# Lecture 6: The simple linear regression model

Nic Rivers

Advanced quantitative research methods, API6319
Fall 2019

# Simple linear regression

- Linear regression models are used to quantify the relationship between a dependent variable and one or more independent variables.
- The 'simple' in Simple linear regression refers to the fact that we only have two variables.
- In reality, we know that outcomes we are interested in can be affected by multiple variables.
- The simple linear regression is a building block towards multiple linear regression, which can control for (some of) these other variables.

# Example: Cigarette excise taxes

- There is an interest in setting appropriate cigarette excise taxes:
  - Discourage smoking
  - Raise revenues
  - But not too high to encourage black market

# Cigarette excise tax effect

- For public policy, it would be very useful to have an estimate of the likely response to a change in the excise tax. One part of the response is the change in consumption resulting from the tax. Define this as:

$$\beta_{CT} = \frac{\text{change in cigarette consumption}}{\text{change in cigarette excise tax per pack}}$$

  (the $_{CT}$ subscript stands for "cigarette tax").
- If we know $\beta_{CT}$, we can figure out the change in cigarette consumption as excise tax changes:

$$\Delta\text{cigarette consumption} = \beta_{CT}\Delta\text{tax rate}$$

- Suppose $\beta_{CT} = -2$. This means that increasing the cigarette excise tax by \$1 per pack reduces cigarette consumption by 2. (Cigarette consumption is measured as the percent of adults who smoke regularly)

# Cigarette tax effect: predictions

- The definition of $\beta_{CT}$ is the definition for the slope of a straight line
- The straight line can be written:

$$\text{cigarettes} = \beta_0 + \beta_{CT} \times \text{tax}$$

- In this equation, $\beta_0$ is the intercept. (What does it mean?)
- If this equation holds, and we know $\beta_0$ and $\beta_{CT}$, we can predict the *change* in cigarette consumption from a *change* in the excise tax rate. We could also predict the average cigarette consumption for a region for a given tax rate.

# Other factors - errors

- We can postulate a straight-line relationship between cigarette consumption and the excise tax
- However, we know that many factors affect cigarette consumption other than the excise tax (like what?).
- So a better way to summarize the relationship might be:

$$\text{cigarettes} = \underbrace{\beta_0 + \beta_{CT} \times \text{tax}}_{\text{average effect}} + \underbrace{\text{other factors}}_{\text{individual effect}}$$

# Visualizing the relationship

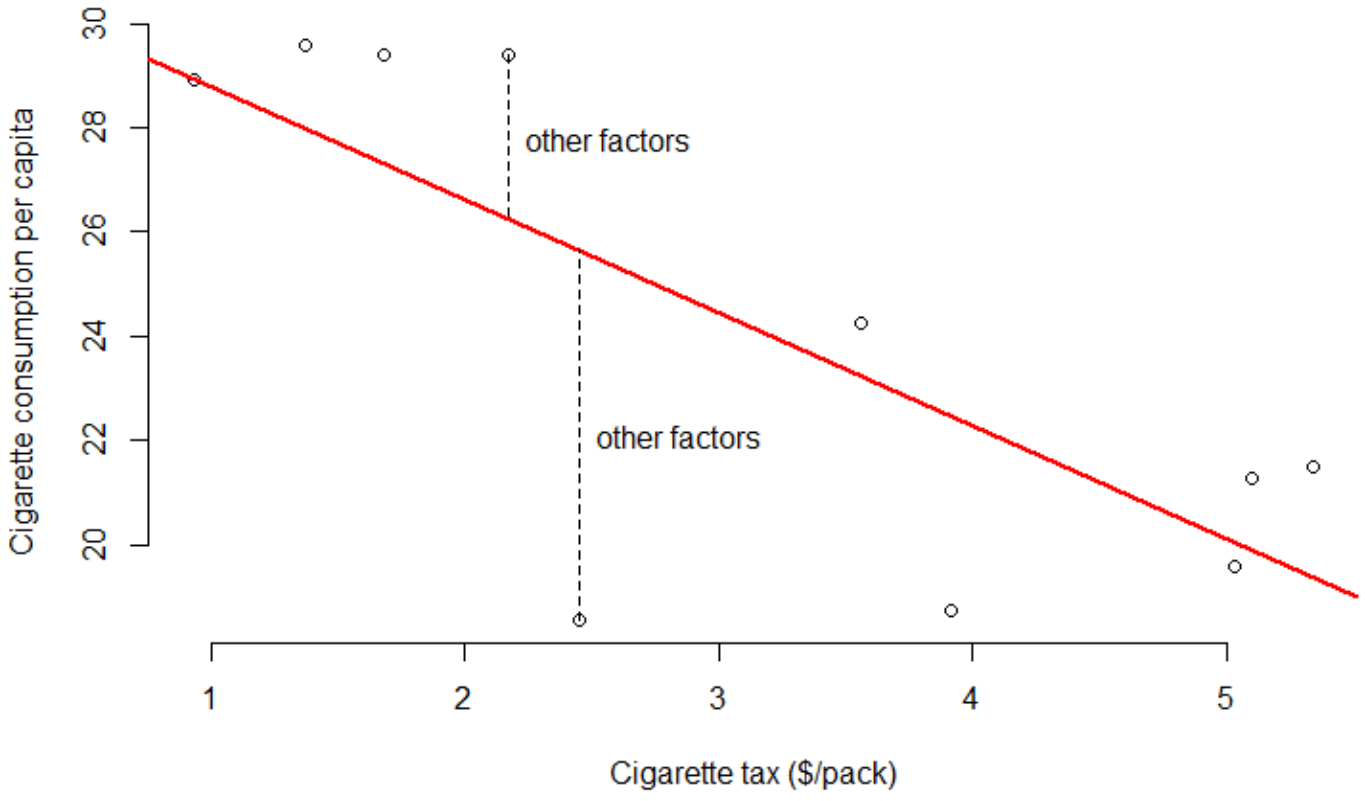
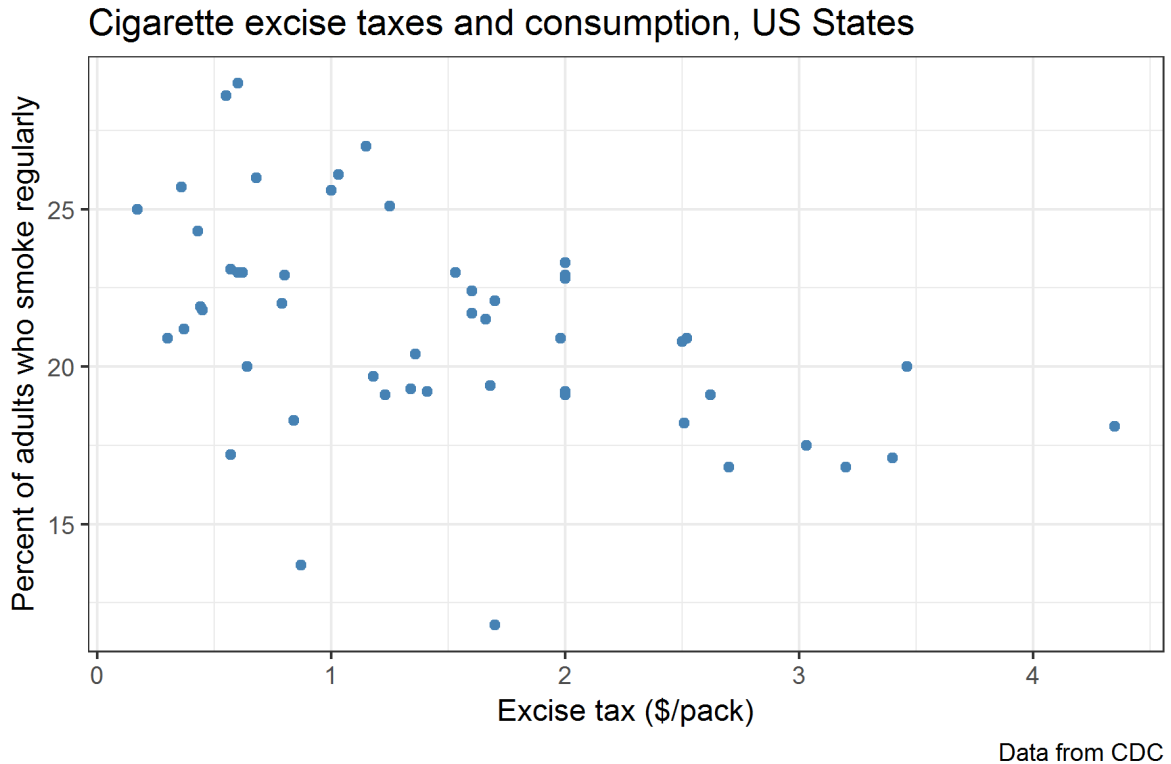



