else, then maybe $\mathrm{E}[\epsilon X] \neq 0$ (and $\mathrm{E}[\epsilon|X] \neq 0$), e.g.
  - Demand curve

$$p_i = \beta_0 + \beta_1 q_i + \epsilon_i$$

ϵ_i = everything that affects price other than quantity. q_i determined in equilibrium implies $\mathrm{E}[\epsilon_i|q_i] \neq 0$

– $\mathrm{E}[\hat{\beta}_1] \neq \beta_1$ and $\hat{\beta}_1$ does not tell us what we want



Code

Exogeneity is the most important assumption underlying regression. In fact, estimating any economic model using any method will involve some kind of exogeneity assumption. By this, we mean that every estimation method requires assuming some error term is either completely independent of some observable ($F_{\epsilon|x}(\epsilon|x) = F_\epsilon(\epsilon)$), mean independent of some observable ($\mathrm{E}[\epsilon|X] = 0$), or at least uncorrelated with an observable ($\mathrm{E}[\epsilon X] = 0$). Much of what separates good empirical work in economics from bad is how plausible are the exogeneity assumptions. Often, economic theory can help us decide whether or not an exogeneity assumption is plausible. Consider the production function example from earlier,

$$\underbrace{\log Y_i}_{y_i} = \underbrace{\mathrm{E}[\log A]}_{\beta_0} + \underbrace{\alpha}_{\beta_1} \underbrace{\log L_i}_{x_i} + \underbrace{(\log A_i - \mathrm{E}[\log A])}_{\epsilon_i}.$$

To think about whether mean independence of the error term, $\mathrm{E}[\underbrace{(\log A - \mathrm{E}[\log A])}\,|\log L] = 0$, makes sense in this model, we should think about how L is determined. The firm chooses how much labor to use. Suppose the firm faces output price p and wage w. If the firm chooses L knowing its productivity, then the firm solves,

$$\max_L pAL^{\alpha} - wL$$

The first order condition is

$$pA\alpha L^{\alpha-1} - w = 0.$$

If we solve for A, we get

$$A = \frac{w}{p\alpha} L^{1-\alpha}$$

$$\log A = \log\left(\frac{w}{p\alpha}\right) + (1-\alpha)\log L$$

so

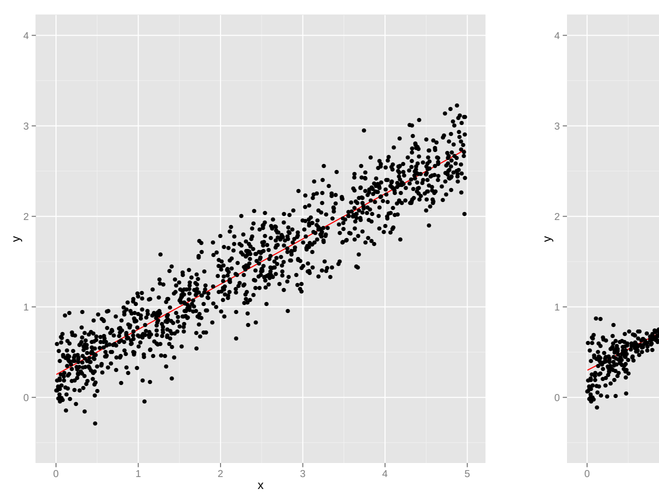
$$\mathrm{E}[\log A|\log L] = (1-\alpha)\log L + \mathrm{E}\left[\log\left(\frac{w}{p\alpha}\right)\middle|\log L\right]$$

This will not be a constant unless $\alpha = 1$ and $\frac{w}{p}$ is mean independent of $\log L$. Both of these are unlikely to hold. Unless $\mathrm{E}[\log A|\log L]$ is constant, $\mathrm{E}[\underbrace{(\log A - \mathrm{E}[\log A])}\,|\log L] \neq 0$. Therefore, exogeneity is not a good assumption in this model, and regression will not give an unbiased estimate of the production function.

Discussion of assumptions

SLR.5 Homoskedasticity: variance of ϵ does not depend on X

Homoskedastic



Code

- *Heteroskedasticity* is when $\mathrm{Var}(\epsilon|X)$ varies with X
- If there is heteroskedasticity, the variance of $\hat{\beta}_1$ is different, but we can fix it

– "robust standard errors" / "heteroscedasticity-consistent (HC) standard errors" / "Eicker–Huber–White standard errors"

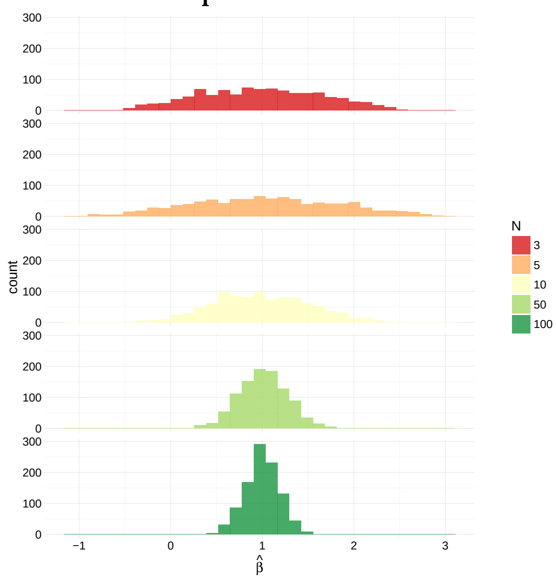
Homoskedasticity is a strong assumption that is usually not very plausible. Therefore, in practice, economists almost always calculate heteroscedasticity-robust standard errors.

