Welcome to Mastering Metrics!

Econ 140, Section 1

Jonathan Old

Roadmap

1. Getting Started & Housekeeping
2. Time for your questions
3. Econ 140: The Big Picture
4. Correlation and Causation

Getting Started & Housekeeping

Introduce yourself …

… and an interesting fact about you that you would like us to know!

Housekeeping

- What you can expect in this section and from me as GSI:

Housekeeping

- What you can expect in this section and from me as GSI:
  - Safe Space

Housekeeping

- What you can expect in this section and from me as GSI:
  - Safe Space
  - Diversity

Housekeeping

- What you can expect in this section and from me as GSI:
  - Safe Space
  - Diversity
  - Challenges in this course

Housekeeping

- What you can expect in this section and from me as GSI:
  - Safe Space
  - Diversity
  - Challenges in this course
  - Mental Health

Housekeeping

- What you can expect in this section and from me as GSI:
  - Safe Space
  - Diversity
  - Challenges in this course
  - Mental Health
- Core values: Transparency, Integrity, Respect

Housekeeping

- What you can expect in this section and from me as GSI:
  - Safe Space
  - Diversity
  - Challenges in this course
  - Mental Health
- Core values: Transparency, Integrity, Respect
- I am your learning companion. I see grades as feedback, not as an evaluation of your abilities or personality.

Your questions

Any questions?

… Remember – this is a safe space! Every question is useful!

Econ 140: The Big Picture

The Big Picture

- Why is applied econometrics so cool and so important?
- Why do we NOT use Excel? A cautionary tale
- Why econometrics and not data science?
    - Causality vs prediction
    - Economic models

Correlation and Causation

Dissecting Bad Causal Claims

Discuss in groups of 2: Why is this statement problematic?



Over the past 60 years, more spending on police hasn't meant less crime

The Washington Post

Intuitively, one might worry that reducing police spending would lead to a spike in crime. A review of spending on state and local police over the past 60 years, though, shows no correlation nationally between spending and crime rates.

In 1960, about $2 billion was spent by state and local governments on police. There were about 1,887 crimes per 100,000 Americans, including 161 violent crimes. By 1980, spending had increased to $14.6 billion — and crime rates had soared to 5,950 crimes per 100,000 Americans and 597 violent crimes. Over the next two decades, those rates thankfully fell, down to about 4,120 crimes per 100,000 people and 507 violent crimes. Spending spiked to more than $87 billion. Eighteen years later — by 2018, the most recent year for which full data are available — crime rates had fallen further to 2,580 crimes per 100,000, including 381 violent crimes.

Figure 1: Police spending and Crime (Source)

Dissecting Bad Causal Claims II

Discuss in groups of 2: Why is this statement problematic?

Three's a crowd: Having more than 2 kids linked to weaker brain function

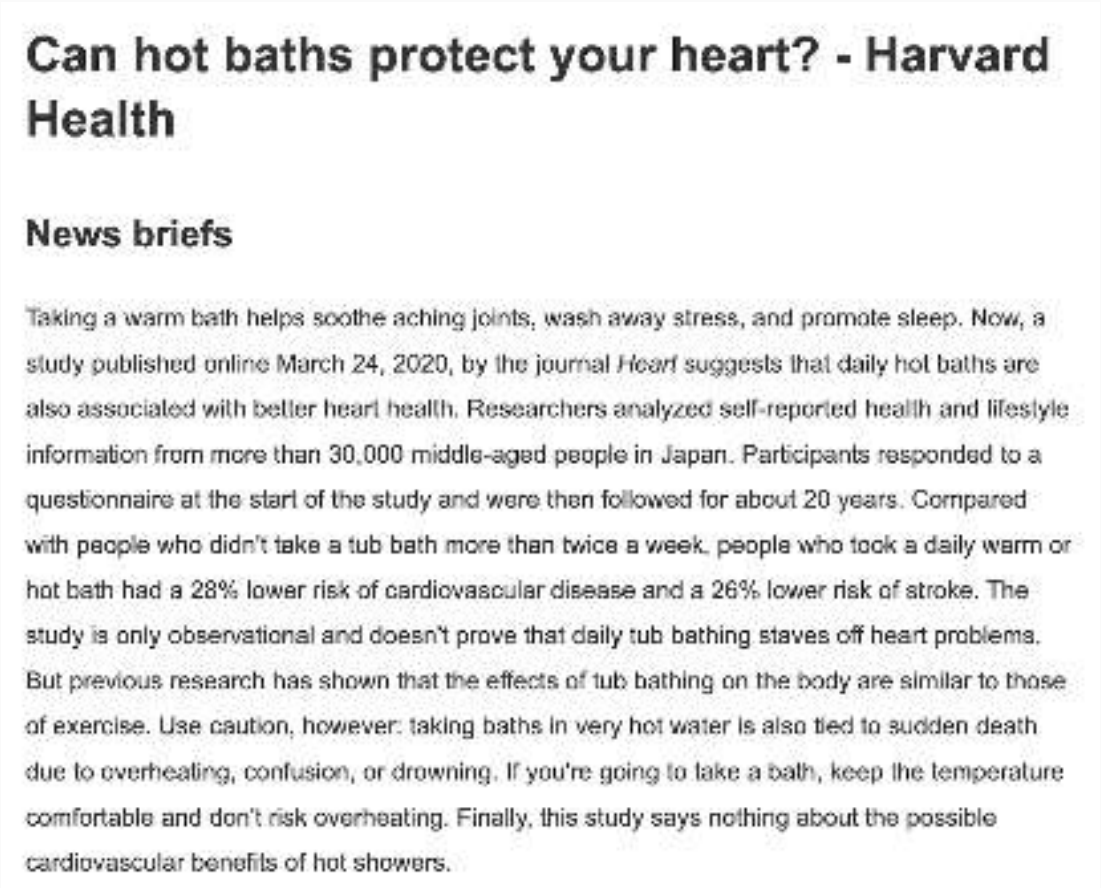


NEW YORK – Everything in moderation – even kids? Researchers from Columbia University and Université Paris-Dauphine report having more than two kids may have a negative impact on late-life cognition. The study shows that older parents with just two children appeared sharper cognitively than those with three.

Figure 2: Number of Children and Cognitive Function(Source)

Dissecting Bad Causal Claims III

Discuss in groups of 2: Why is this statement problematic?



Can hot baths protect your heart? - Harvard Health

News briefs

Taking a warm bath helps soothe aching joints, wash away stress, and promote sleep. Now, a study published online March 24, 2020, by the journal *Heart* suggests that daily hot baths are also associated with better heart health. Researchers analyzed self-reported health and lifestyle information from more than 30,000 middle-aged people in Japan. Participants responded to a questionnaire at the start of the study and were then followed for about 20 years. Compared with people who didn't take a tub bath more than twice a week, people who took a daily warm or hot bath had a 28% lower risk of cardiovascular disease and a 26% lower risk of stroke. The study is only observational and doesn't prove that daily tub bathing staves off heart problems. But previous research has shown that the effects of tub bathing on the body are similar to those of exercise. Use caution, however: taking baths in very hot water is also tied to sudden death due to overheating, confusion, or drowning. If you're going to take a bath, keep the temperature comfortable and don't risk overheating. Finally, this study says nothing about the possible cardiovascular benefits of hot showers.

Figure 3: Hot baths and health (Source)

Dissecting Bad Causal Claims IV

Discuss in groups of 2: Why is this statement problematic?



The New York Times

Another Benefit to Going to Museums? You May Live Longer

Researchers in Britain found that people who go to museums, the theater and the opera were less likely to die in the study period than those who didn't.

Leonardo da Vinci's "The Virgin and Child" at the Louvre in Paris. A British study found that people who engaged with the arts had a lower likelihood of dying during the study period than those who did not. Thibault Camus/Associated Press

Figure 4: Museums and longevity (Source)

Dissecting Bad Causal Claims IV

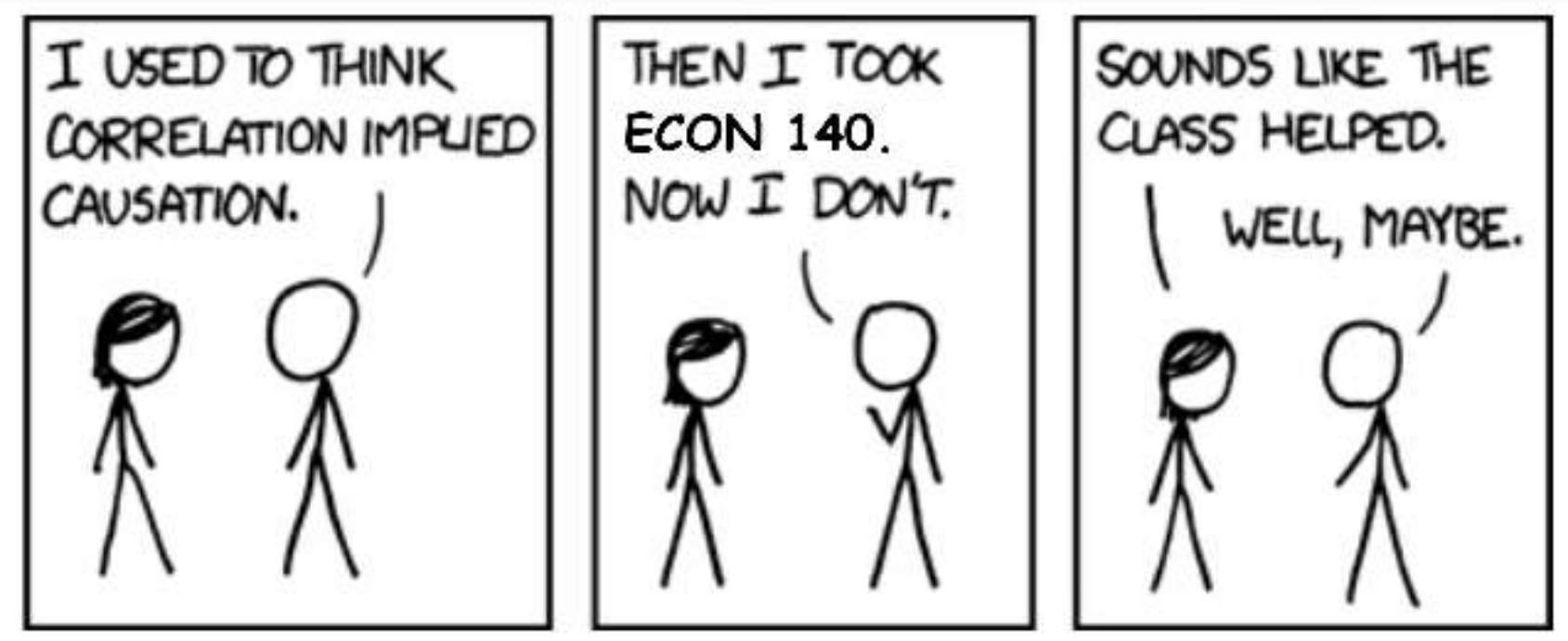


Figure 5: Correlation and causation

Quiz Time!

We have seen that
“Correlation Does Not Imply Causation”
What about:
“No Correlation Implies No Causation”?

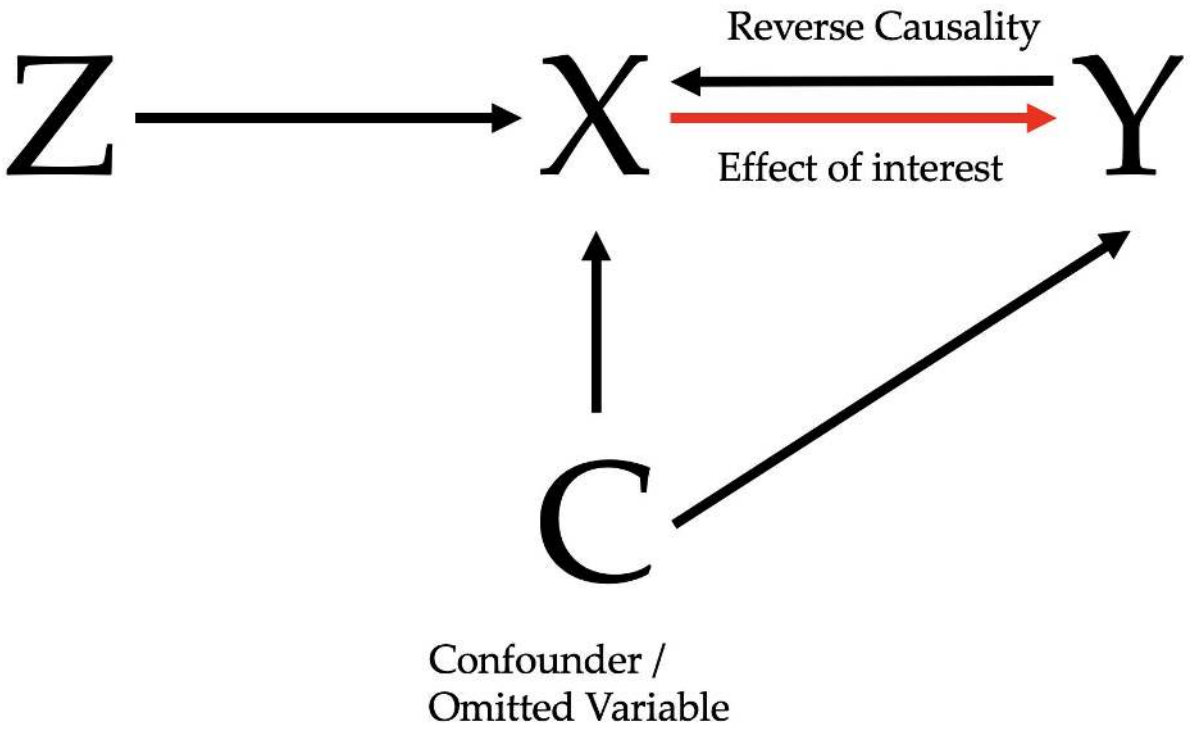
On Counterfactuals, Clones, and Cool (?) Regression Stuff

https://pollev.com/jonathanold608

Roadmap

- Your questions
- Recap: Selection bias and counterfactuals
- A closer look: how RCTs solve selection bias problem
- A small intro to regression and OLS

Recap - Selection Bias



Potential outcomes

Y_{0i} = Outcome of individual i, with "status" 0

Y_{1i} = Outcome of individual i, with "status" 1



Counterfactual outcomes!