Figure: Cigarette consumption and excise taxes, US states (Hypothetical)

# Estimating $\beta_{CT}$

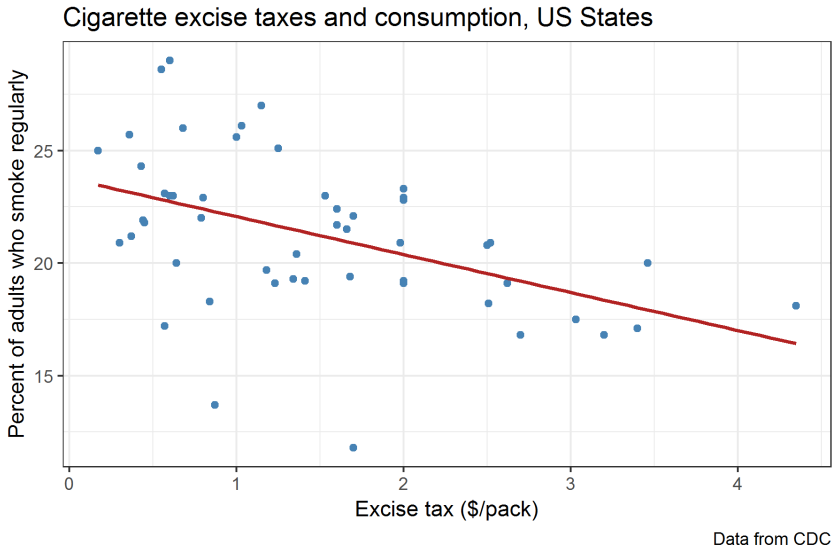
- One way to get some insight into $\beta_{CT}$ is to look at cigarette excise taxes and consumption
- Data from US states



# Estimating the effect of cigarette taxes on consumption

- We don't observe the true relationship between excise taxes and cigarette consumption
- We do have a **sample** of data (from US states)
- We will use our sample of data to estimate the relationship in the sample between excise tax and cigarette consumption: we call this **regression**
- Using this estimated relationship, we will make **inferences** about the relationship in the population
- As with estimates of the mean, we will express our (lack of) **confidence** in our estimates about the relationship we observe

# Regression results




- Estimated sample regression line is: cigarettes $= 23.7 - 1.7 \times$ tax
- A \$1 increase in cigarette tax is associated with a reduction in smoking rates by 1.7 percentage points.
- This equation is a *conditional mean*: mean cigarette consumption conditional on excise tax
- The sample regression line is an estimate of the population (true) regression line

# Definition

- We are interested in studying how some variable $y$ changes with another variable $x$.
  - How an early childhood education program ($x$) affects income later in life ($y$)
  - Whether the presence of a large natural resource endowment ($x$) affects the level of corruption in a country ($y$)
  - Whether tougher sentences ($x$) reduce crime rates ($y$)
- A way to capture such a relationship is with the *simple linear regression model*:

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

# Nomenclature: True regression model

$$y = \beta_0 + \beta_1 x + u \tag{1}$$

- In (1), $y$ is called the **dependent variable**, and $x$ is called the **independent variable** or the **explanatory variable**.
- The first part, $\beta_0 + \beta_1 x$, is called the **population regression line**. This is the relationship that holds on average between $x$ and $y$ in the population.
- The **intercept** $\beta_0$ is the average value of $y$ in the population when $x = 0$
- The **slope** $\beta_1$ is the average change in $y$ in the population associated with a one unit change in $x$
- The variable $u$ is called the **error term** or **disturbance**. It captures all factors other than $x$ that affect $y$.
- Consider the case where $y$ is student test score (at country level) and $x$ is average educational expenditure. What sign do you expect on $\beta_1$? What 'other factors' might $u$ capture?

# Interpreting the simple regression model

- Consider a specific example:

$$grade = \beta_0 + \beta_1 classsize + u \tag{2}$$

  which suggests a relationship between class size (measured in number of students per class) and grade scores (measured on a scale of 0 to 100).
- Here, the parameter $\beta_1$ captures the change in grade scores that occurs when class size is changed. What might be a reasonable value for $\beta_1$?

# Interpreting the simple regression model

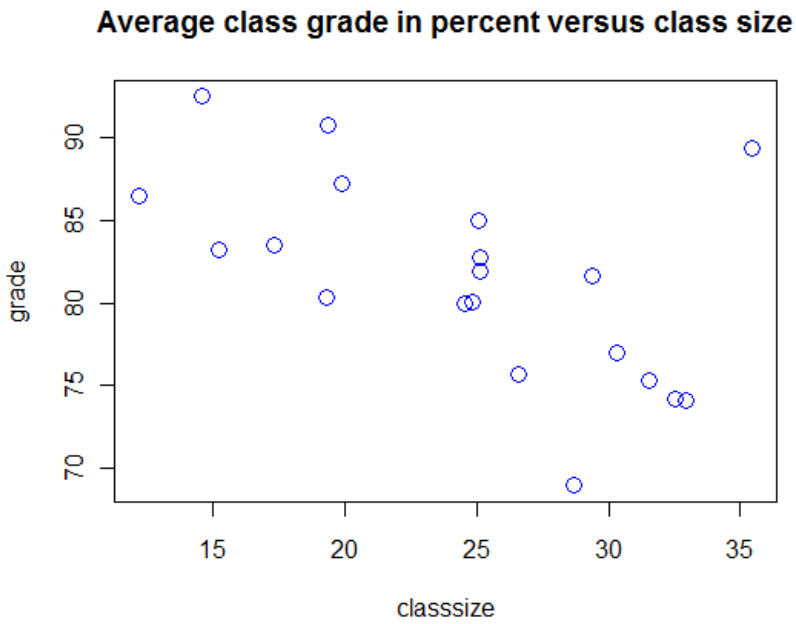
- Consider a specific example:

$$grade = \beta_0 + \beta_1 classsize + u \tag{2}$$

  which suggests a relationship between class size (measured in number of students per class) and grade scores (measured on a scale of 0 to 100).
- Here, the parameter $\beta_1$ captures the change in grade scores that occurs when class size is changed. What might be a reasonable value for $\beta_1$?
- If $\beta_1 = -0.5$, what does the model suggest is the impact of adding 4 more students to a class, holding all other factors equal?
- Note that the model suggests a **linear** relationship: each increment in class size has the same effect on grades. This is why we refer to the model as a **linear** regression model.
- The parameter $\beta_0$ is not normally the focus of our attention. Here, it measures the grade for a hypothetical class of zero students.