Discussion of assumptions

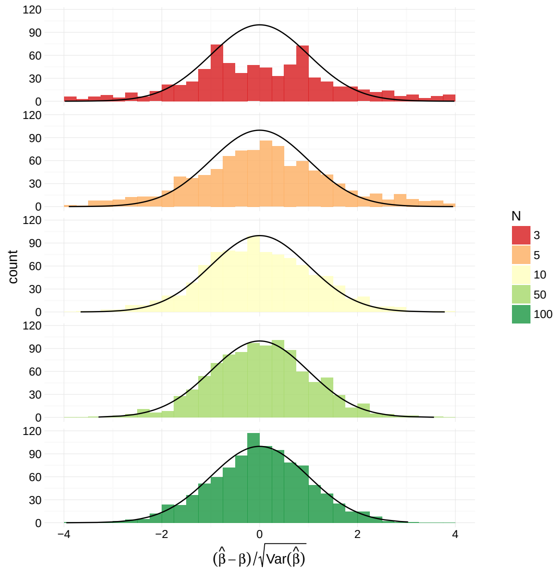
SLR.6 If $\epsilon_i | x_i \sim N$, then $\hat{\beta}_1 \sim N$

- What if ϵ_i not normally distributed?
- We will see that $\hat{\beta}_1$ still asymptotically normal
- Simulation

Discussion of assumptions



Discussion of assumptions

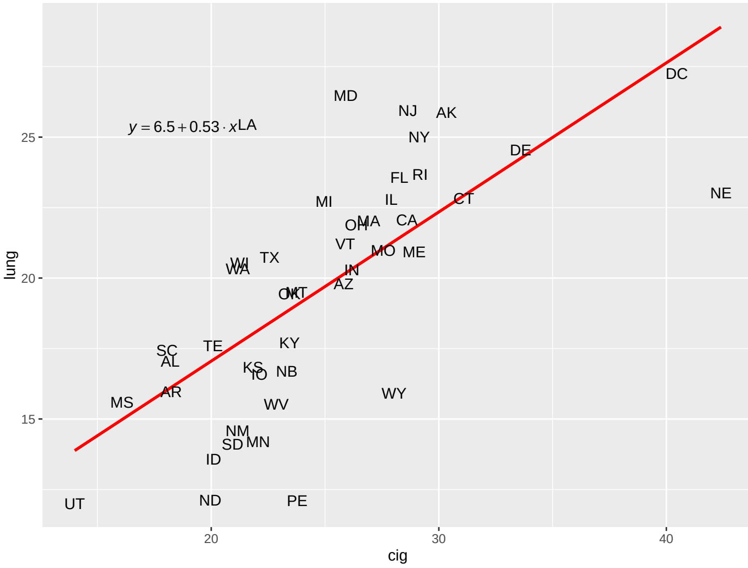


8 Examples

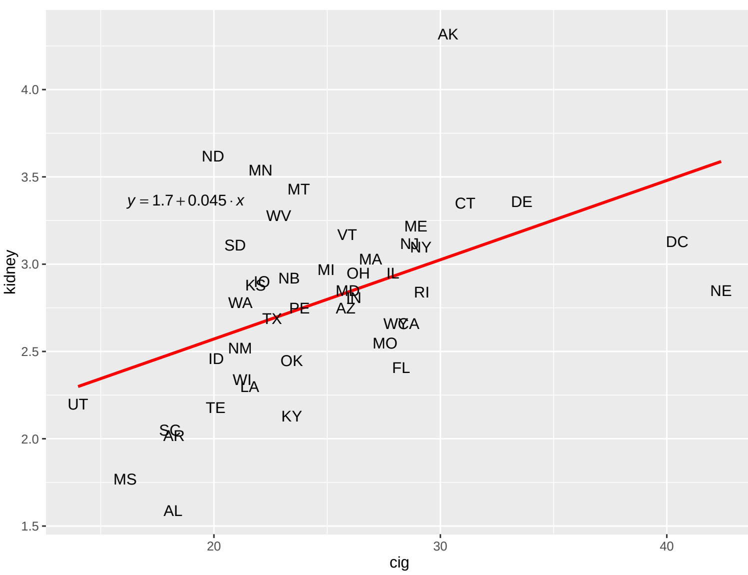
Example: smoking and cancer

- Data on per capita number of cigarettes sold and death rates per thousand from cancer for U.S. states in 1960
- `http://lib.stat.cmu.edu/DASL/Datafiles/cigcancerdat.html`
- Death rates from: lung cancer, kidney cancer, bladder cancer, and leukemia Code