- "Status" can be anything:
  - Treatment assignment
  - Actual treatment
  - Having 0 or 1 children
  - Drinking expensive whiskey or not

BUT: We **NEVER** observe an individual at more than one status at the same time! ➔ The fundamental problem of unobservability!

Does giving you an iPad improve your test scores?
Potential outcomes

Difference in Potential Outcomes

Group

Potential outcome if i has iPad

$$E[Score_{1i} - Score_{0i} | iPad_i = 0]$$

$$= E[Score_{1i} | iPad_i = 0] - E[Score_{0i} | iPad_i = 0]$$

$$= E[Score_{1i} | iPad_i = 0] - E[Score_{0i} | iPad_i = 0]$$

$$+ E[Score_{1i} | iPad_i = 1] - E[Score_{1i} | iPad_i = 1]$$

$$= E[Score_{1i} | iPad_i = 1] - E[Score_{0i} | iPad_i = 0]$$

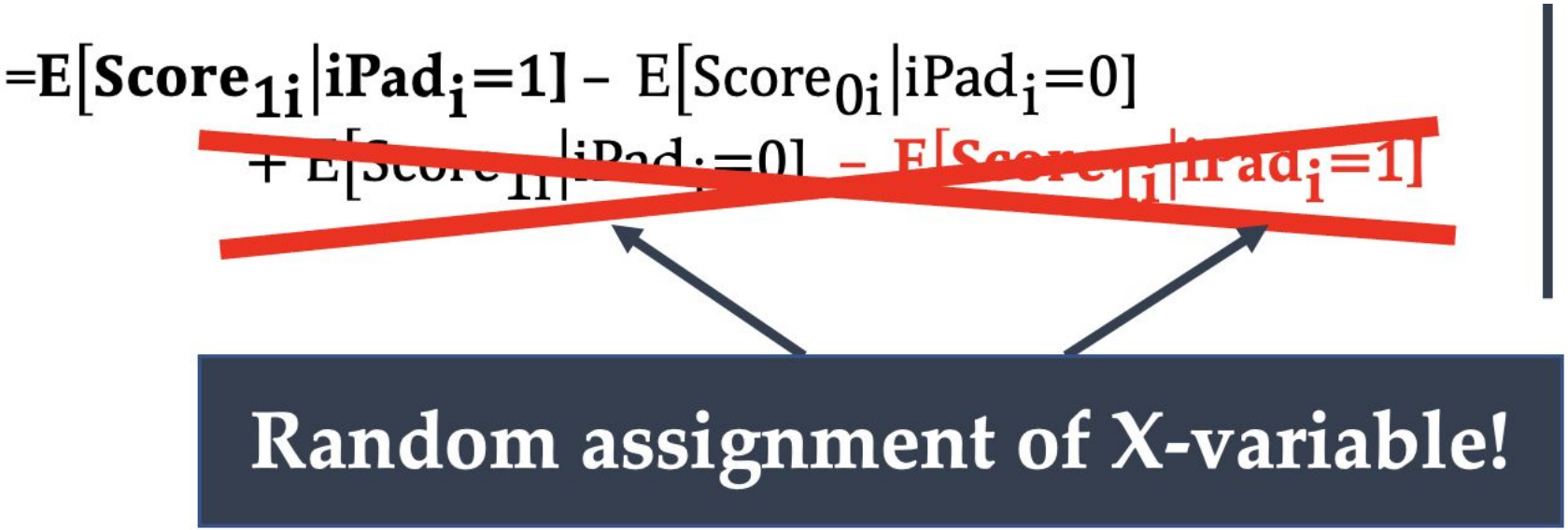
$$+ E[Score_{1i} | iPad_i = 0] - E[Score_{1i} | iPad_i = 1]$$

Treatment effect =

Observed difference +

Selection Bias

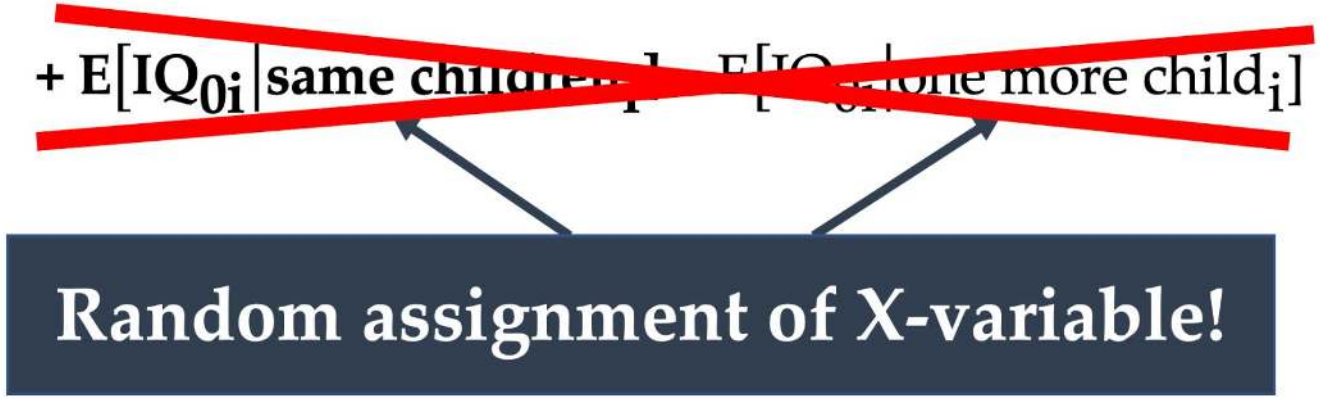
How RCTs solve selection bias



How RCTs solve selection bias

$$E[IQ_{1i} - IQ_{0i} | \text{one more child}_i]$$

$$= E[IQ_{1i} | \text{one more child}_i] - \mathbf{E[IQ_{0i} | same\ children_i]}$$



Epic RCTs of History

On the 20th of May 1747, **I selected twelve patients in the scurvy**, on board the Salisbury at sea. **Their cases were as similar as I could have them**. They all in general had putrid gums, the spots and lassitude, with weakness of the knees. They lay together in one place, being a proper apartment for the sick in the fore-hold; and had one diet common to all, viz. water gruel sweetened with sugar in the morning; fresh mutton-broth often times for dinner; at other times light puddings, boiled biscuit with sugar, etc., and for supper, barley and raisins, rice and currants, sago and wine or the like.

Two were ordered each a quart of cyder a day. **Two others** took twenty-five drops of elixir vitriol three times a day … **Two others** took two spoonfuls of vinegar three times a day … **Two** of the worst patients were put on a course of sea-water … **Two others had each two oranges and one lemon given them every day** … **The two remaining patients,** took … an electary recommended by a hospital surgeon …

The consequence was, **that the most sudden and visible good effects were perceived from the use of oranges and lemons**; one of those who had taken them, being at the end of six days fit for duty … The other was the best recovered of any in his condition; and … was appointed to attend the rest of the sick. Next to the oranges, I thought the cyderhad the best effects …"

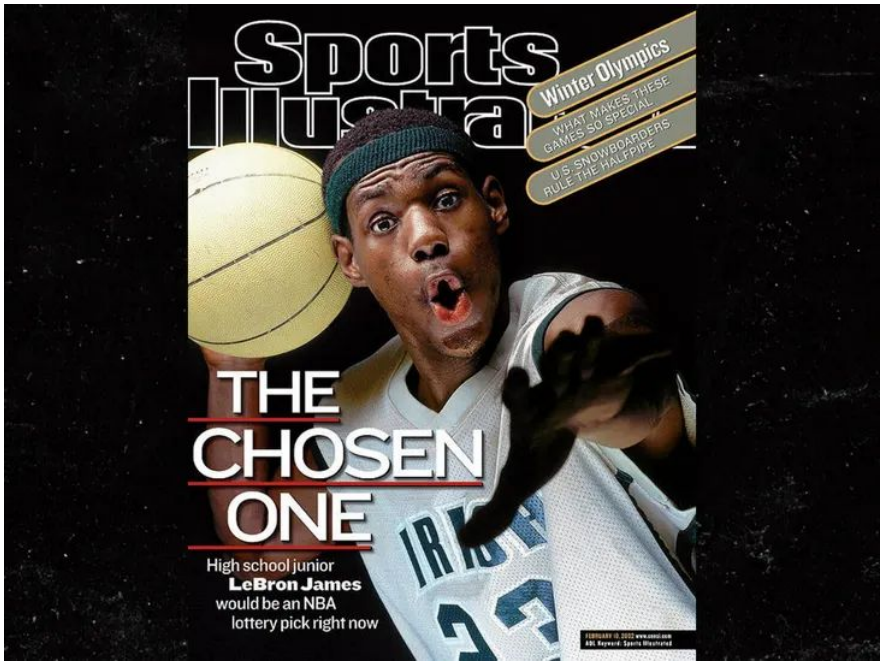
–Dr.James Lind "Treatise on Scurvy" (1753)

So... Do RCTs solve everything?

- Recall: We **NEVER** observe an observation unit at both status at the same time
- Weird things can happen:
  - Hawthorne effect (being aware of being in the study)
  - Pygmalion effect (self-fulfilling prophecies)
  - John Henry effect (control group works harder than treatment)
  - Spillover effects (e.g., from treatment to control)
- It all comes down to: Is your “control group” a **VALID COUNTERFACTUAL** for the “treatment group”

Now... What is a regression?

The name comes from: **regression to the mean**



What we mean by it

A specific **model of the world**

relating two (or more) variables

(in a **linear** fashion - for linear regressions)

$$\mathbf{y_i = a + bx_i + e_i}$$

Model of the world

Are models right?

Are models useful?

Linearity

Which relationships in nature are linear?
Is linearity useful?

We can be more flexible:

$$y_i = a + bx_i^2 + e_i$$
$$\log(y_i) = a + bx_i + e_i$$
$$y_i = a + b_1x_{1i} + b_2x_{2i} + b_3(x_{1i} * x_{2i}) + e_i$$

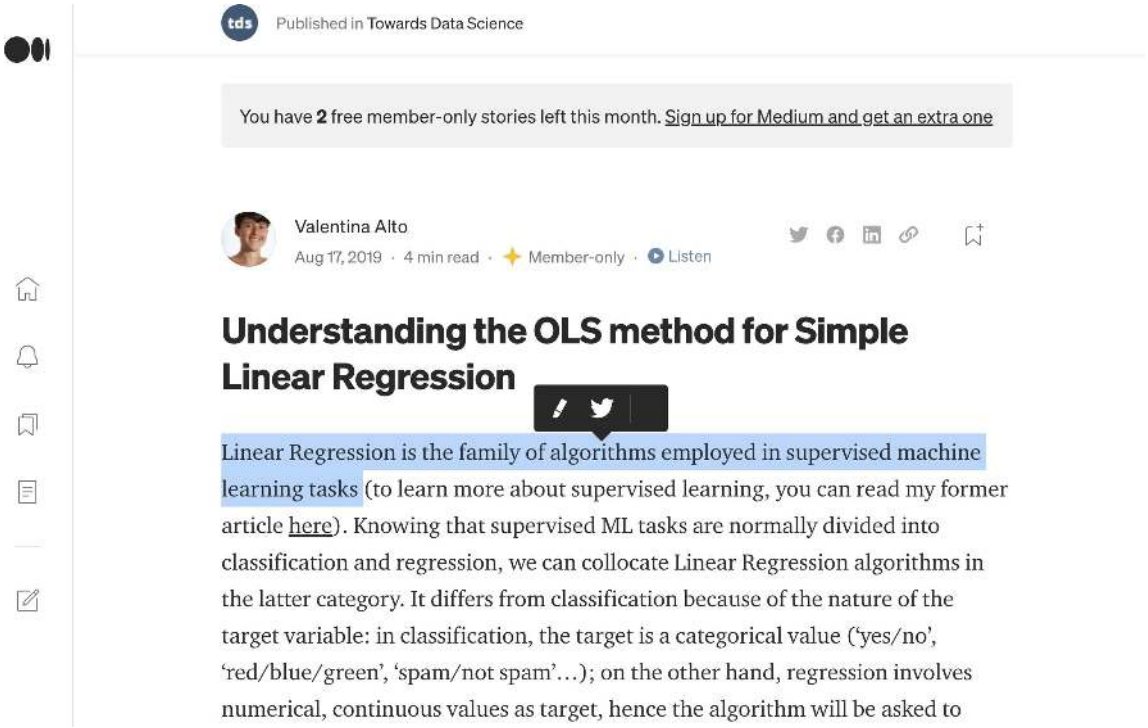
But linearity is a **strong** assumption:
https://www.nature.com/articles/431525a

Other assumptions

Does the model have to be complete?