# Estimating the model

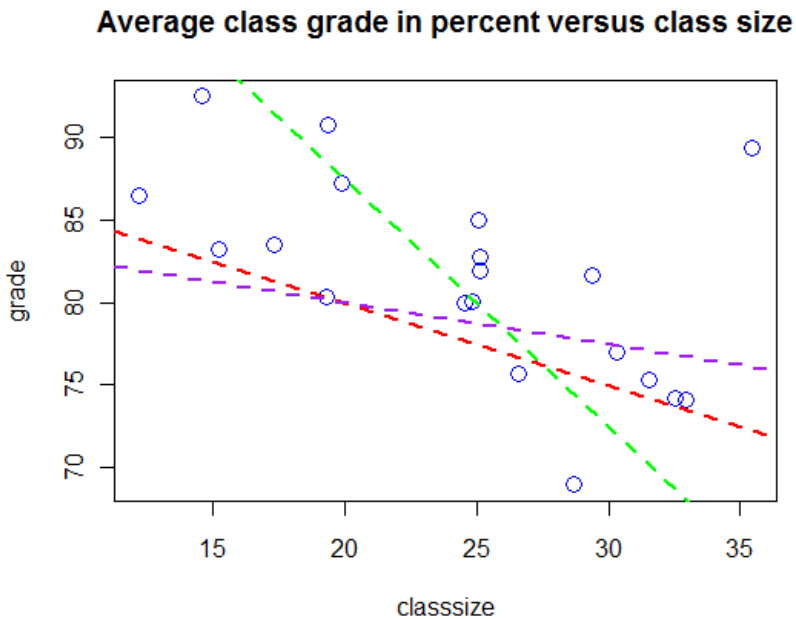
- Knowing $\beta_0$ and $\beta_1$ (i.e., the population regression line) can be very useful for policy analysis
- But we don't know these
- However, we do have a sample of data: we can use that to estimate the unknown slope and intercept of population regression line
- Suppose a sample of 20 classes from the population can produces the following data set:



# Estimating the model

- We aim to choose coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ that "fit" the sampled data well, since these should provide a good approximation of the true relationship in the population.
- We define **residuals** as the distance between the sample regression function and the data points. Note:
  - Residuals: difference between estimated regression line and data
  - Errors: difference between true population regression line and data
- It is natural to think that a good fit should minimize residuals.

Do any of these lines seem like 'good' candidates for summarizing the data?



# Ordinary least squares

- We want to choose $\hat{\beta}_0$ and $\hat{\beta}_1$ so that estimated regression line is as "close" as possible to observed data.

# Ordinary least squares

- We want to choose $\hat{\beta}_0$ and $\hat{\beta}_1$ so that estimated regression line is as "close" as possible to observed data.
- (Note the 'hats' – these distinguish the sample regression line from the true regression line).

# Ordinary least squares

- We want to choose $\hat{\beta}_0$ and $\hat{\beta}_1$ so that estimated regression line is as "close" as possible to observed data.
- (Note the 'hats' – these distinguish the sample regression line from the true regression line).
- How should we measure "closeness"?

# Ordinary least squares

- We want to choose $\hat{\beta}_0$ and $\hat{\beta}_1$ so that estimated regression line is as "close" as possible to observed data.
- (Note the 'hats' – these distinguish the sample regression line from the true regression line).
- How should we measure "closeness"?
- The ordinary least squares (OLS) estimator chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ so that the sum of squared residuals is minimized.

# Ordinary least squares

- We want to choose $\hat{\beta}_0$ and $\hat{\beta}_1$ so that estimated regression line is as "close" as possible to observed data.
- (Note the 'hats' – these distinguish the sample regression line from the true regression line).
- How should we measure "closeness"?
- The ordinary least squares (OLS) estimator chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ so that the sum of squared residuals is minimized.
- The OLS regression line is called the **sample regression line**

# Ordinary least squares

- We want to choose $\hat{\beta}_0$ and $\hat{\beta}_1$ so that estimated regression line is as "close" as possible to observed data.
- (Note the 'hats' – these distinguish the sample regression line from the true regression line).
- How should we measure "closeness"?
- The ordinary least squares (OLS) estimator chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ so that the sum of squared residuals is minimized.
- The OLS regression line is called the **sample regression line**
- Using the sample regression line, we can predict values of $y$ for different values of $x$

# Ordinary least squares

- We want to choose $\hat{\beta}_0$ and $\hat{\beta}_1$ so that estimated regression line is as "close" as possible to observed data.
- (Note the 'hats' – these distinguish the sample regression line from the true regression line).
- How should we measure "closeness"?
- The ordinary least squares (OLS) estimator chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ so that the sum of squared residuals is minimized.
- The OLS regression line is called the **sample regression line**
- Using the sample regression line, we can predict values of $y$ for different values of $x$
- All statistical programs have functions that construct OLS estimators from data

# Finding regression coefficients

- We won't calculate regression coefficients by hand - R will do this for us.
- When R calculates OLS regression coefficients, it is finding the line that minimizes the sum of the squared difference between the sample regression line and the data points.
- This involves finding estimates for $\hat{\beta}_1$ and $\hat{\beta}_2$.

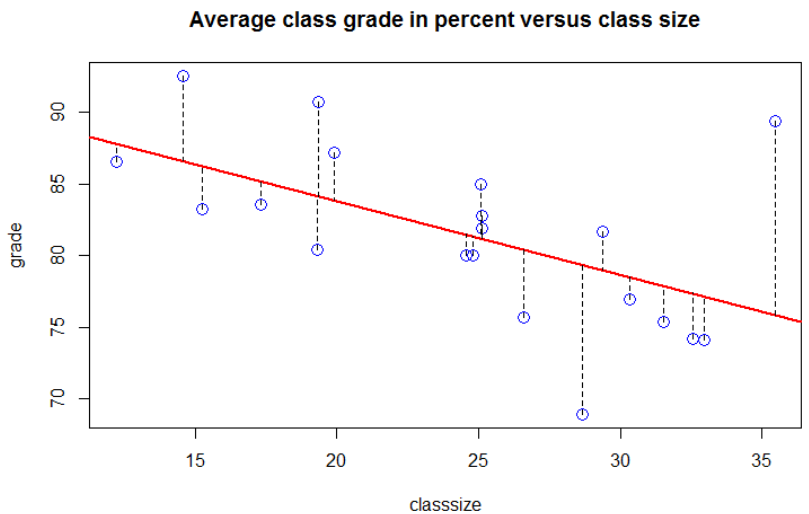
# Ordinary least squares

- The ordinary least squares (OLS) estimate minimizes the sum of squared residuals.
- When we talk about estimating a regression, we normally mean that we have estimated a regression using OLS.
- Fitted regression line:

$$\text{grade} = 94.1 - 0.513 \times \text{classsize}$$

- We can use this to predict grades in classes of different sizes

- Predictions will not be perfect because many factors influence grades
- Prediction does give average grade in class absent other factors
- Regression line is a *conditional mean*



# Measures of fit

- Once we've estimated a regression, it would be nice to know how well that regression line describes the data
- Are observations tightly clustered around regression line? Or spread out?