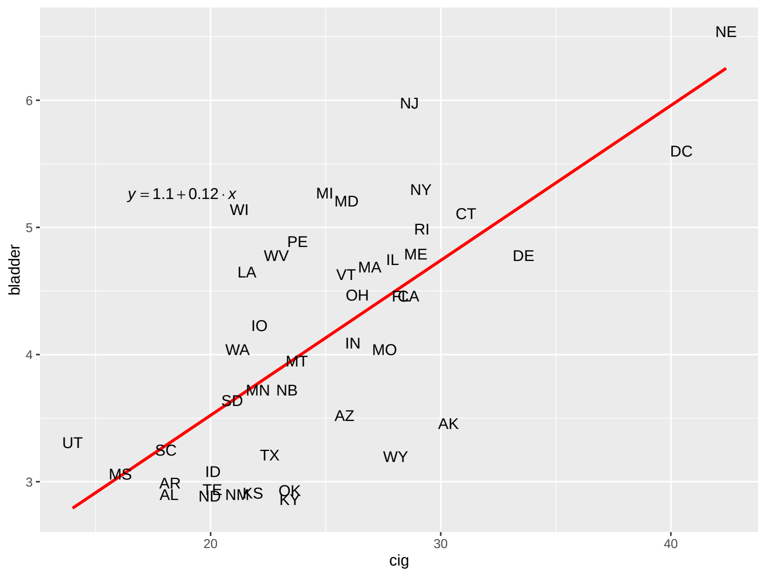
Smoking and lung cancer



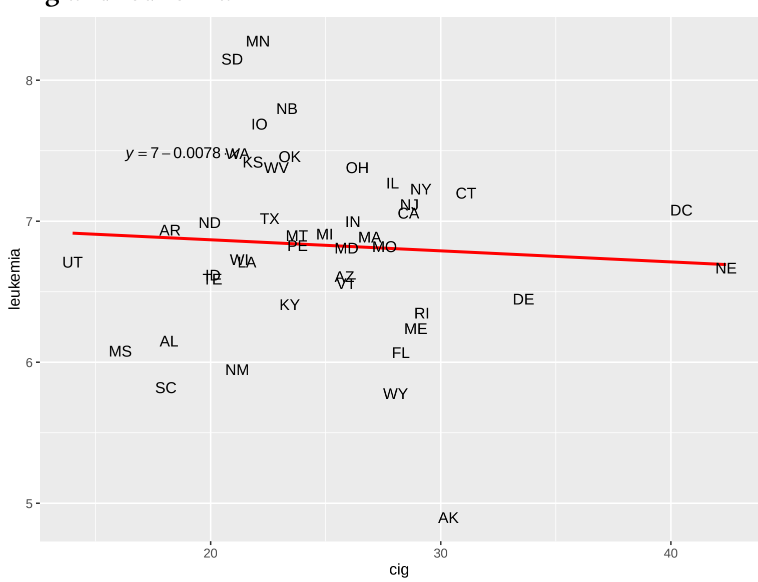
Smoking and kidney cancer



Smoking and bladder cancer



Smoking and leukemia

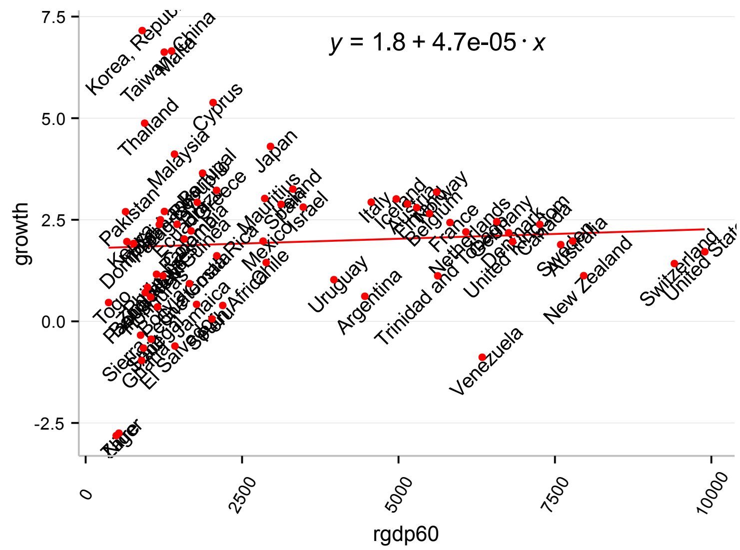


Example: convergence in growth

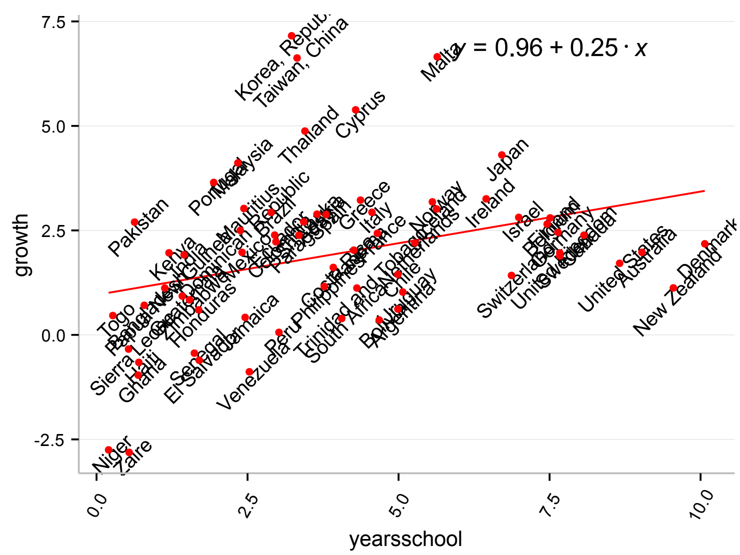
- Data on average growth rate from 1960-1995 for 65 countries along with GDP in 1960, average years of schooling in 1960, and other variables
- From `http://wps.aw.com/aw_stock_ie_2/50/13016/3332253.cw/index.html`, originally used in Beck, Levine, and Loayza (2000)