Well... it depends if we want to interpret it causally

Maths? Estimating regressions with OLS



Published in Towards Data Science

You have **2** free member-only stories left this month. Sign up for Medium and get an extra one

Valentina Alto

Aug 17, 2019 · 4 min read · Member-only · Listen

Understanding the OLS method for Simple Linear Regression

Linear Regression is the family of algorithms employed in supervised machine learning tasks (to learn more about supervised learning, you can read my former article here). Knowing that supervised ML tasks are normally divided into classification and regression, we can collocate Linear Regression algorithms in the latter category. It differs from classification because of the nature of the target variable: in classification, the target is a categorical value ('yes/no', 'red/blue/green', 'spam/not spam'…); on the other hand, regression involves numerical, continuous values as target, hence the algorithm will be asked to

Maths? Estimating regressions with OLS

(Blackboard)

Section 3 - Full

September 14, 2022

0.0.1 Section 3: Let's code

Today, we will look at some coding in R. You will be learning * how to look at and summarize your data * how to run basic regressions * how to interpret regression outputs * how to interpret some of the most important techniques when running regressions * how to visualize data and regression results (more on that next time!)

We will be using a (fictional) dataset on wages in California. We are interested whether there is a "gender wage gap", and how education and age affect one's earnings.

By now, you should know how to read in data. I will not be going over this. Instead, I will just show a piece of code that generates the dataset which we will be using. This can be useful for various things.

You do not need to understand everything in this section.

[1]:
```
set.seed(123456789)
dataset = data.frame(seq(1, 3000))
dataset$n = dataset$seq.1..3000.
dataset$seq.1..3000.=NULL
dataset$male = 0
dataset$male[dataset$n>=1613]=1
dataset$n = NULL

# Individual effect
dataset$individual = rnorm(3000,0,1)
# Gender effect (related to education)
dataset$rgender = rbinom(3000, 1, 0.1)*dataset$male


sampleDist = function(n) {
    sample(x =␣
 ↪c(8,9,10,11,12,13,14,15,16,9,10,11,12,13,14,10,11,15,12,17,17,10,12,11,13,15,15,14,␣
 ↪12, 12, 16, 11, 10, 10, 7, 10, 12, 12, 15, 15, 8, 9, 10, 12), n, replace = T)
    }

# Make education variable
dataset$education = sampleDist(3000)
```

```
dataset$education =␣
 ↪round(dataset$education+dataset$rgender*4*runif(3000,1,2)+rnorm(3000,0,0.
 ↪1),digits=0)
dataset$education[dataset$education<=4] = 7
dataset$education[dataset$education==5] = 7
dataset$education[dataset$education==6] = 8

# Played videogames in childhood
dataset$videogames_childhood = round(runif(3000,0,8),digits=1)

# Education squared
dataset$educationsquared = (dataset$education)*(dataset$education)

# Wages
dataset$wage_monthly = 2000 + 100*dataset$individual + 50*dataset$male +␣
 ↪100*(dataset$education)

dataset$meduc = dataset$male*dataset$education
# Hourly wages
dataset$wage_hourly = 10 + 2*dataset$individual + 2*dataset$male + 1.
 ↪5*(dataset$education) - 1*dataset$meduc
# Age
dataset$age = round(runif(3000,25,70),digits=0)
dataset$agesquared = dataset$age*dataset$age
# Yearly wages
dataset$wage_yearly = 15000 + 1000*dataset$education + 1000*dataset$age -␣
 ↪9*dataset$agesquared + 1000*dataset$individual

dataset$individual = NULL
dataset$rgender = NULL
dataset$n = NULL
dataset$agesquared = NULL
dataset$educationsquared=NULL
dataset$meduc=NULL

dataset$male[dataset$male==0]="Female"
dataset$male[dataset$male==1]="Male"
print("Dataset successfully generated")
```

```
[1] "Dataset successfully generated"
```

Let us have a look at the data:

[2]:
```
colnames(dataset)
head(dataset)
```

1. 'male' 2. 'education' 3. 'videogames_childhood' 4. 'wage_monthly' 5. 'wage_hourly' 6. 'age' 7. 'wage_yearly'

A data.frame: 6 × 7

	male <chr>	education <dbl>	videogames_childhood <dbl>	wage_monthly <dbl>	wage_hourly <dbl>	age <dbl>
1	Female	15	6.2	3550.487	33.50974	67
2	Female	12	7.5	3239.588	28.79175	35
3	Female	15	1.4	3641.554	35.33108	51
4	Female	9	5.2	2827.768	22.05535	41
5	Female	12	5.8	3138.164	26.76329	39
6	Female	12	0.1	3043.738	24.87476	27

A couple of warm-up quiz questions for you: * How many observations are there in the dataset? * How many variables are there in the dataset? * What proportion of the data is male? * What is the average monthly wage in the data?

[3]:
```
dim(dataset)

# Male-female

# mean(data$male)
# Ouch! Need to convert data - this is character, not numeric

# Alternative 1: Recode values using dplyr and recode function
library(dplyr)
dataset$male = as.numeric(recode(dataset$male, "Male" = 1, "Female"=0))

# Alternative 2: Manually
# dataset$male[dataset$male=="Male"]=1
# dataset$male[dataset$male=="Female"]=0
# dataset$male = as.numeric(dataset$male)


mean(dataset$male)

mean(dataset$wage_monthly)

# Let us format these in a nicer way
print(round(mean(dataset$male),2))
print(round(mean(dataset$wage_monthly),10))
```

1. 3000 2. 7

```
Attaching package: ‘dplyr’


The following objects are masked from ‘package:stats’:

    filter, lag
```

```
The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

0.462666666666667

3256.69066908736

```
[1] 0.46
[1] 3256.691
```

Let us summarize the education data. What are mean and standard deviation? How does the data look like? Is there anything you notice?

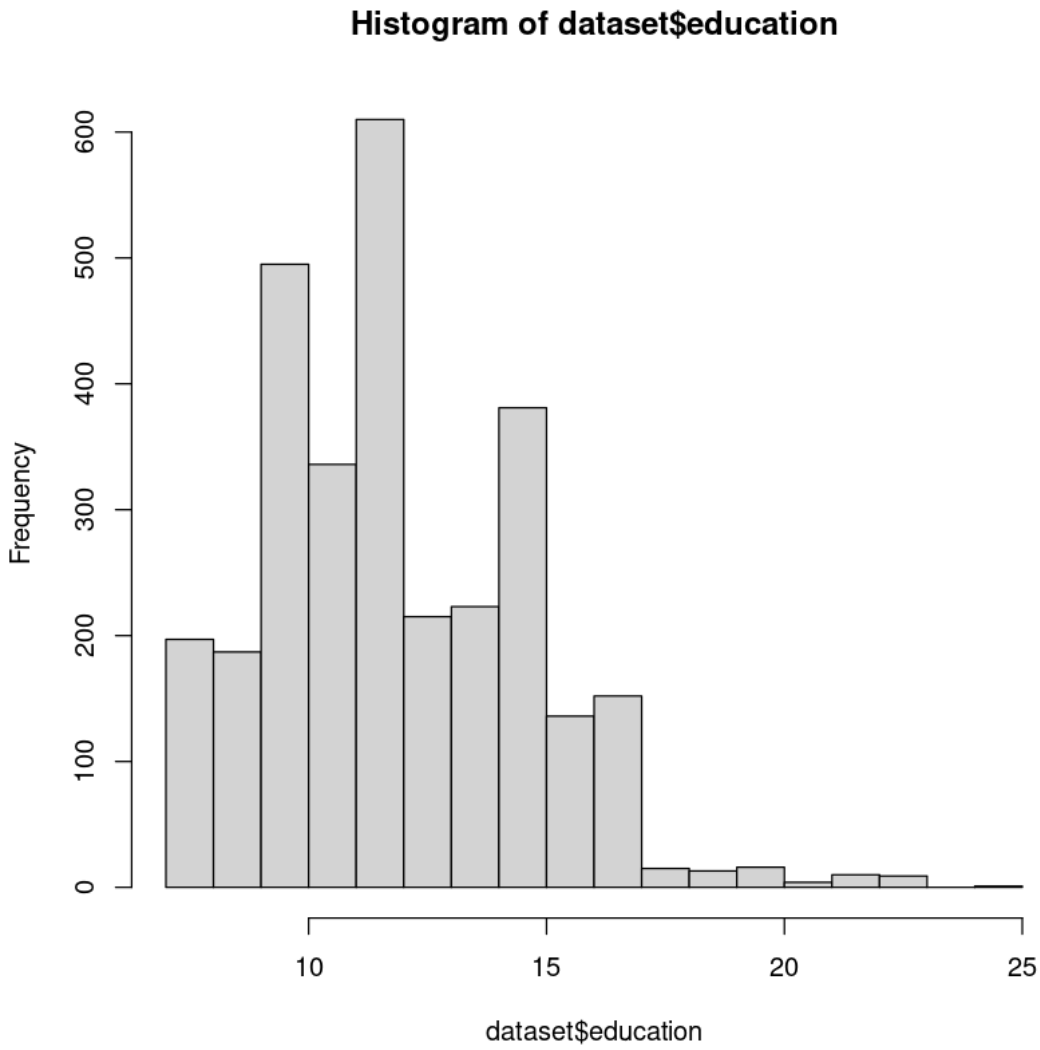
You can do the same thing for monthly wages. Are there any outliers? How would you treat them?

[4]:
```
mean(dataset$education)
summary(dataset$education)
hist(dataset$education)

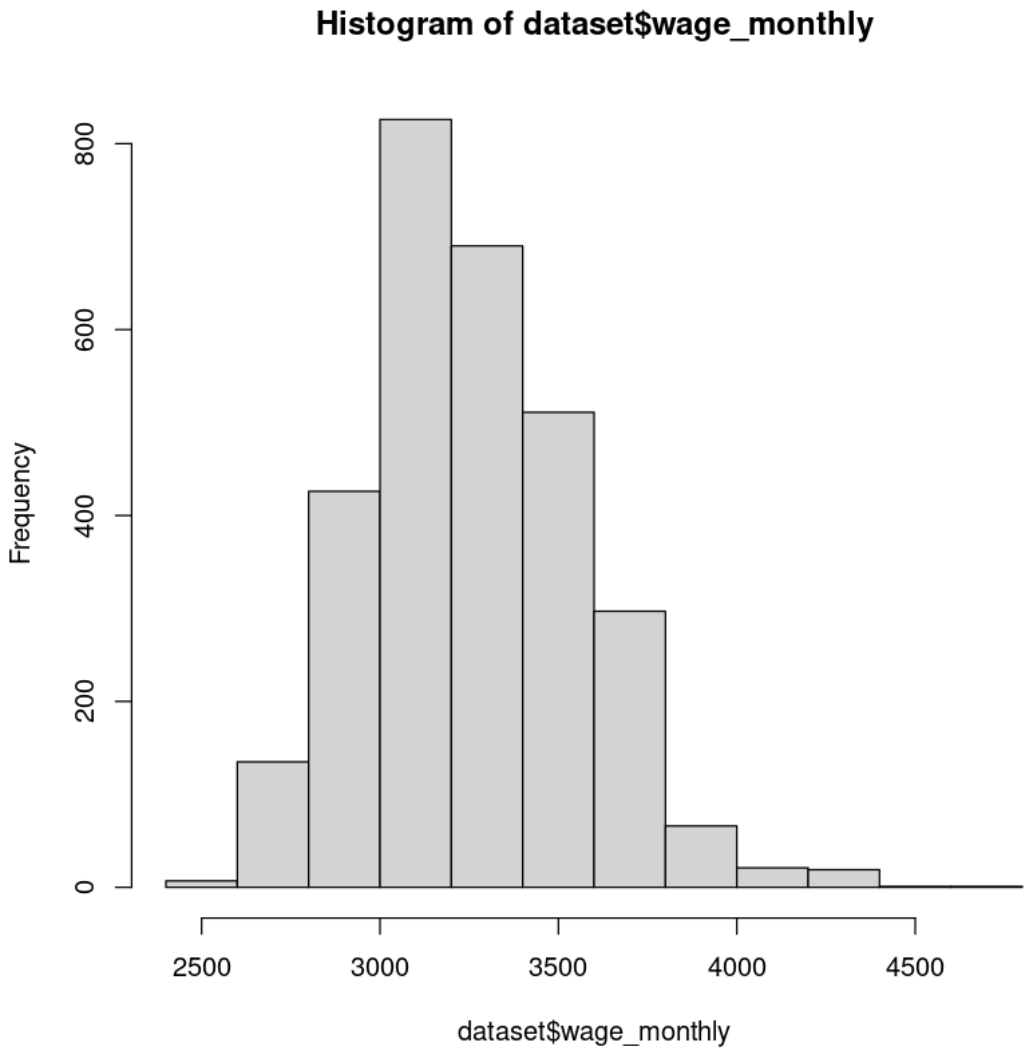
hist(dataset$wage_monthly)
summary(dataset$wage_monthly)
```

12.306

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   7.00   10.00   12.00   12.31   14.00   25.00
```




```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
2445    3046    3225    3257    3460    4704
```



We don't see any outliers in the wage data. We could either drop outliers from our analysis, or we could recode them as a high non-outlier value.

0.1 Digging more into the data: Is there a gender wage gap?

We start with a basic linear model of wages (w_i), which we regress on gender (m_i):

$$w_i = \alpha + \beta_m \cdot \text{male}_i + \epsilon_i$$

How would you run this in R?