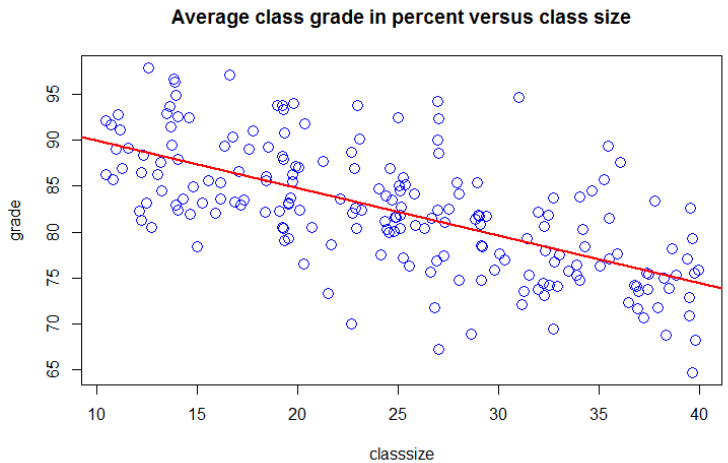



Figure: A worse fit

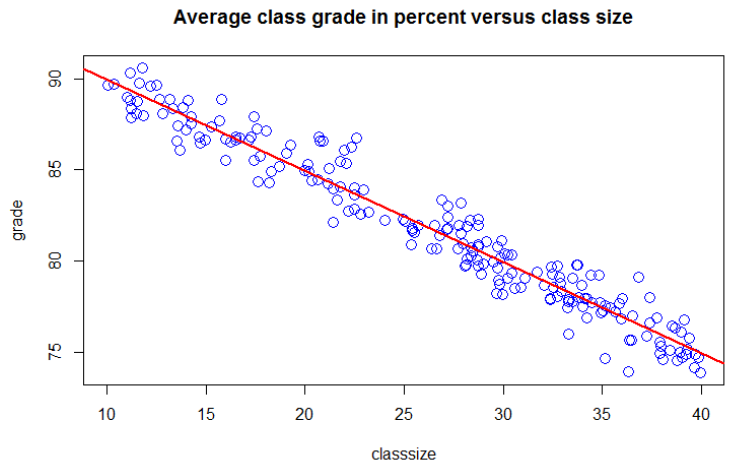



Figure: A better fit

# OLS: Some definitions

Total Sum of Squares (SST) is a measure of the total amount of variance in the dependent variable:

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

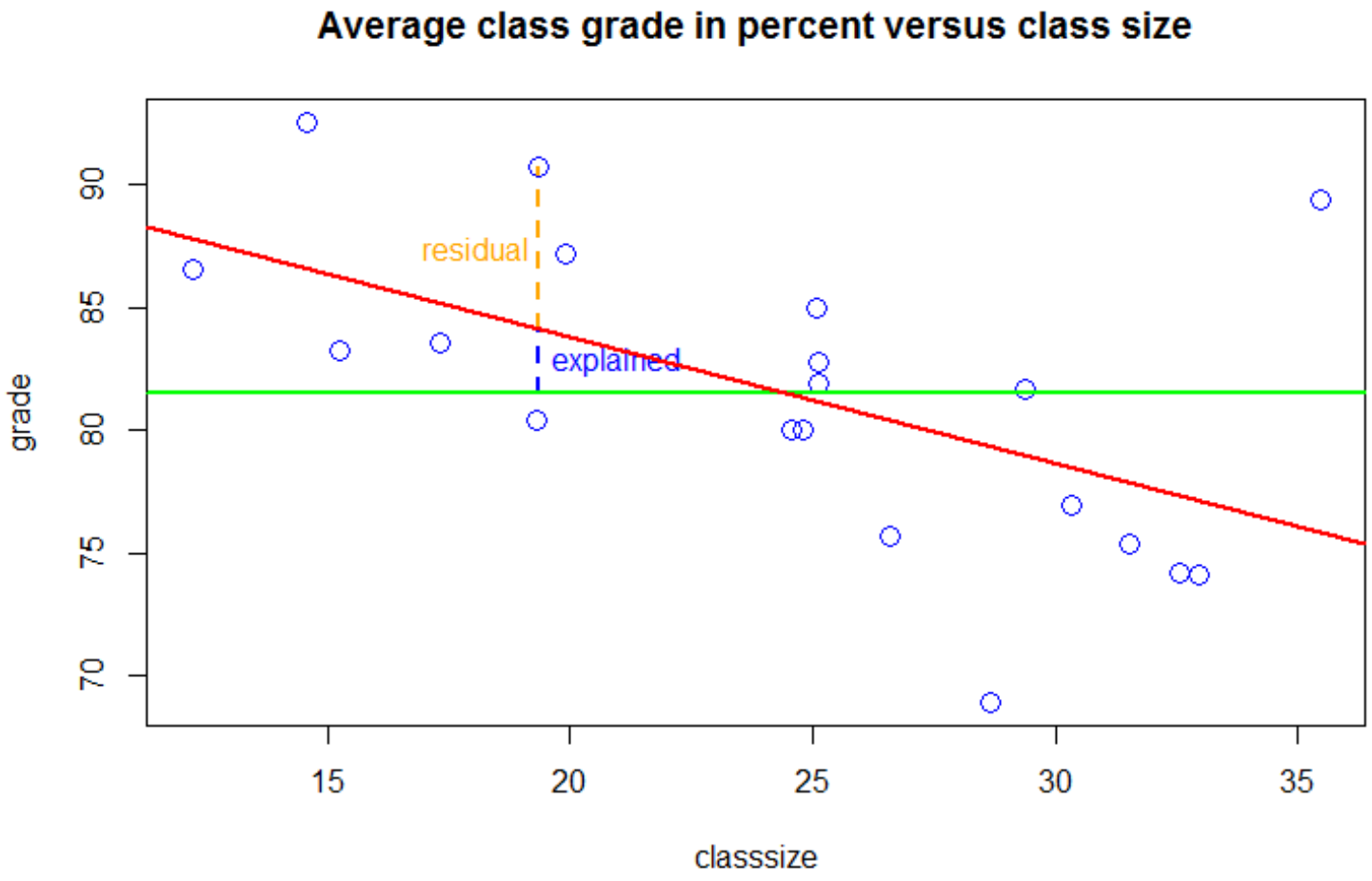
Explained Sum of Squares (SSE) is a measure of the variance in the dependent variable that is explained by the independent variable:

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

Residual Sum of Squares (SSR) is a measure of the variance in the dependent variable that is not explained by the independent variable:

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

# Definitions in pictures



# Goodness of fit

- We often summarize the amount of variance in the dependent variable that is explained by the independent variable using *R*-**squared**:

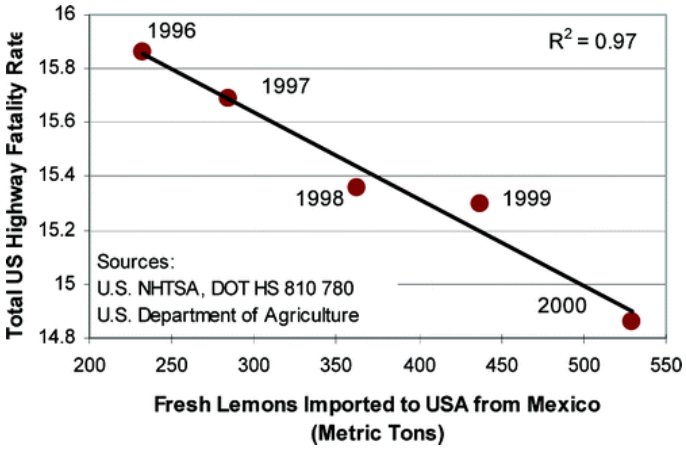
$$R^2 = \frac{SSE}{SST}$$

- $R^2$ can vary between 0 and 1. A value of 0 implies that none of the variance in the dependent variable is explained by changes in the independent variable (the two are completely unrelated). A value of 1 implies that all of the variance in the dependent variable is explained by changes in the independent variable (the dependent variable is completely predicted by the independent variable without any residual).