- Question: has there been in convergence, i.e. did poorer countries in 1960 grow faster and catch-up?
- Code

GDP in 1960 and growth



Years of schooling in 1960 and growth



- Things look different 1995-2014
- Code to download and recreate results using updated growth data through 2014 from the World Bank

9 Inference

Inference with normal errors

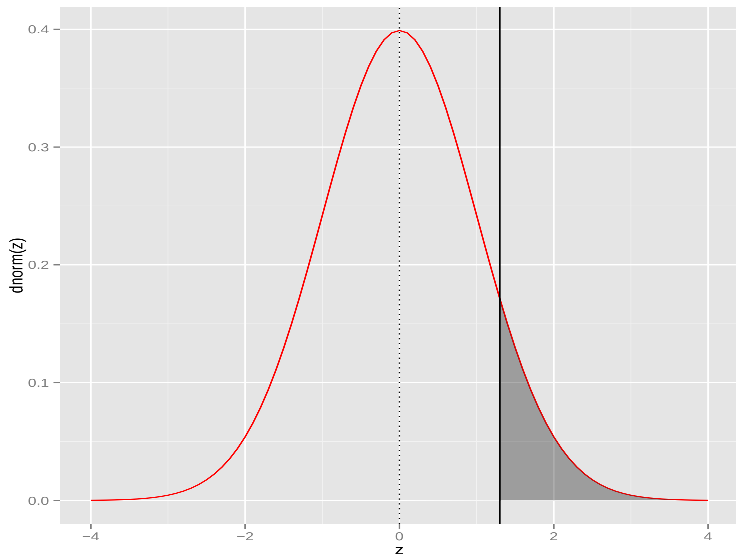
- Regression estimates depend on samples, which are random, so the regression estimates are random
  - Some regressions will randomly look "interesting" due to chance
- Logic of hypothesis testing: figure out probability of getting an interesting regression estimate due solely to change
- Null hypothesis, H_0 : the regression is uninteresting, usually $\beta_1 = 0$
- With assumptions SR.1-SR.6 and under $H_0 : \beta_1 = \beta_1^*$, we know

$$\hat{\beta} \sim N\left(\beta_1^*, \frac{\sigma_\epsilon^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)$$

  or equivalently,

$$t \equiv \frac{\hat{\beta} - \beta_1^*}{\sigma_\epsilon / \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} \sim N(0,1)$$

- P-value: the probability of getting a regression estimate as or more "interesting" than the one we have
  - As or more interesting = as far or further away from β_1^*
  - If we are only interested when $\hat{\beta}_1$ is on one side of β_1^*, then we have a one sided alternative, e.g. $H_a : \beta_1 > \beta_1^*$
  - If we are equally interested in either direction, then $H_a : \beta_1 \neq \beta_1^*$



Inference with normal errors

- One-sided p-value: $p = \Phi(-|t|) = 1 - \Phi(|t|)$
- Two-sided p-value: $p = 2\Phi(-|t|) = 2(1 - \Phi(|t|))$
- Interpretation:
  - The probability of getting an estimate as strange as the one we have if the null hypothesis is true.

	Model 1	Model 2	Model 3	Model 4
(Intercept)	1.09^{*}	6.47^{**}	1.66^{***}	7.03^{***}
	(0.48)	(2.14)	(0.32)	(0.45)
cig	0.12^{***}	0.53^{***}	0.05^{***}	-0.01
	(0.02)	(0.08)	(0.01)	(0.02)
R^2	0.50	0.49	0.24	0.00
Adj. R^2	0.48	0.47	0.22	-0.02
Num. obs.	44	44	44	44
RMSE	0.69	3.07	0.46	0.64

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 1: Smoking and cancer

  - It is *not* about the probability of β_1 being any particular value. β_1 is not a random variable. It is some unknown number. The data is what is random. In particular, the p-value is *not* the probability that that H_0 is false given the data.

- Hypothesis testing: we must make a decision (usually reject or fail to reject H_0)
  - Choose significance level α (usually 0.05 or 0.10)
  - Construct procedure such that if H_0 is true, we will incorrectly reject with probability α
  - Reject null if p-value less than α

9.1 Examples (continued)

Smoking and cancer

Growth and GDP

Caution: multiple testing

- We just looked at 6 regressions, if $H_0 : \beta_1 = 0$ is true in all of them the probability that correctly fail to reject all 6 null hypotheses with a 5% test is $0.95^6 = 0.74$ (assuming the 6 tests are independent)
- A quarter of the time if we look at 6 regressions, we will randomly find at least significant relationship; if we look at 14 regressions the probability that we incorrectly reject a null is more than 0.5

	Model 1	Model 2
(Intercept)	1.80***	0.96*
	(0.38)	(0.42)
rgdp60	0.00	
	(0.00)	
yearsschool		0.25**
		(0.09)
R^2	0.00	0.11
Adj. R^2	-0.01	0.10
Num. obs.	65	65
RMSE	1.91	1.80

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 2: Growth and GDP and education in 1960

Caution: economic significance $\neq$ statistical significance

---

9.2 Estimating σ_ϵ^2

Estimating σ_ϵ^2

- Recall that $\text{Var}(\hat{\beta}|x_1, ..., x_n) = \frac{\sigma_\epsilon^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sigma_\epsilon^2}{n\widehat{\text{Var}}(x)}$
- σ_ϵ^2 unknown
- We estimate σ_ϵ^2 using the residuals,

$$\hat{\sigma}_\epsilon^2 = \frac{1}{n-2}\sum_{i=1}^n \underbrace{\hat{\epsilon}_i^2}_{=(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}$$