[5]:
```
gender_regression = summary(lm(wage_monthly ~ male , data=dataset))
print(gender_regression)
```

```
Call:
lm(formula = wage_monthly ~ male, data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-755.30 -209.43  -26.06  201.42 1381.96

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3200.333      7.262  440.68   <2e-16 ***
male         121.810     10.677   11.41   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 291.6 on 2998 degrees of freedom
Multiple R-squared:  0.04161,   Adjusted R-squared:  0.04129
F-statistic: 130.2 on 1 and 2998 DF,  p-value: < 2.2e-16
```

Let us interpret this regression output. * **What is the meaning of the intercept?** Answer: The intercept is always the 'conditional mean': The expected value of the outcome variable, conditional on all other variables being zero. Here, it is simply the average wage for females (male=0). * **What is the meaning of the coefficient on "male"?** This shows the difference in average wages between males and females. * **What do the other parts of the output mean?** Interpretation of standard error (measure of precision of estimate: The smaller the standard error, the more precise the estimate), and t-statistic (coefficient divided by standard error). A t-statistic above 2 is equivalent to a p-value below 0.05. In this case, we could also reject the null hypothesis that males earn 100$ more than females - 121 (coefficient) minus 2*10 (standard error) gives a bit more than 100.

- **Can you interpret the coefficient as the "gender wage gap"?** Maybe. This is an "unconditional wage gap" - but we cannot interpret this as a causal or discrimination effect.

Below we see that just taking means exactly corresponds to the regression output - but we can do so much more with regressions.

[6]:
```
mean(dataset$wage_monthly[dataset$male==0])
mean(dataset$wage_monthly[dataset$male==1])

mean(dataset$wage_monthly[dataset$male==1]) -␣
  ↪mean(dataset$wage_monthly[dataset$male==0])
```

3200.333429824

3322.14302477363

121.809594949637

Let us consider whether videogames are determinants of wages. Let's include this in the regression as well.

[7]:
```
# Include videogames
example_regression = summary(lm(wage_monthly ~ education + male +␣
 ↪videogames_childhood , data=dataset))
print(example_regression)
```

```
Call:
lm(formula = wage_monthly ~ education + male + videogames_childhood,
    data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-361.40  -67.91   -0.56   70.07  364.62

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          2001.1694     9.1823  217.94   <2e-16 ***
education              99.8057     0.6789  147.02   <2e-16 ***
male                   54.8529     3.7536   14.61   <2e-16 ***
videogames_childhood    0.4836     0.8064    0.60    0.549
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 101.8 on 2996 degrees of freedom
Multiple R-squared:  0.8834,    Adjusted R-squared:  0.8833
F-statistic:  7564 on 3 and 2996 DF,  p-value: < 2.2e-16
```

How would you interpret this regression? How do you interpret the constant here - and what about the coefficients on male and videogames? Answer: The coefficient on videogame is not statistically different from zero. The intercept now is the expected earnings for individuals with zero years of education, who are female, and played zero hours of videogames in their childhood.

0.1.1 Omitted variable bias

Since we also have education in our dataset, let us use it. Do you think education might be an "omitted variable"?

How does our estimate of the "gender wage gap" change if we include education as an additional variable in the regression?

[8]:
```
full_regression = summary(lm(wage_monthly ~ education + male , data=dataset))
print(full_regression)
```

```
Call:
lm(formula = wage_monthly ~ education + male, data = dataset)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-360.68  -68.17   -0.25   70.47  365.97

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2003.2132     8.5255  234.97   <2e-16 ***
education     99.7961     0.6786  147.06   <2e-16 ***
male          54.8694     3.7531   14.62   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 101.7 on 2997 degrees of freedom
Multiple R-squared:  0.8834,    Adjusted R-squared:  0.8833
F-statistic: 1.135e+04 on 2 and 2997 DF,  p-value: < 2.2e-16
```

0.1.2 The OVB formula

We saw above that our estimate of the gender wage gap has become smaller once we control for education.

To formally investigate how the estimate for one coefficient changes if we include more variables in the regression, we do a small experiment.

We will be running three regressions: * 1: Full regression: Include gender and education * 2: Short (biased) regression: Incldue gender only * 3: Auxiliary regression: Regress education on gender

Can you find a relation between the coefficients?

[9]:
```r
full_regression = summary(lm(wage_monthly ~ education + male , data=dataset))
print(full_regression)

true_coefficient = full_regression$coefficients[3,1]
omitted_coefficient = full_regression$coefficients[2,1]

short_regression = summary(lm(wage_monthly ~ male, data=dataset))
print(short_regression)
biased_coefficient = short_regression$coefficients[2,1]

auxiliary_regression =summary(lm(education ~ male, data=dataset))
print(auxiliary_regression)
auxiliary_coefficient = auxiliary_regression$coefficients[2,1]

print(biased_coefficient)
print(true_coefficient + auxiliary_coefficient*omitted_coefficient)
```

```
Call:
lm(formula = wage_monthly ~ education + male, data = dataset)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-360.68  -68.17   -0.25   70.47  365.97

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2003.2132     8.5255  234.97   <2e-16 ***
education     99.7961     0.6786  147.06   <2e-16 ***
male          54.8694     3.7531   14.62   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 101.7 on 2997 degrees of freedom
Multiple R-squared:  0.8834,    Adjusted R-squared:  0.8833
F-statistic: 1.135e+04 on 2 and 2997 DF,  p-value: < 2.2e-16


Call:
lm(formula = wage_monthly ~ male, data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-755.30 -209.43  -26.06  201.42 1381.96

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3200.333      7.262  440.68   <2e-16 ***
male         121.810     10.677   11.41   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 291.6 on 2998 degrees of freedom
Multiple R-squared:  0.04161,   Adjusted R-squared:  0.04129
F-statistic: 130.2 on 1 and 2998 DF,  p-value: < 2.2e-16


Call:
lm(formula = education ~ male, data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-5.6664 -1.9957  0.0043  2.0043 12.3336

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.9957     0.0682  175.89  < 2e-16 ***
male          0.6708     0.1003    6.69 2.65e-11 ***
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.738 on 2998 degrees of freedom
Multiple R-squared:  0.01471,   Adjusted R-squared:  0.01438
F-statistic: 44.76 on 1 and 2998 DF,  p-value: 2.651e-11

[1] 121.8096
[1] 121.8096
```

0.1.3 The Omitted Variable Bias Formula

From these regressions, we see that there is the following relation between the coefficients:

$$\rho_s = \rho + \gamma \cdot \delta$$

where ρ is the "long" coefficient, γ is the coefficient of the omitted variable in the long regression, δ is the coefficient in the auxiliary regression, and ρ_s is the "short" regression coefficient.

This relation is ALWAYS true - no matter whether we consider causality or now.

We can think about this to think about the direction of bias - in a crude way. In our case, γ is positive, and δ is positive as well - so the coefficient from the "short" regression is larger than the coefficient from the "long" regression - as we saw in the regression output!

Let us take a different example. We want to know whether a (non-random) support program for low-income students in a local high school increased their test scores. We regress test scores on the support program and find a negative coefficient. Does this mean that the program had a negative effect on grades?

No! We can use the OVB formula to think about the most important omitted variable in this case: Parents' education (and income, etc.). The higher parents' income, the lower the probability that a child is eligible for the support program ($\delta < 0$). At the same time, the higher parents' education, the better childrens' test scores tend to be ($\gamma > 0$). Therefore, we expect a negative bias: The coefficient when just running test scores on the program is smaller than if we include parents' education! **Neat!**

0.2 Interpreting Interaction Terms (NEXT WEEK)

Maybe the effect of education is different between male and female respondents. To investigate this, we can use a technique called "Interaction Terms". Let us first create a variable that is the product of male and education.

[10]:
```
dataset$male_times_education = dataset$male*dataset$education
```

We will be using hourly wages instead of monthly wages for this analysis. Our regression model looks like this.

$$w_i = \alpha + \beta_m \cdot \text{male}_i + \beta_e \cdot \text{education}_i + \beta_{me} \cdot \text{male}_i \cdot \text{education}_i + \epsilon_i$$

How would we interpret the coefficients?

- Interpret β_m
- Interpret β_e
- Interpret β_{me}

Let us run the regression and have a look at the results:

[11]:
```
interact_regression = summary(lm(wage_hourly ~ education + male +␣
 ↪male_times_education , data=dataset))
print(interact_regression)

# Sidenote: There is a "more correct" way to do it:
interact_regression = summary(lm(wage_hourly ~ education + male +␣
 ↪I(male*education) , data=dataset))
print(interact_regression)
```

```
Call:
lm(formula = wage_hourly ~ education + male + male_times_education,
    data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-7.2015 -1.3734 -0.0039  1.4061  7.2704

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           9.91772    0.25084  39.538  < 2e-16 ***
education             1.50814    0.02048  73.641  < 2e-16 ***
male                  2.36515    0.34440   6.867 7.91e-12 ***
male_times_education -1.02179    0.02735 -37.362  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.035 on 2996 degrees of freedom
Multiple R-squared:  0.8831,    Adjusted R-squared:  0.8829
F-statistic:  7541 on 3 and 2996 DF,  p-value: < 2.2e-16


Call:
lm(formula = wage_hourly ~ education + male + I(male * education),
    data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-7.2015 -1.3734 -0.0039  1.4061  7.2704

Coefficients:
```

```
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           9.91772    0.25084  39.538  < 2e-16 ***
education             1.50814    0.02048  73.641  < 2e-16 ***
male                  2.36515    0.34440   6.867 7.91e-12 ***
I(male * education) -1.02179    0.02735 -37.362  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

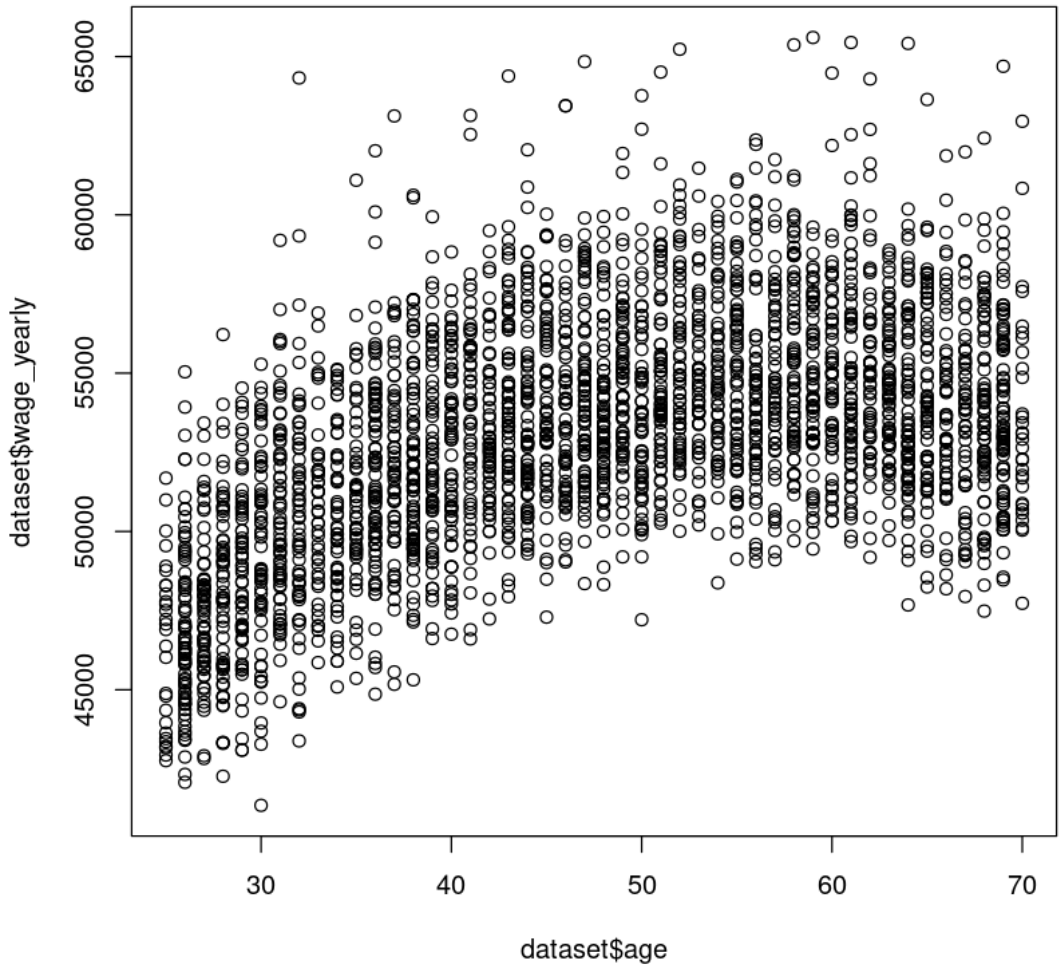
Residual standard error: 2.035 on 2996 degrees of freedom
Multiple R-squared:  0.8831,    Adjusted R-squared:  0.8829
F-statistic:  7541 on 3 and 2996 DF,  p-value: < 2.2e-16
```

0.3 Wage and age: Exploring a life-cycle relationship

We may also be interested in the relation between age and wages (for this part of the analysis, we use yearly wages). What do you think this relationship may look like? Will wages increase in age or become smaller as people become older?

Let us make a plot!

[12]:
```
plot(dataset$age, dataset$wage_yearly)
```



How does this data look like? Do you think a simple linear model would fully capture the relationship between age age wages? What would you do to solve this issue?

Let us generate a new variable that is the square of the age. This looks like a parabola, so maybe we will be able to fit the data with a quadratic model.

[13]:
```
dataset$agesquared = dataset$age*dataset$age
```

We now have a regression model that looks like this:

$$w_i = \alpha + \beta_1 \cdot \text{age}_i + \beta_2 \cdot \text{age}_i^2 + \epsilon_i$$

How do we expect w_i to change if age goes up by one year? This is just so awesome. The effect of age on wage can now be different for different values of age. To see this, we can just take

a derivative:

$$\frac{d\text{wage}_i}{d\text{age}_i} = \beta_1 + 2 * \beta_2 \cdot \text{age}_i$$

The derivative changes with age!