# Interpreting goodness of fit

- For simple linear regression (one independent variable), $R^2$ is the correlation coefficient (r) squared
- $R^2$ tells us how good the regression line is in predicting values for the dependent variable given the independent variable
- Even if $R^2$ is low, the relationship that we are examining may still be important. For example, gender explains only a small portion of differences in wages between individuals, but it is still an important effect.
- $R^2$ is **not** related to causality: two variables can be highly correlated and have no causal relationship; two variables can be loosely correlated and be causally linked.

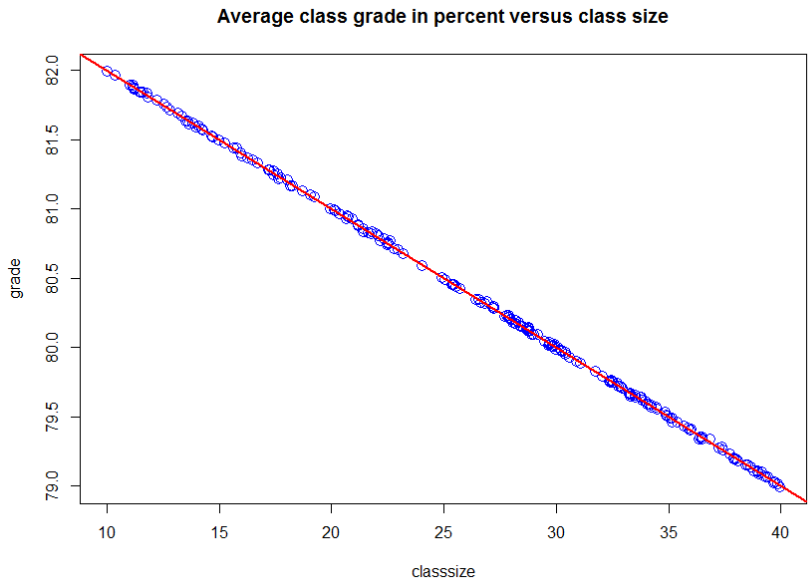
# Correlation and causation

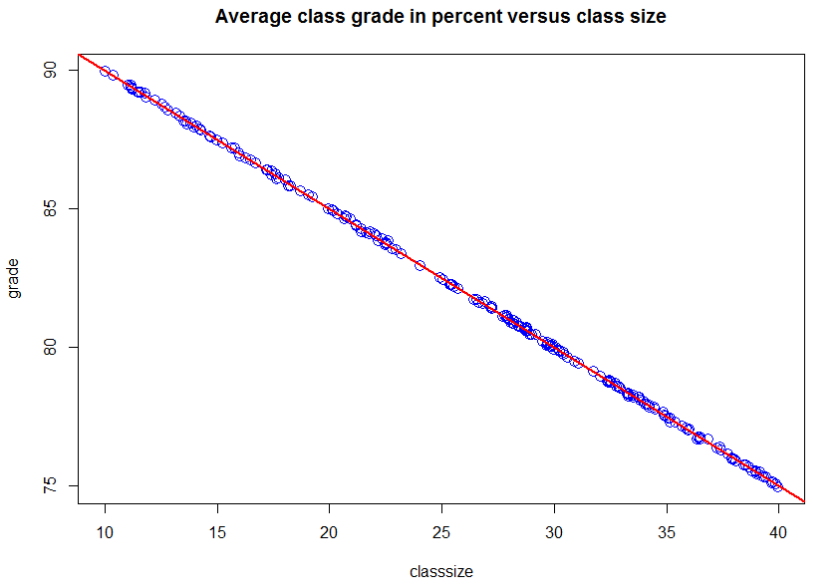



Regression of hourly earnings on gender, LFS2018October: $R^2$=0.017.

# $R^2$: where does it fit in?

- $R^2$ measures how tightly points are clustered around the regression line; closely related to $R^2$ (square)
- It does not measure the slope of the relationship
- Two regressions with similar $R^2$ can have totally different implications.



# Standard error of the regression

- The *standard error of the regression* is the typical[1] deviation between the estimated regression line and the data.
- This is sometimes referred to as the *root mean squared error*.
- It gives us a measure of how far off - on average - the predictions of the model are from reality.

[1] The square root of the average squared deviation

# Population and sample regression functions

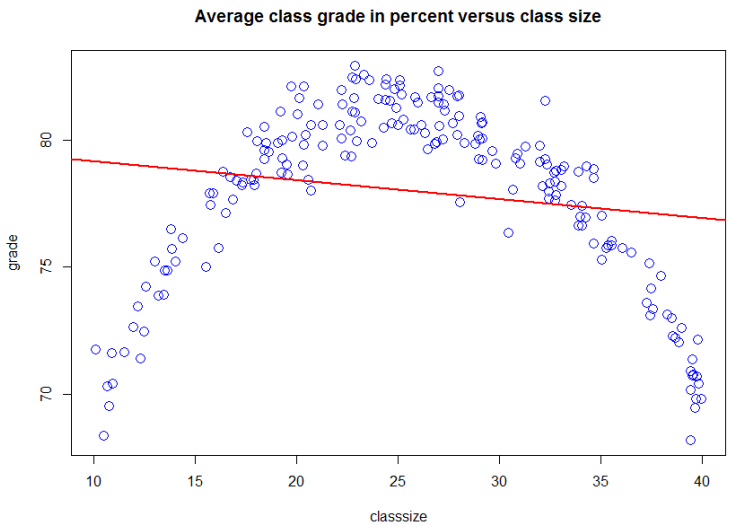
- The population regression function (PRF) describes the true average relationship that we are investigating: $y = \beta_0 + \beta_1 x + u$.
- The sample regression function (SRF) is an attempt to estimate the PRF.
- Because we estimate the SRF from a finite sample rather than the whole population, the SRF is an estimate of the PRF. We denote the fact that is an estimate by using a $\hat{\cdot}$ over the coefficients: $y = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{u}$.
- Given several assumptions, the SRF will be an **unbiased** estimator of the PRF. This means that on average, if we took a large number of samples, the coefficients from the SRF would equal the coefficients in the PRF. In a single sample, they may not be equal (think of the parallel with estimating the mean from a sample).

# Assumptions required for SRF to be unbiased estimator of PRF (1)

**Assumption 1**

The underlying PRF is linear in parameters: $y = \beta_0 + \beta_1 x + u$

- If PRF is non-linear, then a simple linear regression will fail to estimate coefficients



# Assumptions required for SRF to be unbiased estimator of PRF (2)

**Assumption 2**

The sample upon which we are estimating the relationship is representative of the population (i.e., it is a random sample).

- This is the same assumption required for consistently estimating the population mean
- If there are systematic problems with the sample, then the SRF will be a biased estimator of the PRF
- Earlier in the semester, we referred to this concept as "external validity"

# Assumptions required for SRF to be unbiased estimator of PRF (3)

**Assumption 3**

There is variation in $x$.

- If there is no variation in $x$, it is impossible to know the relationship between $x$ and $y$
- More variation in $x$ makes it easier to identify relationship between $x$ and $y$
- This is important to consider in choosing a research topic (all quantitative research is about comparison – you need something to compare to)

# Assumptions required for SRF to be unbiased estimator of PRF (4)

**Assumption 4 (really important one)**

The error $u$ has mean of zero for any value of explanatory variable.

- Consider the PRF: $y = \beta_0 + \beta_1 x + u$.
- This says that 'other factors' ($u$) do not systematically vary with $x$. In other words, there is no relationship between $u$ and $x$.
- There will always be "other factors" that determine $y$. What matters is whether these other factors vary systematically with $x$ or not.
- Think about a randomized controlled experiment (e.g., drug trial). In this context, the researcher randomly assigns $x$ across the population. Thus, even if 'other factors' influence the outcome, there is no relationship between $u$ and $x$.
- In an observational study, for the SRF to be a good estimator of the PRF, the researcher has to make the case that $x$ is assigned 'as if' randomly.
- This is the same as the "fourth hurdle" we talked about earlier in the semester (confounding factors).