- If SLR.1-SLR.5, $\text{E}[\hat{\sigma}_\epsilon^2] = \sigma_\epsilon^2$
- Using $\frac{1}{n-2}$ instead of $\frac{1}{n}$ makes $\hat{\sigma}_\epsilon^2$ unbiased
  - $\hat{\epsilon}_i$ depends on 2 estimated parameters, $\hat{\beta}_0$ and $\hat{\beta}_1$, so only $n-2$ degrees of freedom
- Estimate $\text{Var}(\hat{\beta}_1|x_1, ..., x_n)$ by

$$\widehat{\text{Var}}(\hat{\beta}_1|x_1, ..., x_n) = \frac{\hat{\sigma}_\epsilon^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\frac{1}{n-2}\sum_{i=1}^n \hat{\epsilon}_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

- Standard error of $\hat{\beta}_1$ is $\sqrt{\widehat{\text{Var}}(\hat{\beta}_1|x_1, ..., x_n)}$
- If SLR.1-SLR.6, t-statistic with estimated $\widehat{\text{Var}}(\hat{\beta}_1|x_1, ..., x_n)$ has a $t(n-2)$ distribution instead of $N(0, 1)$

$$t = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\widehat{\text{Var}}(\hat{\beta}_1|x_1, ..., x_n)}} \sim t(n-2)$$

9.3 Confidence intervals

Confidence intervals

- $\hat{\beta}_1$ is random
- $\widehat{\text{Var}}(\hat{\beta}_1)$, p-values, and hypthesis tests are ways of expressing how random is $\hat{\beta}_1$
- Confidence intervals are another
- A $1-\alpha$ confidence interval, $CI_{1-\alpha} = [LB_{1-\alpha}, UB_{1-\alpha}]$ is an interval estimator for β_1 such that

$$\text{P}(\beta_1 \in CI_{1-\alpha} = 1 - \alpha)$$

  ($CI_{1-\alpha}$ is random; β_1 is not)
- Recall: if SLR.1-SLR.6, then

$$\hat{\beta}_1 \sim N\left(\beta_1, \text{Var}(\hat{\beta}_1)\right)$$

  – Implies

$$\begin{aligned}
\text{P}\left(\hat{\beta}_1 < \beta_1 + \sqrt{\text{Var}(\hat{\beta}_1)}\Phi^{-1}(\alpha/2)\right) &= \alpha/2 \\
\text{P}\left(\hat{\beta}_1 - \sqrt{\text{Var}(\hat{\beta}_1)}\Phi^{-1}(\alpha/2) < \beta_1\right) &= \alpha/2
\end{aligned}$$

  and

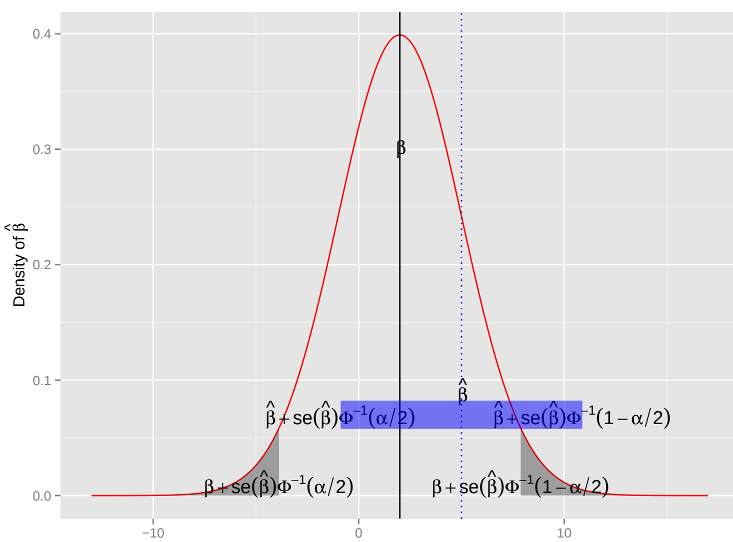
$$\begin{aligned}
\text{P}\left(\hat{\beta}_1 > \beta_1 + \sqrt{\text{Var}(\hat{\beta}_1)}\Phi^{-1}(1-\alpha/2)\right) &= \alpha/2 \\
\text{P}\left(\hat{\beta}_1 - \sqrt{\text{Var}(\hat{\beta}_1)}\Phi^{-1}(1-\alpha/2) > \beta_1\right) &= \alpha/2
\end{aligned}$$

so

$$
\begin{aligned}
&P\begin{pmatrix}\hat\beta_1+\sqrt{\mathrm{Var}(\hat\beta_1)}\Phi^{-1}(\alpha/2)<\beta_1\\ \beta_1<\hat\beta+\sqrt{\mathrm{Var}(\hat\beta_1)}\Phi^{-1}(1-\alpha/2)\end{pmatrix}=\\
=&1-P\left(\hat\beta_1+\sqrt{\mathrm{Var}(\hat\beta_1)}\Phi^{-1}(\alpha/2)<\beta_1\right)-\\
&-P\left(\hat\beta_1+\sqrt{\mathrm{Var}(\hat\beta_1)}\Phi^{-1}(1-\alpha/2)>\beta_1\right)\\
=&1-\alpha
\end{aligned}
$$