Let us run the regression with the squared term:

[14]:
```
quadratic_regression = summary(lm(wage_yearly ~ education + age + agesquared ,␣
 ↪data=dataset))
print(quadratic_regression)

# Sidenote: There is a "more correct" way to do it:
quadratic_regression = summary(lm(wage_yearly ~ education + age + I(age^2) ,␣
 ↪data=dataset))
print(quadratic_regression)
```

```
Call:
lm(formula = wage_yearly ~ education + age + agesquared, data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-3670.7  -676.3    -8.2   704.4  3680.2

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 14853.5572   279.4781   53.15   <2e-16 ***
education     999.1820     6.7372  148.31   <2e-16 ***
age          1005.3263    11.6659   86.18   <2e-16 ***
agesquared     -9.0277     0.1216  -74.27   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1017 on 2996 degrees of freedom
Multiple R-squared:  0.9252,    Adjusted R-squared:  0.9251
F-statistic: 1.236e+04 on 3 and 2996 DF,  p-value: < 2.2e-16


Call:
lm(formula = wage_yearly ~ education + age + I(age^2), data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-3670.7  -676.3    -8.2   704.4  3680.2

Coefficients:
```

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 14853.5572   279.4781   53.15   <2e-16 ***
education     999.1820     6.7372  148.31   <2e-16 ***
age          1005.3263    11.6659   86.18   <2e-16 ***
I(age^2)       -9.0277     0.1216  -74.27   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1017 on 2996 degrees of freedom
Multiple R-squared:  0.9252,    Adjusted R-squared:  0.9251
F-statistic: 1.236e+04 on 3 and 2996 DF,  p-value: < 2.2e-16
```

We can now also calculate the “tipping point” by setting the derivative from above to zero.

$$\frac{d\text{wage}_i}{d\text{age}_i} = 0 \Leftrightarrow 0 = \beta_1 + 2\beta_2 \cdot \text{age}_i \Leftrightarrow \text{age}_i = -\frac{\beta_1}{2\beta_2}$$

Plugging this in, we get $1005/(2 * 9) \approx 56$ years as the tipping point.

0.4 Summary: What we learned

Today, we: * Saw how to get some very basic information (like means, standard deviations, etc.) from a dataset * learned how to generate new variables * learned how to run linear regressions * learned how to interpret the output of linear regressions * looked at what happens if we forget to include an important variable in our regression: **omitted variable bias**! * Learned two key techniques which allow us to analyze much more rich economic patterns in the world: Interaction effects (**NEXT WEEK**) and quadratic terms.

[ ]:

Section 4: Getting data and some additional regression magic

This week, we are going to learn:

- How to download data from two very popular sources in the field of political economy and macroeconomics
- Look at an interesting question related to the effectiveness of international aid.
- Learn about interaction terms and bad controls
- Maybe: Make a quick example about data visualization

In [1]:
```
install.packages(c('WDI','owidR','plotly','showtext','huxtable','jtools'))

library('WDI')
library('owidR')
library('ggplot2')
library('dplyr')
library('plotly')
library('huxtable')
library('tidyr')
library('showtext')
library('jtools')
library('plotly')
print('Installed required packages')

# We set a seed to make sure we get the results if we run the code again, in case we gen
set.seed("1234")
```

```
Installing packages into ‘/opt/r’
(as ‘lib’ is unspecified)

also installing the dependency ‘RJSONIO’



Attaching package: ‘dplyr’


The following objects are masked from ‘package:stats’:

    filter, lag


The following objects are masked from ‘package:base’:

    intersect, setdiff, setequal, union



Attaching package: ‘plotly’


The following object is masked from ‘package:ggplot2’:

    last_plot


The following object is masked from ‘package:stats’:

    filter


The following object is masked from ‘package:graphics’:

    layout



Attaching package: ‘huxtable’


The following object is masked from ‘package:dplyr’:

    add_rownames


The following object is masked from ‘package:ggplot2’:

    theme_grey


Loading required package: sysfonts

Loading required package: showtextdb
```

```
[1] "Installed required packages"
```

We first download some data from the World Development Indicators (https://data.worldbank.org/). We can track down the data online and then download it using the official codes provided by the World Bank. We

can filter by country and years, have a look at the data, etc.

In [2]:
```
dataset = WDI(indicator=c(oda_per_cap='DT.ODA.ODAT.PC.ZS',
                          perc_below_pov_line='SI.POV.DDAY',
                          gdp_per_cap_ppp='NY.GDP.PCAP.PP.CD',
                          gdp_per_cap='NY.GDP.PCAP.CD',
                        inflation='FP.CPI.TOTL.ZG'),
              country="all",
              start=1960,
              end=2022,
              extra=TRUE)

# Sort by country-year
dataset = dataset[order(dataset$country, dataset$year),]

colnames(dataset)
head(dataset)
print("Done")
```

'iso2c' · 'country' · 'year' · 'status' · 'lastupdated' · 'oda_per_cap' · 'perc_below_pov_line' · 'gdp_per_cap_ppp' · 'gdp_per_cap' · 'inflation' · 'iso3c' · 'region' · 'capital' · 'longitude' · 'latitude' · 'income' · 'lending'

A data.frame: 6 × 17

	iso2c	country	year	status	lastupdated	oda_per_cap	perc_below_pov_line	gdp_per_cap_ppp	gdp_
	<chr>	<chr>	<int>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	
437	AF	Afghanistan	1960		2022-09-16	1.909532	NA	NA	5
446	AF	Afghanistan	1961		2022-09-16	3.781052	NA	NA	5
447	AF	Afghanistan	1962		2022-09-16	1.810416	NA	NA	5
448	AF	Afghanistan	1963		2022-09-16	3.842526	NA	NA	7
449	AF	Afghanistan	1964		2022-09-16	4.737925	NA	NA	8
450	AF	Afghanistan	1965		2022-09-16	5.416661	NA	NA	10

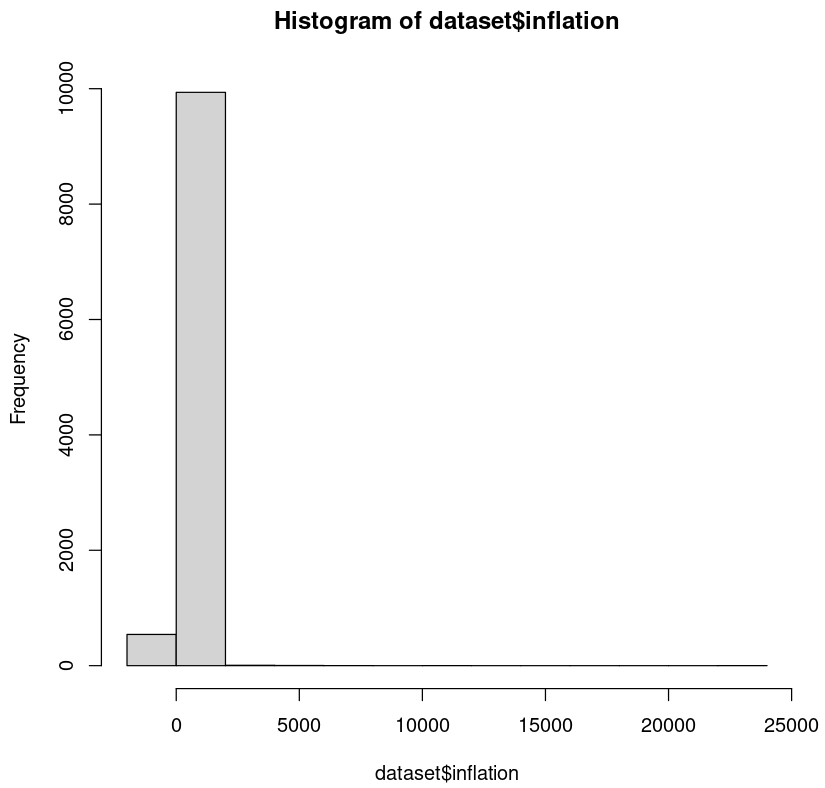
```
[1] "Done"
```

In [3]:
```
# We can, like last week, summarize the data and make histograms to look at its distribu
summary(dataset$gdp_per_cap_ppp)

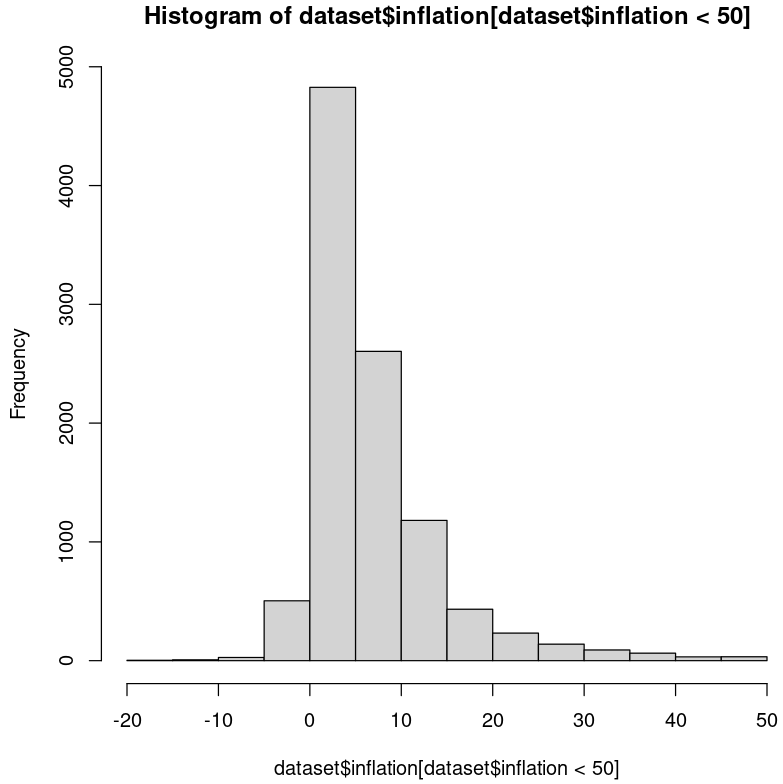
hist(dataset$inflation)
# This one looks weird because of outliers, let's only look at cases with inflation belo
hist(dataset$inflation[dataset$inflation<50])

summary(dataset$perc_below_pov_line)
```

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
   285.3   2947.8   7692.0  14597.4  19250.6 153563.9     9023
```




```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
 0.0     0.5     3.6    13.7    19.7    91.5   14170
```



Some of the data have missing values in some years. Today, we will solve this pragmatically by just using the previous years' values if an observation is missing. We have to do this *grouped by country*.

In [4]:
```
dataset = dataset %>%
  group_by(iso3c) %>%
      fill(perc_below_pov_line, .direction = "down")

dataset = dataset %>%
```

```
  group_by(iso3c) %>%
  fill(oda_per_cap, .direction = "down")

dataset = dataset %>%
  group_by(iso3c) %>%
  fill(gdp_per_cap_ppp, .direction = "down")
```

We are interested in whether the effect of international aid (ODA - official development assistance) is different in countries with democracies vs. in non-democratic countries. For this we need data on demcracy. We will also get data from schooling for some later analysis. We can get this from https://ourworldindata.org/.

We can first search for terms using the search engine in the package. Next, we can read in the data, rename it, and filter by years (and countries, if we want to).

In [5]:
```
owid_search("democracy")

democracy <- owid("democracy-polity",
                  rename = "democracy") %>%
  filter(year %in% 1960:2018)


owid_search("schooling")
yos <- owid("mean-years-of-schooling-long-run",
                  rename = "years_of_schooling") %>%
  filter(year %in% 1960:2018)
```

A matrix: 54 × 2 of type chr

titles	chart_id
Child mortality rate vs. liberal democracy	child-mortality-vs-liberal-democracy
Age of democracy	age-of-democracy-bmr
Age of democracy	age-of-democracy-womsuffr-bmr
Age of democracy	age-of-democracy-polity
Age of electoral democracy	age-of-electoral-democracy
Age of electoral democracy	age-of-electoral-democracy-lied
Age of liberal democracy	age-of-liberal-democracy-row
Citizen satisfaction with democracy	citizen-satisfaction-with-democracy
Citizen support for democracy	citizen-support-for-democracy
Deliberative democracy	deliberative-democracy-vdem
Deliberative democracy weighted by population	deliberative-democracy-popw-vdem
Democracy	democracy-polity
Democracy	democracy-eiu
Democracy weighted by population	democracy-popw-polity
Democracy weighted by population	popw-democracy-eiu
Distribution of deliberative democracy	distribution-deliberative-democracy-vdem
Distribution of deliberative democracy weighted by population	distribution-deliberative-democracy-popw-vdem
Distribution of democracy	distribution-democracy-eiu
Distribution of democracy	distribution-democracy-polity
Distribution of democracy weighted by population	distribution-democracy-popw-polity
Distribution of democracy weighted by population	distribution-democracy-popw-eiu
Distribution of egalitarian democracy	distribution-egalitarian-democracy-vdem
Distribution of egalitarian democracy weighted by population	distribution-egalitarian-democracy-popw-vdem
Distribution of electoral democracy	distribution-electoral-democracy-vdem
Distribution of electoral democracy weighted by population	distribution-electoral-democracy-popw-vdem
Distribution of liberal democracy	distribution-liberal-democracy-vdem
Distribution of liberal democracy weighted by population	distribution-liberal-democracy-popw-vdem
Distribution of participatory democracy	distribution-participatory-democracy-vdem
Distribution of participatory democracy weighted by population	distribution-participatory-democracy-popw-vdem
Egalitarian democracy	egalitarian-democracy-vdem
Egalitarian democracy weighted by population	egalitarian-democracy-popw-vdem
Electoral democracy	electoral-democracy
Electoral democracy	electoral-democracy-lied
Electoral democracy	electoral-democracy-row
Electoral democracy weighted by population	electoral-democracy-popw-vdem
Experience with democracy	experience-with-democracy-bmr
Experience with democracy	experience-with-democracy-womsuffr-bmr
Experience with democracy	experience-with-democracy-polity

titles	chart_id
Experience with electoral democracy	experience-with-electoral-democracy
Experience with electoral democracy	experience-with-electoral-democracy-lied
Experience with liberal democracy	experience-with-liberal-democracy-row
GDP per capita vs. liberal democracy	gdp-per-capita-vs-liberal-democracy
Government effectiveness vs. liberal democracy	govt-effectiveness-vs-liberal-democracy
Human rights protection vs. liberal democracy	human-rights-protection-vs-liberal-democracy
Liberal democracy	liberal-democracy
Liberal democracy	liberal-democracy-row
Liberal democracy today vs. past average years of schooling	liberal-democracy-today-vs-past-schooling
Liberal democracy weighted by population	liberal-democracy-popw-vdem
Life expectancy vs. liberal democracy	life-expectancy-vs-liberal-democracy
Participatory democracy	participatory-democracy-vdem
Participatory democracy weighted by population	participatory-democracy-popw-vdem
Taxation vs. liberal democracy	taxation-vs-liberal-democracy
Varieties of democracy	varieties-democracy-vdem
Varieties of democracy weighted by population	varieties-democracy-popw-vdem