# When assumption #4 is violated

- If $u$ varies systematically with $x$, then the SRF will not be a good estimator of the PRF
- Example: imagine trying to estimate the effect of education on income:

$$\text{income} = \beta_0 + \beta_1 \text{educ} + \text{other factors}$$

- It is possible that individuals with high intelligence obtain lots of education, thus there is a correlation between education and 'other factors.' But high intelligence probably also helps people earn lots of money.
- In this case, a simple regression of income on education is unlikely to produce a good estimate of $\beta_1$.
- (Note that there are always "other factors". This alone doesn't mean that we shouldn't trust the regression. The problem arises if "other factors" vary with $x$.)

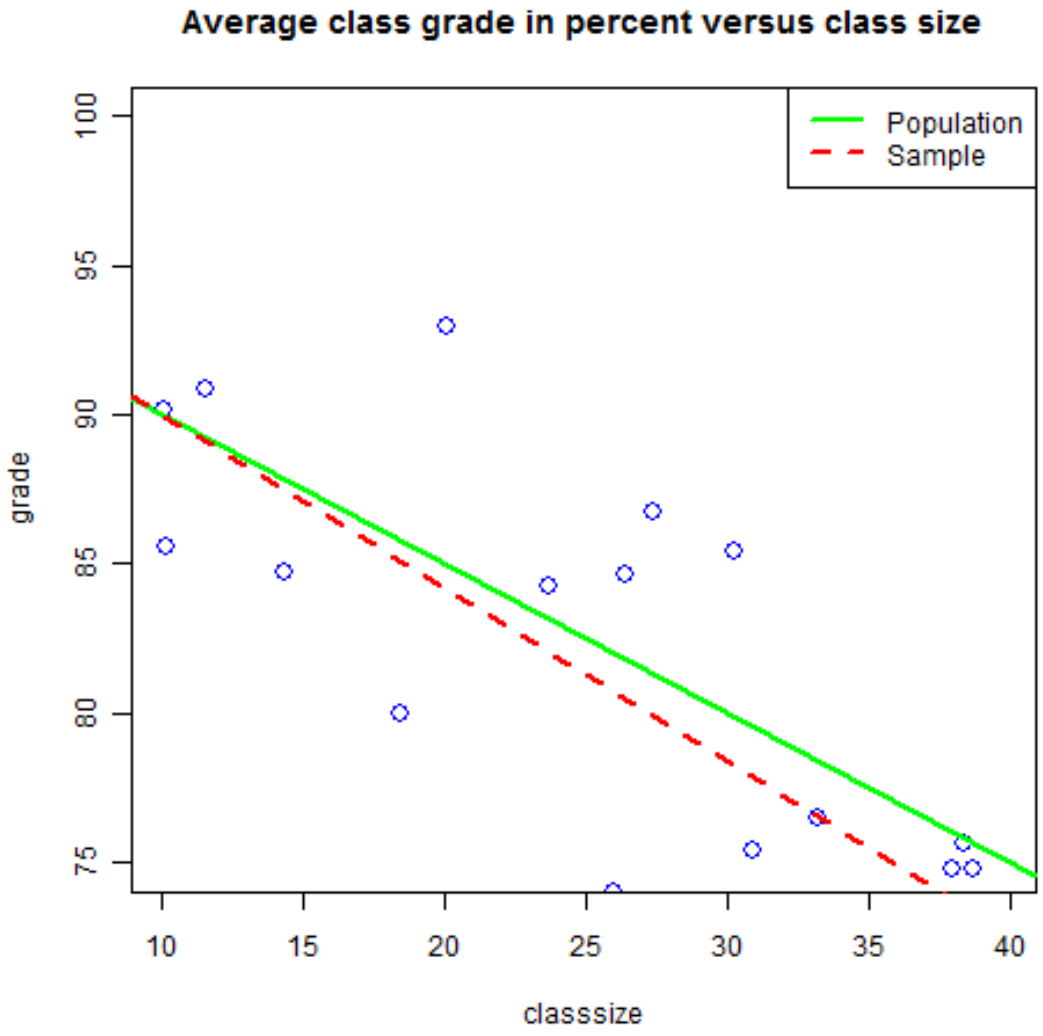
# When assumptions hold

- When the OLS assumptions hold, the SRF will be an unbiased estimate of the PRF
- When the OLS assumptions hold, we can interpret $\hat{\beta}_1$ as an estimate of the effect of a change in $x$ on the value of $y$, *holding all other factors constant*.
- If the "four hurdles" can be cleared also, then $\hat{\beta}_1$ is the causal effect of $x$ on $y$.
- Note that assumption #4 is shared between "4 hurdles for causality" and "4 assumptions for SRF to be unbiased estimator of PRF".

# OLS estimates as random variables

- OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are computed from a sample of the population
- If a different random sample were drawn, the estimators would end up with different values
- Same idea when estimating population mean

# Population and sample regression functions



# Central limit theorem

- When estimating the mean, we appealed to the central limit theorem
- This says that an estimate of the mean is a normally-distributed random variable, no matter what the distribution of the underlying sample
- We use the same theorem for estimates of $\beta_0$ and $\beta_1$. This allows us to generate a measure of confidence in our regression estimates, and to use them in hypothesis tests.

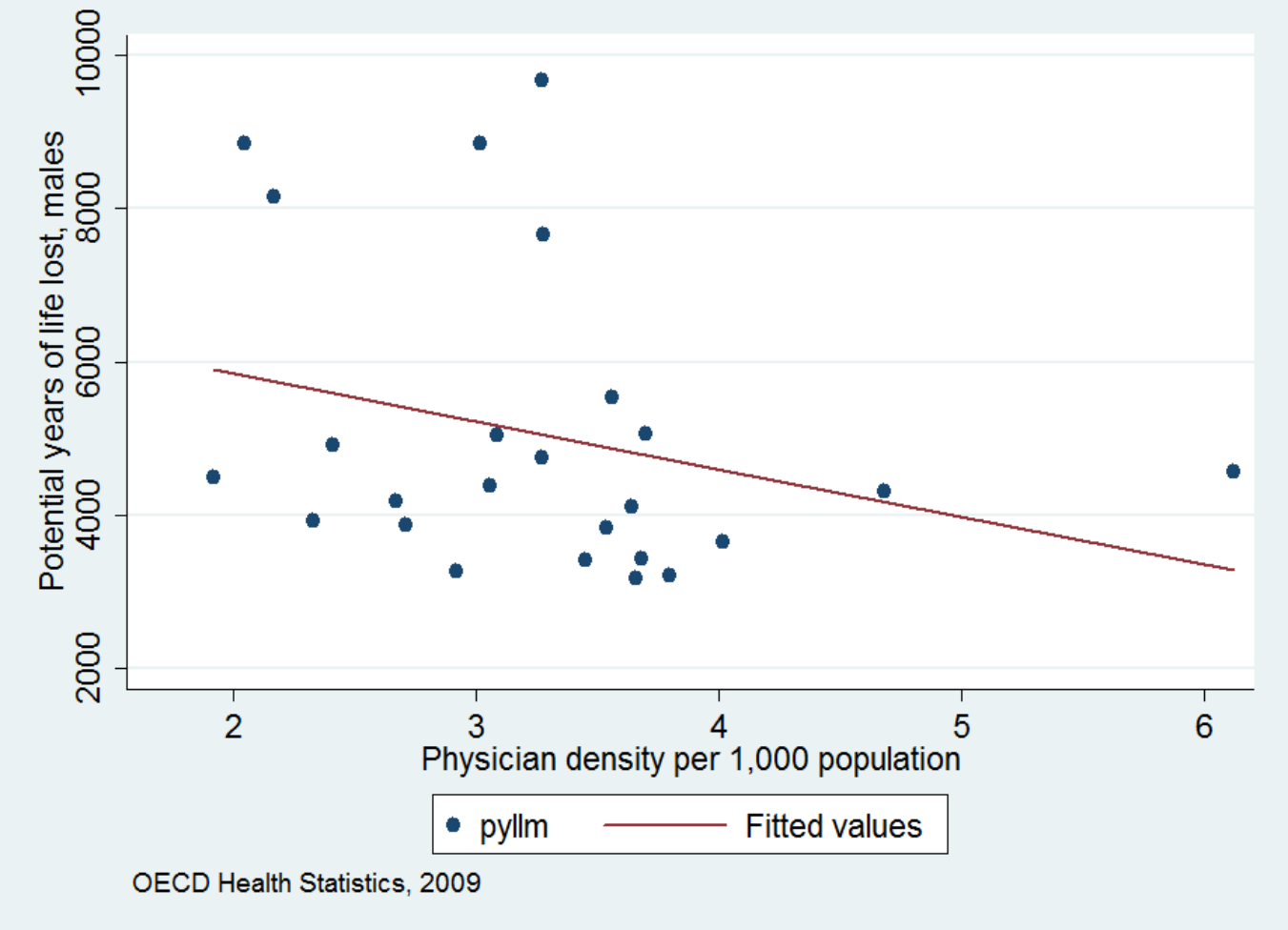
# Confidence in estimates of the mean (refresher)