- For $\alpha=0.05$, $\Phi^{-1}(0.025)\approx -1.96$, $\Phi^{-1}(0.975)\approx 1.96$
- For $\alpha=0.1$, $\Phi^{-1}(0.05)\approx -1.64$

Confidence intervals



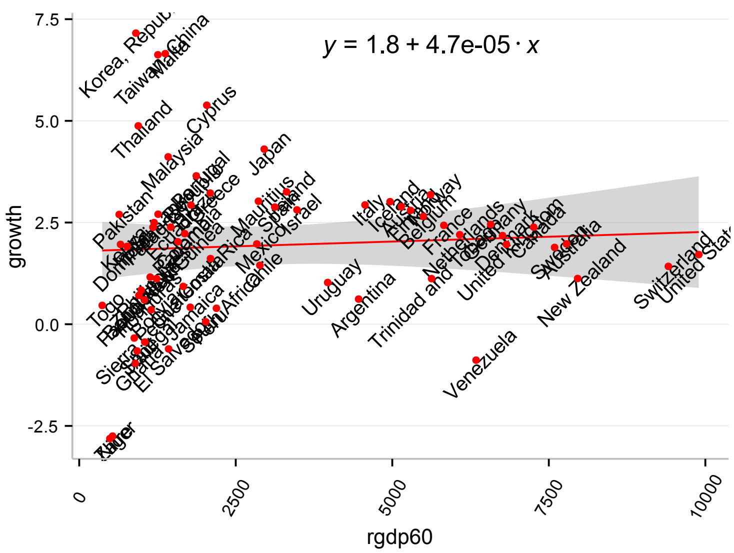
Confidence intervals

- $1-\alpha$ confidence interval
$$\hat\beta_1 \pm \sqrt{\mathrm{Var}(\hat\beta_1)}\Phi^{-1}(\alpha/2)$$
- With estimated $\hat\sigma^2_\epsilon$, use t distribution instead of normal
$$\hat\beta_1 \pm \sqrt{\widehat{\mathrm{Var}}(\hat\beta_1)}F^{-1}_{t,n-2}(\alpha/2)$$

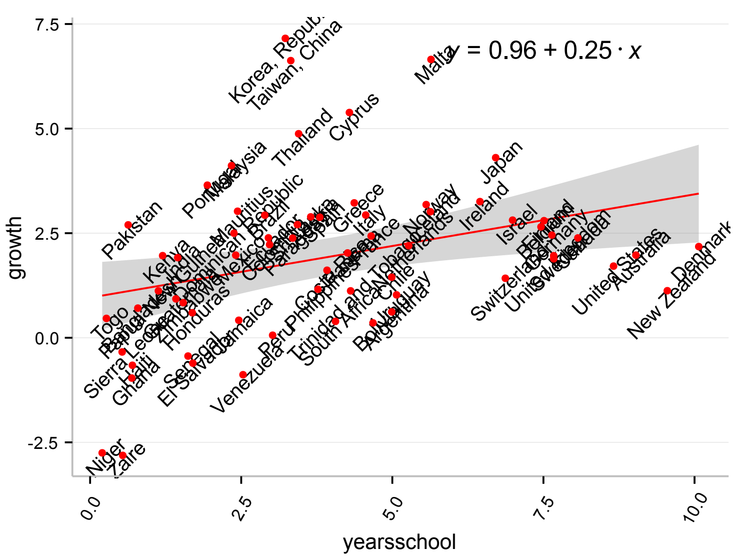
$F^{-1}_{t,n-2}$ = inverse CDF of $t(n-2)$ distribution

	$F_{t,n-2}(1-\alpha/2)$					
	$n-2$					
$\alpha/2$	5	10	20	50	100	∞
0.025	2.57	2.23	2.09	2.01	1.98	1.96
0.05	2.02	1.81	1.72	1.68	1.66	1.64

Example: GDP in 1960 and growth



Example: Years of schooling in 1960 and growth



10 Efficiency

Gauss-Markov theorem

- The sample regression estimator,

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

  also called *O*rdinary *L*east *S*quares (OLS) is not the only unbiased estimator of

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- *Gauss-Markov theorem:* if SLR.1-SLR.5, then OLS is the *B*est *L*inear *U*nbiased *E*stimator
  - Linear means linear in y, $\hat{\beta}_1 = \sum_{i=1}^{n} c_i y_i$ with $c_i = \frac{(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$
  - Unbiased means $\mathrm{E}[\hat{\beta}_1] = \beta_1$
  - Best means that among all linear unbiased estimators, OLS has the smallest variance

---

Proof: setup

- Let $\tilde{\beta}_1$ be a linear unbiased estimator of β_1
  - Linear: $\tilde{\beta}_1 = \sum_{i=1}^{n} c_i y_i$
  - Unbiased: $\mathrm{E}[\tilde{\beta}_1 | x_1, ..., x_n] = \beta_1$ (for all possible β_0, β_1)
- We will show that

$$\mathrm{Var}(\tilde{\beta}_1 | x_1, ..., x_n) \geq \mathrm{Var}(\hat{\beta}_1 | x_1, ..., x_n)$$

The c_i's in the definition of $\tilde{\beta}_1$ are allowed to be nonlinear functions of x_i. For example OLS uses $c_i = \frac{(x_i - \bar{x})}{\sum_{j=1}^{n} (x_j - \bar{x})^2}$.