A matrix: 14 × 2 of type chr

titles	chart_id
Child mortality vs. women's average years of schooling	correlation-between-child-mortality-and-mean-years-of-schooling-for-those-aged-15-and-older
Liberal democracy today vs. past average years of schooling	liberal-democracy-today-vs-past-schooling
Average years of schooling vs. GDP per capita	average-years-of-schooling-vs-gdp-per-capita
Average years of schooling vs. expected years of schooling	average-schooling-vs-expected-schooling
Expected years of schooling vs. GDP per capita	expected-years-of-schooling-vs-gdp-per-capita
Gender ratios for average years of schooling	gender-ratios-for-mean-years-of-schooling
Fertility rate vs. average years of schooling	fertility-rate-vs-mean-years-of-schooling
Average years of schooling	mean-years-of-schooling-long-run
Average years of schooling for men	mean-years-of-schooling-male
Average years of schooling for women	mean-years-of-schooling-female
Correlation between Internet usage and average years of schooling	correlation-between-internet-usage-and-mean-years-of-schooling
Expected years of schooling vs. Share in extreme poverty	expected-years-of-schooling-vs-share-in-extreme-poverty
Primary schooling: Adjusted net attendance rate	primary-schooling-adjusted-net-attendance-rate
Productivity vs. share of tertiary schooling	productivity-vs-share-of-tertiary-schooling

The next thing we want to do is to combine these two datasets. We can do this by a process called **merging** the data. We need to have at least one column of data that have the same identifier (for example, survey ID number, country name, year, etc.). We merge with the command `join` and we can do a `right_join` in this case because we want complete observations in the democracy data.

In [6]:
```
democracy$country = democracy$entity
democracy$iso3c = democracy$code

yos$country = yos$entity
yos$iso3c = yos$code


dataset_panel <- dataset %>%
  right_join(democracy)

dataset_panel <- dataset_panel %>%
  left_join(yos)



dataset_panel = dataset_panel[order(dataset_panel$country, dataset_panel$year),]
head(dataset_panel)
```

```
Joining, by = c("country", "year", "iso3c")
Joining, by = c("country", "year", "iso3c", "entity", "code")
```

A gr

iso2c	country	year	status	lastupdated	oda_per_cap	perc_below_pov_line	gdp_per_cap_ppp	gdp_per_c
<chr>	<chr>	<int>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dl
AF	Afghanistan	1960		2022-09-16	1.909532	NA	NA	59.773
AF	Afghanistan	1961		2022-09-16	3.781052	NA	NA	59.860
AF	Afghanistan	1962		2022-09-16	1.810416	NA	NA	58.458
AF	Afghanistan	1963		2022-09-16	3.842526	NA	NA	78.706
AF	Afghanistan	1964		2022-09-16	4.737925	NA	NA	82.095
AF	Afghanistan	1965		2022-09-16	5.416661	NA	NA	101.108

Next, we have to do some data cleaning and generate some additional variables. I will largely skip over this. (Note: The main problem is that we now have a *panel* (countryXyear)(dataset, which we haven't talked about in class yet. I will convert this into a *cross-sectional* dataset, which only has one observation (year) per unit (country) of analysis. We will talk about the exciting world of panel datasets later in class!)

In [10]:
```
head(dataset_panel)

# Normalize democracy scores: All will lie between 0 and 1
dataset_panel$democracy = (dataset_panel$democracy+10)/20

# Generate cross-sectional data
dataset_cs = dataset_panel %>%
  filter(year==2016)

# Make year averages

data_pov90 = dataset_panel %>%
  filter(year==1990) %>%
  group_by(iso3c) %>%
  dplyr::summarize(pov_1990 = mean(perc_below_pov_line, na.rm=TRUE))
```

```r
data_aid9010 = dataset_panel %>%
  filter(year %in% 1990:2010) %>%
  group_by(iso3c) %>%
  dplyr::summarize(aid_9010 = mean(oda_per_cap, na.rm=TRUE))

# Calculate in per year amounts.
data_aid9010$aid_9010 = data_aid9010$aid_9010*20

data_dem9010 = dataset_panel %>%
  filter(year %in% 1990:2010) %>%
  group_by(iso3c) %>%
  dplyr::summarize(dem9010 = mean(democracy, na.rm=TRUE))


data_gdp1990= dataset_panel %>%
  filter(year %in% 1990) %>%
  group_by(iso3c) %>%
  dplyr::summarize(gdp1990 = mean(gdp_per_cap, na.rm=TRUE))


# Merge in year-averages
dataset_cs = dataset_cs %>%
  left_join(data_pov90)

dataset_cs = dataset_cs %>%
  left_join(data_aid9010)

dataset_cs = dataset_cs %>%
  left_join(data_dem9010)

dataset_cs = dataset_cs %>%
  left_join(data_gdp1990)


# Generate happiness variable
dataset_cs$happiness = dataset_cs$gdp_per_cap_ppp + rnorm(nrow(dataset_cs),0,1000)


head(dataset_cs)
```

A gro

iso2c	country	year	status	lastupdated	oda_per_cap	perc_below_pov_line	gdp_per_cap_ppp	gdp_per_c
<chr>	<chr>	<int>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<db
AF	Afghanistan	1960		2022-09-16	1.909532	NA	NA	59.773
AF	Afghanistan	1961		2022-09-16	3.781052	NA	NA	59.860
AF	Afghanistan	1962		2022-09-16	1.810416	NA	NA	58.458
AF	Afghanistan	1963		2022-09-16	3.842526	NA	NA	78.706
AF	Afghanistan	1964		2022-09-16	4.737925	NA	NA	82.095
AF	Afghanistan	1965		2022-09-16	5.416661	NA	NA	101.108

```
Joining, by = "iso3c"
Joining, by = "iso3c"
Joining, by = "iso3c"
Joining, by = "iso3c"
```

iso2c	country	year	status	lastupdated	oda_per_cap	perc_below_pov_line	gdp_per_cap_ppp	gdp_per_c
<chr>	<chr>	<int>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dl
AF	Afghanistan	2016		2022-09-16	115.00937350	NA	1981.118	512.01
NA	Africa	2016	NA	NA	NA	NA	NA	
AL	Albania	2016		2022-09-16	59.46939738	0.1	12078.843	4124.05
DZ	Algeria	2016		2022-09-16	3.56461208	0.5	11624.341	3946.45
AO	Angola	2016		2022-09-16	6.97824841	14.6	7103.226	1728.02
AR	Argentina	2016		2022-09-16	0.06010502	0.7	20307.870	12790.24

Our final datast includes the following variables of interest for our analysis:

- `gdp_per_cap_ppp` : GDP per capita in the year 2011, in PPP and real 2017 USD
- `perc_below_poverty_line` : Share of the population below the poverty line (2.15 USD per day), in 2011.
- `pov_1990` : Share of the population below the poverty line (2.15 USD per day), in 1990
- `aid_9010` : Total development aid received (in USD), per capita, over the years 1990-2010
- `dem_9010` : Average democracy level between 1990 and 2010 (between 0 and 1)
- `happiness` : A fictional happiness score, in 2011

Our hypothesis is that development aid has an effect on the welfare of a country, measured by GDP per capita and the poverty rate. Maybe aid takes many years to have an effect, so we take the sum over 1990 to 2010 and see how this affected our outcomes in 2011. We are also wondering whether democracy was related to welfare, and so we also average the democracy score between 1990 and 2010 to keep everything comparable.

Analysis: Analyzing interaction terms

Last week, I gave a hint that we will talk about interaction terms. Interaction terms can help us to understand questions such as:

- Does having a child have a **different effect** on wages for men and women?
- Do children from different socio-economic backgrounds **respond differently** to free school meals?
- Does development aid **work better** in democratic as opposed to non-democratic countries?

We quickly analyze the math of this model on the blackboard.

Let's run the regression!

In [11]:
```
nointeraction_pov <- lm(perc_below_pov_line ~ pov_1990 +  aid_9010 + dem9010 , data=data

interaction_pov <- lm(perc_below_pov_line ~  pov_1990 +  aid_9010 + dem9010 + I(aid_9010

nointeraction_gdp <- lm(gdp_per_cap_ppp ~ pov_1990 +  aid_9010 + dem9010 , data=dataset_

interaction_gdp <- lm(gdp_per_cap_ppp ~  pov_1990 +  aid_9010 + dem9010 + I(aid_9010*dem
```

```
export_summs(nointeraction_pov, interaction_pov, nointeraction_gdp, interaction_gdp)
```

A huxtable: 14 × 5

	names	Model 1	Model 2	Model 3	Model 4
	<chr>	<chr>	<chr>	<chr>	<chr>
		Model 1	Model 2	Model 3	Model 4
1	(Intercept)	-61.2374795071691	-128.785951684016	-100473.050623957	743.840399071715
2		(131.151998543101)	(184.46296744571)	(50308.6071500479)	(65918.4780877931)
3	pov_1990	0.465746047021812 ***	0.478817136761546 ***	-170.119670411957 **	-189.705828663991 ***
4		(0.125929692234316)	(0.129811946126957)	(48.3053821940856)	(46.388746992385)
5	aid_9010	0.00143573381067744	0.0889241284377262	1.78207531734834	-129.313462736596 *
6		(0.00237476791877688)	(0.165961798407396)	(0.910937443770791)	(59.3070214754539)
7	dem9010	111.590525363933	236.236210579241	217161.635844753 *	30388.3702126576
8		(243.232174121096)	(341.274800056792)	(93301.451979696)	(121955.727704969)
9	I(aid_9010 * dem9010)		-0.162697591926318		243.791515944305 *
10			(0.308598125471292)		(110.278605258806)
1.1	N	35	35	35	35
2.1	R2	0.310724982601861	0.317052621256712	0.485566742578144	0.557630737582579
.1	*** p < 0.001; ** p < 0.01; * p < 0.05.	*** p < 0.001; ** p < 0.01; * p < 0.05.	*** p < 0.001; ** p < 0.01; * p < 0.05.	*** p < 0.001; ** p < 0.01; * p < 0.05.	*** p < 0.001; ** p < 0.01; * p < 0.05.

Bad controls

When we talked about omitted variables, we always talked about how forgetting an important variable can have bad consequences for your analysis. But there is also the opposite problem: Bad controls! The bad controls problem arises when we control for a variable in a regression that is actually an outcome (a variable on the left-hand side).

For example:

- A treatment gives people money at time t, and we want to study how it changes their consumption at time t+1. We control for their wage at time t+1.
- We want to study whether winning the lottery at time t makes you happier at time t+1. We control for whether you bought a Ferrari at time t+1.

Let's make an example. We know that happiness is strongly related to income (at least, to some extent: https://ourworldindata.org/happiness-and-life-satisfaction). Aid in the past will affect income today and happiness today. Therefore, controlling for happiness today introduces a bad control problem!

In [14]:
```
basic_regression <- lm( gdp_per_cap_ppp ~  gdp1990 , data=dataset_cs)

bad_controls <- lm( gdp_per_cap_ppp ~  gdp1990  + happiness, data=dataset_cs)

export_summs(basic_regression, bad_controls)
```

A huxtable: 10 × 3

	names	Model 1	Model 2
	<chr>	<chr>	<chr>
		Model 1	Model 2
1	(Intercept)	7978.42432394203 ***	-47.7777015498555
2		(1239.0877614923)	(138.307999063505)
3	gdp1990	2.04469323680988 ***	0.00333770097061518
4		(0.11501970977029)	(0.021294596424702)
5	happiness		0.999164381516758 ***
6			(0.00893164138740528)
1.1	N	117	117
2.1	R2	0.733189781839178	0.997591438771217
.1	*** p < 0.001; ** p < 0.01; * p < 0.05.	*** p < 0.001; ** p < 0.01; * p < 0.05.	*** p < 0.001; ** p < 0.01; * p < 0.05.