- In the single variable case, we often want to estimate the mean, and provide some confidence about our estimate
- We can use the sample average as an estimator of the population average
- We can calculate the standard deviation in the sample: $sd = \sqrt{\frac{1}{N}\sum_i (x_i - \bar{x})}$
- We can calculate the standard error of the mean estimate: $se = \frac{sd}{\sqrt{N}}$
- We can use the standard error to express our confidence in the mean estimate, and to test hypotheses

# Testing hypotheses about regression coefficients

- In a similar way that we tested hypotheses about the mean, we can test hypotheses about regression coefficients
- Just as estimates of the mean are normally distributed, estimates of regression coefficients are normally distributed
- Null hypothesis: there is no relationship between $x$ and $y$
- We can use a $t$ test based on the standard error of regression coefficients to test the hypothesis
- Same rules as we learned earlier in the semester.
- Similarly, we can construct confidence intervals around regression estimates

# An example: physicians and health




Did this regression satisfy OLS assumptions? Is this the causal effect of physicians on health?

# Interpreting results

- Assume that the regression did satisfy OLS assumptions.
- We can look at parameters:

| | Estimate | Standard error |
|---|---|---|
| Intercept | 11,609 | 362 |
| Slope | -1,455 | 142 |

- What do these mean?
- What is the null hypothesis?

# Interpreting results

- Assume that the regression did satisfy OLS assumptions.
- We can look at parameters:

| | Estimate | Standard error |
|---|---|---|
| Intercept | 11,609 | 362 |
| Slope | -1,455 | 142 |

- What do these mean?
- What is the null hypothesis?
  - $H_0$: increasing the number of physicians has no effect on years of lost life
- Can we reject the null hypothesis? With what degree of confidence?
- What is a 95% confidence interval?

# Practical vs. statistical significance

- Our focus has been on assessing the "statistical significance" of our estimates
- Generally, in a regression context, the slope parameter is declared "statistically significant" if it differs from 0 with 95 percent confidence (this is evidence that there is a real relationship and not just a quirk in the data)
- But a statistically significant relationship doesn't necessarily indicate a practically significant relationship
- Recall the relationship between standard error and sample size
- With a large sample size, we can detect relationships with more and more precision
- Lesson - think about practical as well as statistical significance

# Reporting regression results

- Regression results are presented in a "standard" fashion because they are so widely used
- Typical to do the following:
  - Present coefficient and (standard error in brackets below)
  - Present $R^2$ value and number of observations
  - Present stars for statistical significance
  - Present results for a number of comparable models side by side (different samples, different variable definitions)
- Try R packages huxtable or stargazer.

| | log income per capita in 2000 CE | | | | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Predicted diversity (ancestry adjusted) | 541.792*** (130.250) | 248.699*** (86.798) | | | 524.240*** (172.284) | 374.297** (189.015) |
| Predicted diversity square (ancestry adjusted) | −387.026*** (91.148) | −172.552*** (61.446) | | | −370.660*** (123.664) | −264.700* (137.333) |
| Predicted diversity (unadjusted) | | | 140.903*** (51.614) | 10.152 (52.732) | −1.063 (74.681) | −67.278 (84.783) |
| Predicted diversity square (unadjusted) | | | −107.686*** (38.133) | −7.418 (38.000) | −2.002 (57.317) | 52.844 (67.248) |
| Continent fixed effects | No | Yes | No | Yes | No | Yes |
| Observations | 143 | 143 | 143 | 143 | 143 | 143 |
| $R^2$ | 0.13 | 0.47 | 0.08 | 0.45 | 0.14 | 0.48 |
| *p*-value for joint significance of linear and quadratic terms in: | | | | | | |
| Adjusted diversity | | | | | 0.010 | 0.039 |
| Unadjusted diversity | | | | | 0.419 | 0.748 |

*Notes*: This table establishes that, when explaining log income per capita in 2000 CE, the ancestry-adjusted measure of genetic diversity outperforms the unadjusted measure in terms of (i) the qualitative robustness of the hump-shaped effect to continent fixed effects, and (ii) maintaining explanatory power in regressions that perform a horse race between the two measures of diversity. Bootstrap standard errors, accounting for the use of generated regressors, are reported in parentheses.

*** Significant at the 1 percent level.

# Estimating a regression with R

- Use the `lm()` function in R.
- This function expects a regression "formula" and a data set.
- Specify the regression formula in this format:
  `variableY ~ variableX`.
- So the code in R might look like:
  `lm(income ~ education, data=lfs_data)`
- You can assign a regression to an object (call it whatever you want):
  `my_regression <- lm(income ~ education, data=lfs_data)`

# R regression output

- You can look at the regression output using:
  `summary(my_regression)`
- You can also get nicely formatted regression output, which you can include in a word document or in a RMardown document, using packages that are designed for format regression output. For example,
  `library(huxtable)`
  `huxreg(my_regression)`
  will produce a nicely formatted regression table.

# Example

- Candidate spending on elections is very large. In 2004, candidates in the (Canadian) federal election spent over $100 million. Spending in the 2012 US election was several billion dollars; spending in the US continues to grow very quickly.
- There are fairness concerns about large campaign spending
- Partly in response to these concerns, Canada has introduced campaign spending limits.
- To determine the impact of these rules, as well as other potential rules, it is necessary to know the impact of campaign spending on election results.
- We use a data set covering US campaign spending and election results in two-party contests from the 1998 House of Representatives election

# A statistical model

Key variables in the data set:

| | |
|---|---|
| voteA | The percentage of votes received by candidate A |
| expshareA | The percentage of total campaign expenditures due to candidate A |

Estimate a simple regression model with 173 observations using the following model:

$$voteA = \beta_0 + \beta_1 share_A + u$$

Questions:

- What is the interpretation of $\beta_0$?
- What is the interpretation of $\beta_1$?
- Do you think that there are any problems with estimating this equation? (OLS assumptions? 4 hurdels)

# A statistical model

Key variables in the data set:

| | |
|---|---|
| voteA | The percentage of votes received by candidate A |
| expshareA | The percentage of total campaign expenditures due to candidate A |

Estimate a simple regression model with 173 observations using the following model:

$$voteA = \beta_0 + \beta_1 share_A + u$$

Questions:

- What is the interpretation of $\beta_0$?
- What is the interpretation of $\beta_1$?
- Do you think that there are any problems with estimating this equation? (OLS assumptions? 4 hurdels)
  - What if there is a variable 'likeability' included in $u$. We would expect more likeable candidates to garner more of the vote, and perhaps also more campaign donations. So our regression would confound likeability with expenditures.