Proof: outline

1. Show that $\sum_{i=1}^{n} c_i = 0$ and $\sum_{i=1}^{n} c_i x_i = 1$
2. Show $\mathrm{Cov}(\tilde{\beta}_1, \hat{\beta}_1 | x_1, ..., x_n) = \mathrm{Var}(\hat{\beta}_1 | x_1, ..., x_n)$
3. Show $\mathrm{Var}(\tilde{\beta}_1 | x_1, ..., x_n) \geq \mathrm{Var}(\hat{\beta}_1 | x_1, ..., x_n)$
4. Show $\mathrm{Var}(\tilde{\beta}_1 | x_1, ..., x_n) = \mathrm{Var}(\hat{\beta}_1 | x_1, ..., x_n)$ only if $\tilde{\beta}_1 = \hat{\beta}_1$

Theorem 5 (Gauss-Markov). *If SLR.1-SLR.5, then OLS is the Best Linear Unbiased Estimator. In other words if*

$$\tilde{\beta}_1 = \sum_{i=1}^{n} c_i y_i$$

is another linear (in y) unbiased ($\mathrm{E}[\tilde{\beta}_1 | x_1, ..., x_n] = \beta_1$) estimator, then

$$\mathrm{Var}(\tilde{\beta}_1 | x_1, ..., x_n) \geq \mathrm{Var}(\hat{\beta}_1 | x_1, ..., x_n)$$

where the inequality is strict unless $\tilde{\beta}_1 = \hat{\beta}_1$.

Proof. **Step 1: show that $\sum_{i=1}^{n} c_i = 0$ and $\sum_{i=1}^{n} c_i x_i = 1$.** We do this by using the fact that $\tilde{\beta}_1$ is unbiased. Note that

$$\begin{aligned}
\mathrm{E}[\tilde{\beta}_1 | x_1, ..., x_n] =& \mathrm{E}[\sum_{i=1}^{n} c_i y_i | x_1, ..., x_n] \\
& \text{substitute in the model for } y \\
=& \mathrm{E}[\sum_{i=1}^{n} c_i(\beta_0 + \beta_1 x_i + \epsilon_i) | x_1, ..., x_n] \\
& \text{use linearity of expectation} \\
=& \beta_0 \sum_{i=1}^{n} \mathrm{E}[c_i | x_1, ..., x_n] + \beta_1 \sum_{i=1}^{n} x_i \mathrm{E}[c_i | x_1, ..., x_n] + \sum_{i=1}^{n} \mathrm{E}[\epsilon_i c_i | x_1, ..., x_n] \\
& c_i \text{ is constant given } x_1, ..., x_n \\
=& \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} x_i c_i
\end{aligned}$$

Hence, unbiasedness requires

$$\beta_1 = \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} x_i c_i.$$

An estimator being unbiased means that it is unbiased for any possible value of the true parameters. For the previous equation to hold for all possible β_0 and β_1, we need $\sum_{i=1}^{n} c_i = 0$ and $\sum_{i=1}^{n} c_i x_i = 1$.

[1] We will go over the proof in class. See Marmer's slides or Wooldridge (2013) 3A for details

Step 2: show $\text{Cov}(\tilde{\beta}_1, \hat{\beta}_1|x_1, ..., x_n) = \text{Var}(\hat{\beta}_1|x_1, ..., x_n)$. Substituting in $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ let's us write $\tilde{\beta}_1$ as

$$\begin{aligned}
\tilde{\beta}_1 =& \sum_{i=1}^{n} c_i y_i \\
=& \sum_{i=1}^{n} c_i(\beta_0 + \beta_1 x_i + \epsilon_i) \\
=& \beta_0(\sum_{i=1}^{n} c_i) + \beta_1(\sum_{i=1}^{n} c_i x_i) + (\sum_{i=1}^{n} c_i \epsilon_i) \\
& \text{Using result of step 1} \\
=& \beta_1 + \sum_{i=1}^{n} c_i \epsilon_i
\end{aligned}$$

Similarly for $\hat{\beta}_1$,

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

Let $w_i = \frac{(x_i - \bar{x})}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$ so that we can write

$$\hat{\beta}_1 = \beta_1 + \sum_{i=1}^{n} w_i \epsilon_i.$$

Now let's calculate the covariance of $\tilde{\beta}_1$ and $\hat{\beta}_1$.

$$\begin{aligned}
\mathrm{Cov}(\tilde{\beta}_1, \hat{\beta}_1 | x_1, \ldots, x_n) =& \mathrm{E}[(\tilde{\beta}_1 - \beta_1)(\hat{\beta}_1 - \beta_1)] \\
=& \mathrm{E}\left[\left(\sum_{i=1}^{n} c_i \epsilon_i \right) \left(\sum_{i=1}^{n} w_i \epsilon_i \right) \middle| x_1, \ldots, x_n \right] \\
=& \sum_{i=1}^{n} c_i w_i \mathrm{E}[\epsilon_i^2 | x_1, \ldots, x_n] + \sum_{i=1}^{n} \sum_{j \neq i} c_i w_j \mathrm{E}[\epsilon_i \epsilon_j | x_1, \ldots, x_n] \\
& \text{second term} = 0 \text{ because of independent observations} \\
=& \sum_{i=1}^{n} c_i w_i \mathrm{E}[\epsilon_i^2 | x_1, \ldots, x_n] \\
& \text{using homoskedasticity} \\
=& \sigma^2 \sum_{i=1}^{n} c_i w_i \\
& \text{substituting } w_i = \frac{(x_i - \bar{x})}{\sum_{j=1}^{n} (x_j - \bar{x})^2} \\
=& \sigma^2 \sum_{i=1}^{n} c_i \frac{(x_i - \bar{x})}{\sum_{j=1}^{n} (x_j - \bar{x})^2} \\
=& \frac{\sigma^2}{\sum_{j=1}^{n} (x_j - \bar{x})^2} \left((\sum_{i=1}^{n} c_i x_i) - (\bar{x} \sum_{i=1}^{n} c_i) \right) \\
& \text{using result of step 1} \\
=& \frac{\sigma^2}{\sum_{j=1}^{n} (x_j - \bar{x})^2} (1 - \bar{x}0) \\
=& \mathrm{Var}(\hat{\beta}_1 | x_1, \ldots, x_n)
\end{aligned}$$