We can interpret bad controls through the OVB formula we learnt last time:

$$\rho_s = \rho + \gamma \cdot \delta$$

where ρ is the "long" coefficient, γ is the coefficient of the omitted variable in the long regression, δ is the coefficient in the auxiliary regression (regress omitted variable on included variable), and ρ_s is the "short" regression coefficient.

Remember? It told you that this formula **ALWAYS** holds - no matter what variable I include.

Extreme case: Include GDP$\times$10 on the RHS - what will happen?

Do you need an OVB revision?

In [15]:
```
# Revise OVB
a = lm(log(gdp_per_cap_ppp) ~  democracy  , data=dataset_cs)
b = lm(log(gdp_per_cap_ppp) ~ years_of_schooling +  democracy  , data=dataset_cs)
c = lm(years_of_schooling  ~ democracy  , data=dataset_cs)
export_summs(a,b,c)
```

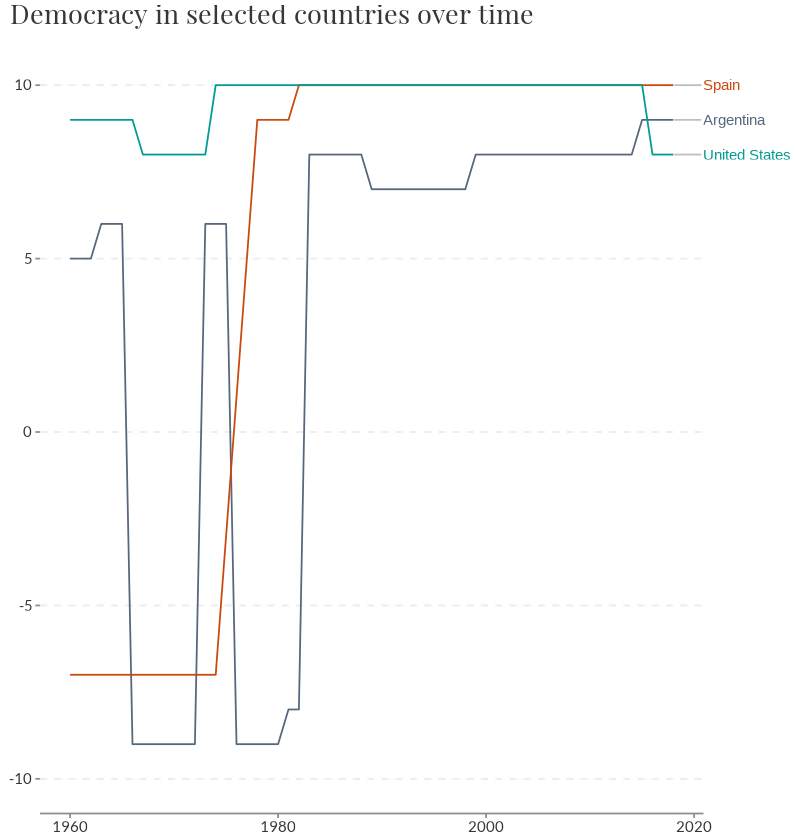
A huxtable: 10 × 4

	names	Model 1	Model 2	Model 3
		<chr>	<chr>	<chr>
		Model 1	Model 2	Model 3
1	(Intercept)	1.09009171786149	8.96440800542877 ***	-25.4438428120717 **
2		(3.5845626524535)	(1.94872769203634)	(8.62215636672121)
3	democracy	15.2527368966298 *	-4.28199504589645	63.1104742397465 ***
4		(6.67806650043998)	(3.69542563257363)	(16.0966995706918)
5	years_of_schooling		0.313394531780118 ***	
6			(0.0166755807608285)	
1.1	N	145	144	162
2.1	R2	0.035196299293719	0.725080867999958	0.0876534587488713
.1	*** p < 0.001; ** p < 0.01; * p < 0.05.	*** p < 0.001; ** p < 0.01; * p < 0.05.	*** p < 0.001; ** p < 0.01; * p < 0.05.	*** p < 0.001; ** p < 0.01; * p < 0.05.

Data visualizations

An important skill - this is just a quick example and we'll do more in the next weeks!

In [16]:
```
a = c("Argentina","Spain", "United States")
owid_plot(democracy, filter = a, summarise=FALSE) +
  labs(title = paste0("Democracy in selected countries over time")) +
  scale_y_continuous(limits = c(-10, 10))
```

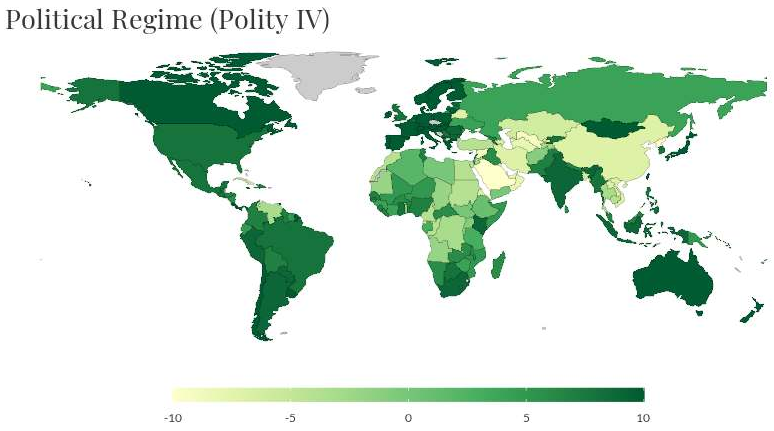


In [17]:
```
# Make a map

democracy2 <- owid("democracy-polity",
                   rename = "democracy") %>%
```

```
  filter(year==2018)

owid_map(democracy2, palette = "YlGn") +
  labs(title = "Political Regime (Polity IV)")
```



In [ ]:

Interpreting regressions + Exam practice

https://pollev.com/jonathanold608

Roadmap - Regression, regression, regression

- Discussion PSet 1
- (Recap:) Control variables and bad controls
- How to approach an Econ 140R exam question
- Practice Exam questions

Problem Set 1:
some quick points

PSet 1 - some notes on interpretation

Whether there is a causal relationship or not **does not depend on**:

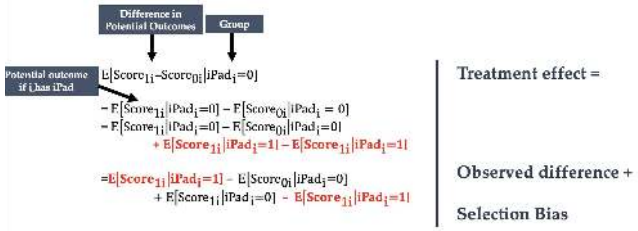
- Outliers
- Heterogeneous treatment effects
- Variance
- The size of a coefficient

But it **does** depend on:

- Selection Bias
- Omitted variable bias
- Reverse causality
- ...

PSet 1 - some notes on interpretation

Recall: Every observed effect in the data is a composite of:



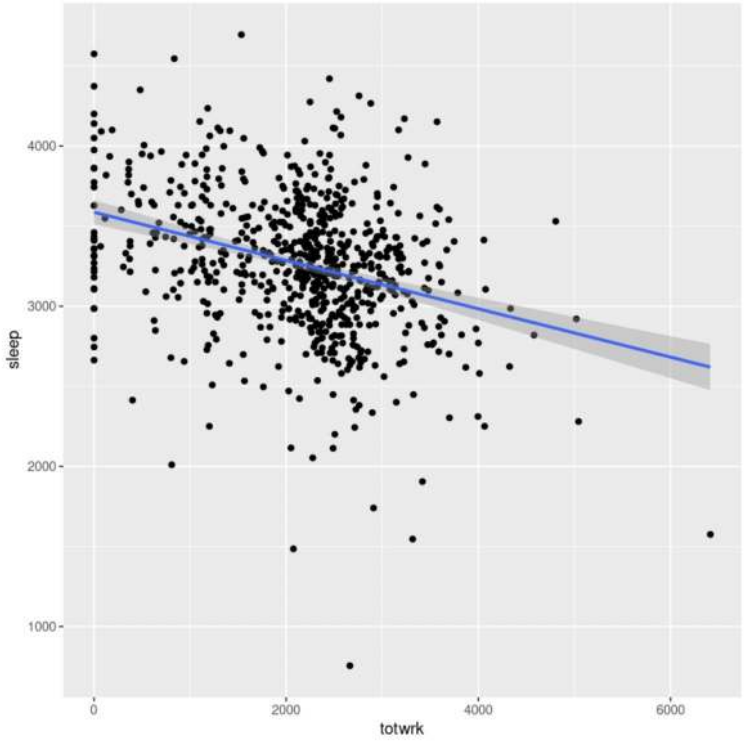
Causal Effect

+

Selection Bias

- We cannot observe either of these - so we can never be sure!
- Econometrics is all about uncertainty. You can state that there are different possibilities and you cannot know for sure!

PSet 1 - let's try together



- What does the data say?
- How do we interpret this?
- What would an RCT do?

Recap and big picture: Control variables

Why do we use control variables?

In decreasing order of importance:

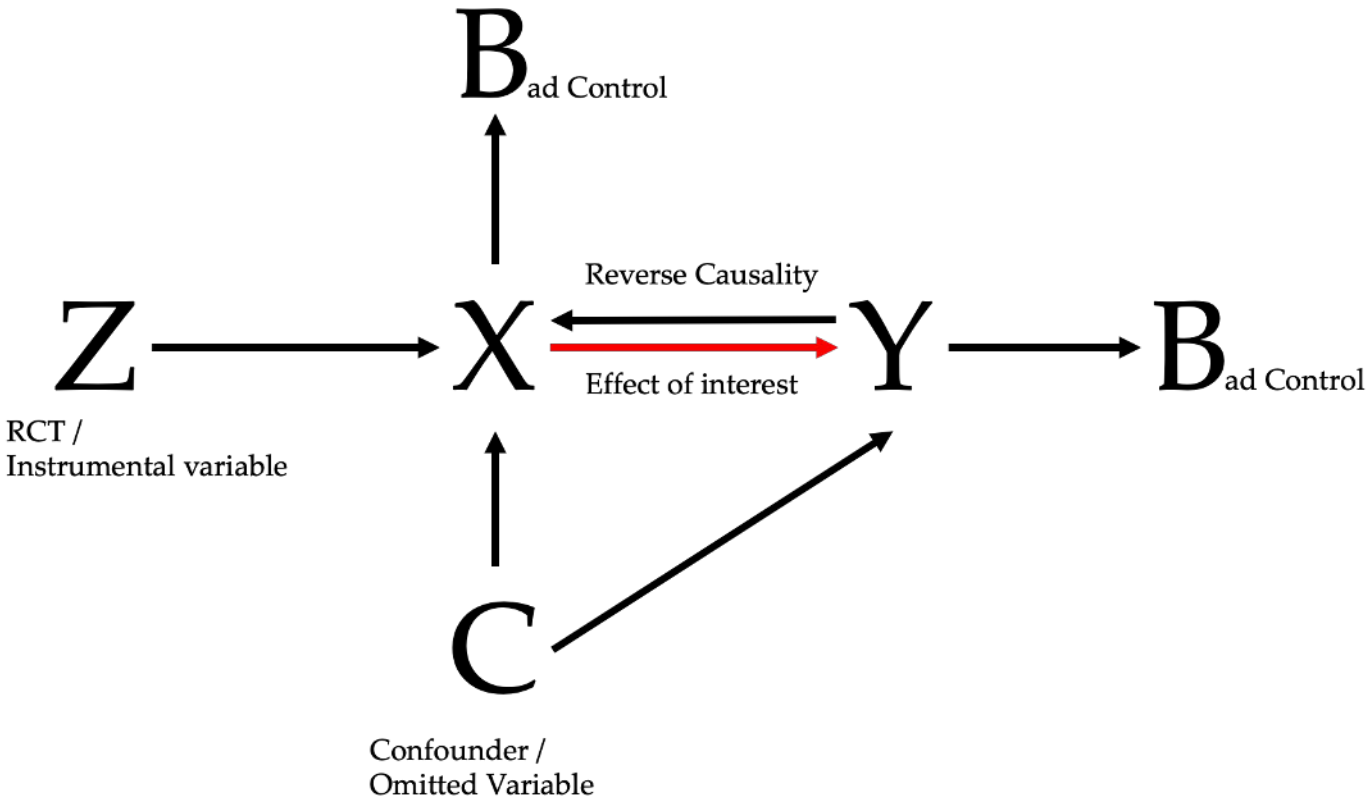
- To remove selection bias / omitted variable bias
- To increase precision of our estimates
- To know about the (conditional/partial) correlation of other variables
- To better predict the outcome

How control variables work

- See here: https://nickchk.com/causalgraphs.html

Are more controls always good?

Are more controls always good?



Are more controls always good?

No, because:

- Some controls are **BAD controls**
- Some controls are unrelated to the outcome of interest - such variables are **unnecessary** and increase the risk of spurious correlation and overfitting

So what are “good” controls?

- Predetermined (before the “treatment” of interest)
- Immutable characteristics (not changed by treatment)

Interpreting regressions: A guide

[See separate slides]

Exam practice

A systematic way to approach an exam question

1. **Think**: About the question, about the real world
2. Start with the numbers **you see**
3. **Interpret** the numbers:
   a. Direction (positive or negative?)
   b. Magnitude (big or small?)
   c. Significance (significant or insignificant, at what level?)
4. Establish whether estimated relationship is **causal or not**
   a. Is X-variable randomized or as good as random? Do we have valid counterfactuals?
   b. If not: Do you expect **bias**? Of which sort (**OVB, reverse causality**, ...)?
   c. Find a plausible story around bias (using the OVB formula)
5. How would you find a causal effect?
   a. Can you run an RCT?