# A statistical model

Key variables in the data set:

| | |
|---|---|
| voteA | The percentage of votes received by candidate A |
| expshareA | The percentage of total campaign expenditures due to candidate A |

Estimate a simple regression model with 173 observations using the following model:

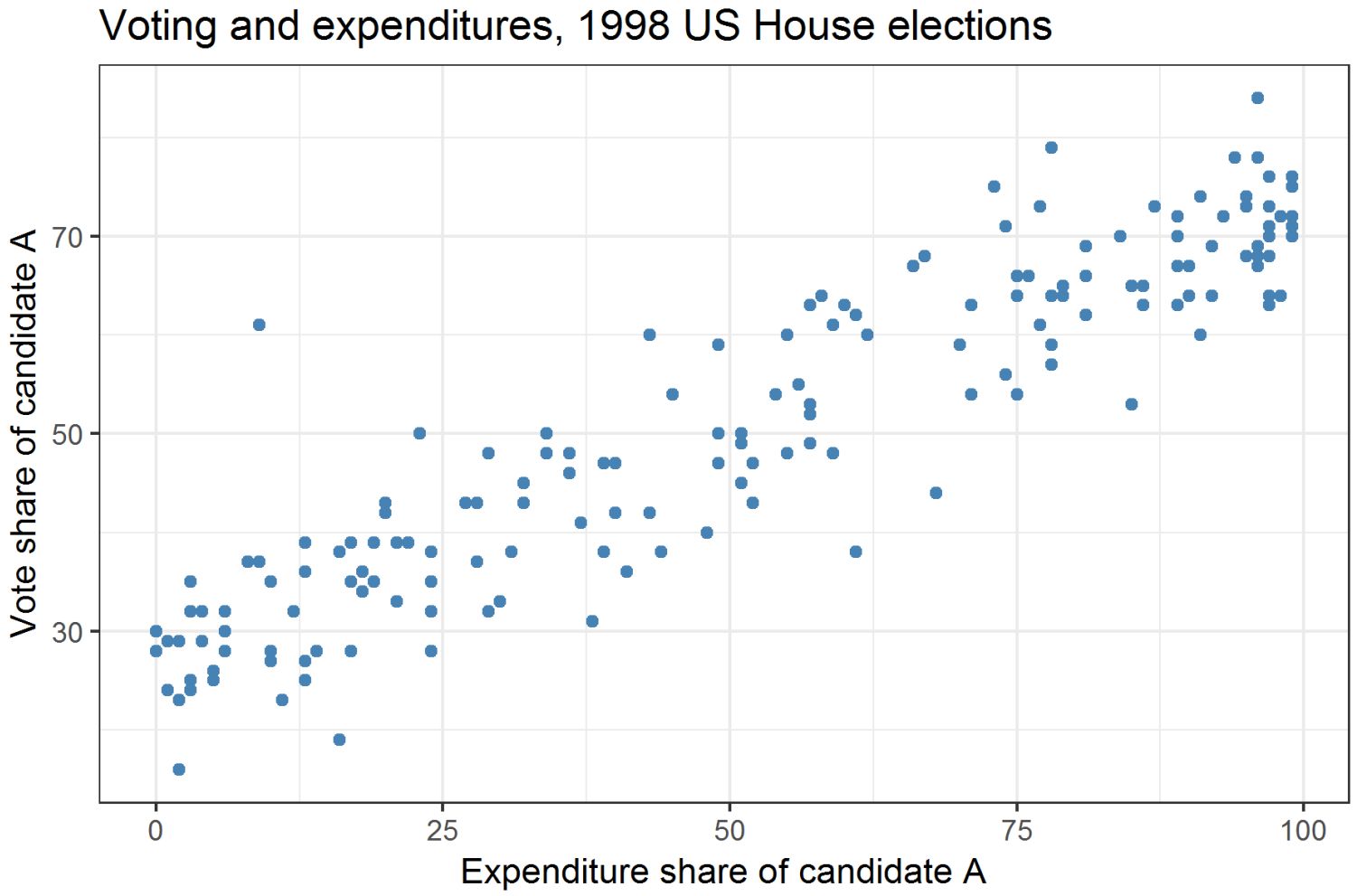
$$voteA = \beta_0 + \beta_1 share_A + u$$

Questions:

- What is the interpretation of $\beta_0$?
- What is the interpretation of $\beta_1$?
- Do you think that there are any problems with estimating this equation? (OLS assumptions? 4 hurdels)
  - What if there is a variable 'likeability' included in $u$. We would expect more likeable candidates to garner more of the vote, and perhaps also more campaign donations. So our regression would confound likeability with expenditures.
  - We can still interpret the regression as an association, but need to be cautious about inferring causation.

# What does the data look like?

`Excel: Insert - Scatter Chart`



# Regression analysis

```
> summary(lm(votea ~ sharea, data=dat))

Call:
lm(formula = votea ~ sharea, data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
-17.1021  -4.1062  -0.2452   3.5727  30.0370

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 26.78513    0.88813   30.16   <2e-16 ***
sharea       0.46421    0.01456   31.89   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.386 on 171 degrees of freedom
Multiple R-squared:  0.8561,    Adjusted R-squared:  0.8552
F-statistic:  1017 on 1 and 171 DF,  p-value: < 2.2e-16
```

# Coefficient estimates

- The first column gives the estimated model coefficients. Here, the model is:

$$\widehat{votea} = 26.81 + 0.464 shareA$$

  The 'hat' on the variable indicates that this is a predicted value. This says that a one unit increase in the *shareA* (i.e., a one percent increase in the share of expenditures) is associated with a 0.464 unit increase in *votea* (i.e., a 0.464 percent increase in the share of the vote).
- The second column is the standard error of the coefficient estimate. It gives us an indication of how precisely the coefficient is estimated. Low standard errors imply high precision.
- The t statistic is the t-value associated with 0. Essentially, this value measures the number of standard errors the value of zero is from the mean (coefficient) estimate. A large value means that zero is distant from the coefficient estimate.
- The P-value is the t-test that compares zero to the coefficient estimate.
- The 95% confidence intervals report the lower and upper 95% confidence intervals. Interpretation: if we drew another sample from the same population, 95 times out of 100 the coefficient estimate would fall in the 95%-CI.

# Example

- The file ‘‘`fertility-contraception.csv`’’ contains estimates of the fertility rate (number of children born per woman) and estimates of use of contraception for 15-49 year old women at the country level
- I obtained the data from the World Bank, and constructed averages of each variable from the 2000-2015 period
- Based on the data, what happens to fertility rates when the prevalence of contraception use increases by 10 percentage points? How confident are you in your estimate?
- What assumptions are required to treat this as a causal estimate?

# Example 2

- The file `schoolSizes.xls` contains the average class size (number of students per $3^{rd}$ grade class) in each public elementary school in the Ottawa-Carleton District schoolboard as well as the percentage of grade 3 students in each school that have reached the provincial level of math proficiency.
- I obtained the data from the EQAO.
- Do smaller classes cause better math scores?
- (Note: you can read an excel file using the package readxl and the function read_excel)