Step 3: show $\mathrm{Var}(\tilde{\beta}_1 | x_1, \ldots, x_n) \geq \mathrm{Var}(\hat{\beta}_1 | x_1, \ldots, x_n)$. Consider $\mathrm{Var}(\tilde{\beta}_1 - \hat{\beta}_1 | x_1, \ldots, x_n)$. We know that this is ≥ 0 because it is a variance. Also,

$$\begin{aligned}
\mathrm{Var}(\tilde{\beta}_1 - \hat{\beta}_1 | x_1, \ldots, x_n) =& \mathrm{Var}(\tilde{\beta}_1 | x_1, \ldots, x_n) + \mathrm{Var}(\hat{\beta}_1 | x_1, \ldots, x_n) - 2\mathrm{Cov}(\tilde{\beta}_1, \hat{\beta}_1 | x_1, \ldots, x_n) \\
& \text{Using result of step 2} \\
=& \mathrm{Var}(\tilde{\beta}_1 | x_1, \ldots, x_n) - \mathrm{Var}(\hat{\beta}_1 | x_1, \ldots, x_n)
\end{aligned}$$

Hence,

$$\begin{aligned}
\mathrm{Var}(\tilde{\beta}_1 - \hat{\beta}_1 | x_1, \ldots, x_n) = \mathrm{Var}(\tilde{\beta}_1 | x_1, \ldots, x_n) - \mathrm{Var}(\hat{\beta}_1 | x_1, \ldots, x_n) \geq 0 \\
\mathrm{Var}(\tilde{\beta}_1 | x_1, \ldots, x_n) \geq \mathrm{Var}(\hat{\beta}_1 | x_1, \ldots, x_n)
\end{aligned}$$

Step 4: show $\mathrm{Var}(\tilde{\beta}_1 | x_1, \ldots, x_n) = \mathrm{Var}(\hat{\beta}_1 | x_1, \ldots, x_n)$ **only if** $\tilde{\beta}_1 = \hat{\beta}_1$ This step is only needed for the "where the inequality is strict only if $\tilde{\beta}_1 = \hat{\beta}_1$ part of the theorem. If $\mathrm{Var}(\tilde{\beta}_1 | x_1, \ldots, x_n) = \mathrm{Var}(\hat{\beta}_1 | x_1, \ldots, x_n)$, then

$$\mathrm{Var}(\tilde{\beta}_1 - \hat{\beta}_1 | x_1, \ldots, x_n) = \mathrm{Var}(\tilde{\beta}_1 | x_1, \ldots, x_n) - \mathrm{Var}(\hat{\beta}_1 | x_1, \ldots, x_n) = 0,$$

so $\tilde{\beta}_1 - \hat{\beta}_1$ is constant. Since both are unbiased, the constant difference must be zero. Thus, $\tilde{\beta}_1 = \hat{\beta}_1$. $\square$

References

Abbring, Jaap. 2001. "An Introduction to Econometrics: Lecture notes." URL http://jabbring.home.xs4all.nl/courses/b44old/lect210.pdf.

Angrist, J.D. and J.S. Pischke. 2009. *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Angrist, Joshua D and Jörn-Steffen Pischke. 2014. *Mastering 'Metrics: The Path from Cause to Effect*. Princeton University Press.

Baltagi, BH. 2002. *Econometrics*. Springer, New York. URL http://gw2jh3xr2c.search.serialssolutions.com/?sid=sersol&SS_jc=TC0001086635&title=Econometrics.

Beck, T., R. Levine, and N. Loayza. 2000. "Finance and the Sources of Growth." *Journal of financial economics* 58 (1):261–300. URL http://www.sciencedirect.com/science/article/pii/S0304405X00000726.

Bierens, Herman J. 2012. "The Two-Variable Linear Regression Model." URL http://personal.psu.edu/hxb11/LINREG2.PDF.

Diez, David M, Christopher D Barr, and Mine Cetinkaya-Rundel. 2012. *OpenIntro Statistics*. OpenIntro. URL http://www.openintro.org/stat/textbook.php.

Linton, Oliver B. 2017. *Probability, Statistics and Econometrics*. Academic Press. URL http://gw2jh3xr2c.search.serialssolutions.com/?sid=sersol&SS_jc=TC0001868500&title=Probability%2C%20statistics%20and%20econometrics.

Stock, J.H. and M.W. Watson. 2009. *Introduction to Econometrics, 2/E*. Addison-Wesley.

Wooldridge, J.M. 2013. *Introductory econometrics: A modern approach*. South-Western.