There are no traps!

Practice Exam: Question 1 (RCT, coefficient interpretation)

You are analyzing the results from a carefully constructed randomized evaluation of the effect of workplace incentives on worker productivity in a large moving company. You regress the log of worker output (measured in kg moved per hour) on a constant and an indicator for whether the individual was randomly selected to receive a financial incentive for high productivity.

A. **(4 marks)** The coefficient on the indicator variable is 0.20 with a standard error of 0.05. Explain, as if to a non-technical audience, what this estimate implies.

B. **(4 marks)** A colleague notices that because of the physical nature of the work, men are likely to be more productive at this job than women, all else equal. As such, your colleague claims, the worker's gender constitutes an omitted variable. You have data on each employee's gender so you rerun the regression including a gender dummy. How do you expect your estimate for the effect of incentives to change and why?

C. **(4 marks)** Another colleague claims that women may be more or less responsive to workplace financial incentives than men. Describe the regression you would run to test this claim and how you would interpret the results.

Practice Exam: Question 2 (Regression and conceptual)

Consider the following hypothetical example. The Department of Education wants to evaluate the effectiveness of a program from 16- and 17-year olds designed to increase the number of students studying science, technology, engineering and mathematics (STEM) at college and university. Some secondary schools offer the program and others do not.

A. **(10 marks)** You have data from a sample of 25-year olds. The data include each individual's current labour market income and whether or not the secondary school the individual attended offered the STEM program at the time the individual attended. Write down the equation for the bivariate regression using these two variables, with labour market income as the outcome variable. Define the variables precisely. How would you interpret the regression coefficients (describe them as if to a general audience)? Would you expect the slope coefficient to capture the causal effect of explanatory variable on the outcome? Why or why not?

B. **(7.5 marks)** A colleague makes the following claim: "You do not need to worry about establishing causality. The data were gathered using a random sample, so they are not biased." Critically assess this claim.

Practice Exam: Question 2 (Regression controls)

Consider the following hypothetical example. The Department of Education wants to evaluate the effectiveness of a program from 16- and 17-year olds designed to increase the number of students studying science, technology, engineering and mathematics (STEM) at college and university. Some secondary schools offer the program and others do not.

D. **(7.5 marks)** The program was funded by tax revenues raised by local districts. Therefore, you gather additional data and control for average district income in your regression. First, explain as if to a general audience what it means to control for a variable in a regression. Second, describe what you expect would happen to the slope coefficient you described above when you control for income. Clearly state any assumptions you make.

Practice Exam: Question 3 (Regression)

The Ministry of Truth is interested in a rumour that air pollution could impact mental health. One of the most harmful pollutants is fine particulate matter PM2.5, which comes from operations that involve the burning of fuels such as wood, oil, coal, gas, or grass fires. A research team is sent to investigate the rumour. The team randomly selects and surveys 19,920 people across 71 districts of the country. The key variable, $Exposure_i$, is a dummy variable equal to 1 if the individual i is exposed to a large amount of PM2.5 in the last two years, and 0 otherwise. The team also conducts a standardised questionnaire to record depressive symptoms in the last month, called the Kessler Psychological Distress scale (K6). The questionnaire results in a score, $Depression_i$, that ranges from 0 to 24; and the higher the score, the more severe the depressive symptoms for individual i. The variable has a sample average of 2.96. Running regressions with $Depression_i$ as the dependent variable, the analyst Winston Smith obtains the following results:

Practice Exam: Question 3 (Regression)

Dependent variable: $Depression_i$

Regressor	(1)	(2)	(3)	(4)	(5)	(6)
$Exposure_i$	0.834	0.635	0.614	0.598	0.422	0.554
	(0.032)	(0.039)	(0.045)	(0.021)	(0.020)	(0.042)
$Exposure_i \times Female_i$		-0.834	0.065	0.054		
		(0.013)	(0.024)	(0.011)		
$Female_i$			-0.739	-0.732	-0.745	-0.825
			(0.036)	(0.018)	(0.019)	(0.066)
$Manual\ Job_i$				0.324	0.050	
				(0.122)	(0.008)	
Age_i					0.324	0.452
					(0.111)	(0.132)
Age_i^2					0.421	0.524
					(0.122)	(0.121)

Notes. All estimations contain a constant term. Robust standard errors are in the parentheses. $Manual\ Job_i = 1$ if i works in a manual job, and 0 otherwise. $Female_i = 1$ if i is a female, 0 otherwise. Age_i is the age (years old) of individual i, and Age^2 is the square of Age_i.

Practice Exam: Question 3

Dependent variable: $Depression_i$

Regressor	(1)	(2)	(3)	(4)	(5)	(6)
$Exposure_i$	0.834	0.635	0.614	0.598	0.422	0.554
	(0.032)	(0.039)	(0.045)	(0.021)	(0.020)	(0.042)
$Exposure_i \times Female_i$		-0.834	0.065	0.054		
		(0.013)	(0.024)	(0.011)		
$Female_i$			-0.739	-0.732	-0.745	-0.825
			(0.036)	(0.018)	(0.019)	(0.066)
$Manual\ Job_i$				0.324	0.050	
				(0.122)	(0.008)	
Age_i					0.324	0.452
					(0.111)	(0.132)
Age_i^2					0.421	0.524
					(0.122)	(0.121)

Notes. All estimations contain a constant term. Robust standard errors are in the parentheses. $Manual\ Job_i = 1$ if i works in a manual job, and 0 otherwise. $Female_i = 1$ if i is a female, 0 otherwise. Age_i is the age (years old) of individual i, and Age^2 is the square of Age_i.

a) Interpreting the coefficient in Column (1), a journalist, Katherine, claims: "Since participants are randomly selected, we can infer that exposure to a large amount of PM2.5 does cause depression."

i. Explain carefully why Katherine is wrong, specifying the direction of bias(es) if there is any. Which assumption(s) would she need to impose for the causality claim to hold?

[13 marks]

ii. What is the correct interpretation from Column (1) that Katherine should have made?

[2 marks]

Practice Exam: Question 3

Dependent variable: $Depression_i$

Regressor	(1)	(2)	(3)	(4)	(5)	(6)
$Exposure_i$	0.834	0.635	0.614	0.598	0.422	0.554
	(0.032)	(0.039)	(0.045)	(0.021)	(0.020)	(0.042)
$Exposure_i \times Female_i$		-0.834	0.065	0.054		
		(0.013)	(0.024)	(0.011)		
$Female_i$			-0.739	-0.732	-0.745	-0.825
			(0.036)	(0.018)	(0.019)	(0.066)
$Manual\ Job_i$				0.324	0.050	
				(0.122)	(0.008)	
Age_i					0.324	0.452
					(0.111)	(0.132)
Age_i^2					0.421	0.524
					(0.122)	(0.121)

Notes. All estimations contain a constant term. Robust standard errors are in the parentheses. $Manual\ Job_i = 1$ if i works in a manual job, and 0 otherwise. $Female_i = 1$ if i is a female, 0 otherwise. Age_i is the age (years old) of individual i, and Age^2 is the square of Age_i.

c) Another analyst, Julia, suggests that the team should include into the regressions a variable, $Health\ Expenditure_i$, which captures individual i's total expenditures on health-related services in the last two years. Her rationale is that financial distress may cause depression and is related to air pollution in the area where people with financial distress work. Should the team follow her suggestion? Explain why or why not.

[10 marks]

Exam practice

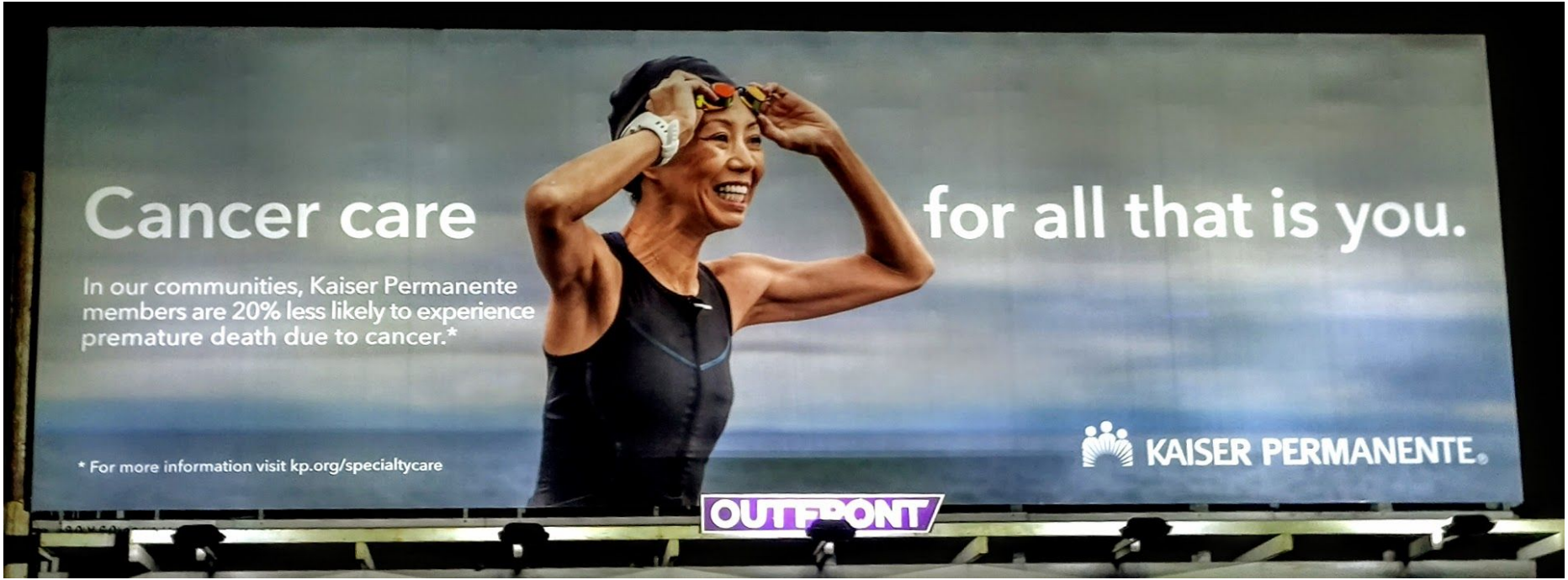
https://pollev.com/jonathanold608

Today is all about Midterm exam practice

- Your questions
- A very detailed mock exam question that talks through:
  - The fundamental problem of causal identification
  - Regression table interpretation
  - Omitted variable bias
  - The potential of RCTs

Mock exam question

Mock exam



Seen this week on the I-80!

Question 1.a: Conceptual



You see the following billboard on
way back from San Francisco. Based on the billboard, a friend of you advises you to switch to Kaiser Permanente insurance, as this will decrease your risk of premature death due to cancer by 20%.

- **Why is your friend likely wrong?**
- **What would have to hold for your friend's statement to be correct?**
- Bonus question: Is there anything else that's weird about the comparison on the billboard?

Question 1a: Answer



- My friend is wrong because a simple comparison of means does not deliver a causal estimate of the impact of the insurance. **People who are members of KP are likely fundamentally different from people that are not, in ways that are correlated with their risk to die of cancer**. They may, for example, be richer, younger, healthier, or more health-conscious. Thus, the two groups are not valid counterfactuals for each other.
- For my friend's statement two hold, KP-members and non-KP-members would have to be valid counterfactuals for each other. **That is, KP-members' expected cancer death risk without KP insurance would have to be the same as that of non-KP-members.** We call this the 'identifying assumption' of your friend's research design.
- Other weird things: The billboard does not specify the comparison group (all people? All people with insurance?). We also don't know whether this is a conditional or unconditional probability.

Question 1b: Regression tables



- You were able to obtain the data that KP used for their analysis, and you run a regression of whether an individual died from cancer, on a dummy variable whether they were members of KP. Interpret the regression output below.
- Test whether the coefficient on KP member is significant.

Outcome: Died from cancer	Model 1
(Intercept)	**0.020** (0.0010)
KP member	**-0.004** (0.0009)
N	10,324

Robust standard errors in parentheses.

Question 1b: Answer

Outcome: Died from cancer	Model 1
(Intercept)	**0.020** (0.0010)
KP member	**-0.004** (0.0009)
N	10,324

- The intercept indicates that, in this sample, 2% of individuals that were not members of KP died from cancer. The intercept is significantly different from zero, as |0.02/0.001| > 2
- The slope coefficient on KP member indicates that individuals that were members of KP were 0.4 percentage points less likely to die from cancer. As -0.004/0.02=-0.2, this means they were 20% less likely to die from cancer.
- The t-statistic for the slope coefficient is -0.004/0.0009=-4.44. This is less than -2, so the difference is statistically significant from zero.

Question 1c: Omitted variable bias



Another friend suggests ‘controlling for income’.

- Explain intuitively what it means to control for a variable in a regression
- Using the omitted variable bias formula, explain what you would expect to happen to the coefficient on KP member.

Outcome: Died from cancer	Model 1	Model 2
(Intercept)	**0.020** (0.0010)	**0.030** (0.0011)
KP member	**-0.002** (0.0009)	
Annual income (in 10,000)		
N	10,324	8,925

Robust standard errors in parentheses.

Question 1c: Answer

Died from cancer	Model 1	Model 2
(Intercept)	**0.020** (0.0010)	**0.030** (0.0011)
KP member	**-0.002** (0.0009)	
Annual income (10,000)		
N	10,324	8,925

- Controlling for a variable removes omitted variable bias caused by the omission of that variable. Controlling for a variable is the econometric equivalent of “keeping a variable constant”. Mathematically, including a control variable removes systematic differences in both the other independent variables and the dependent variable that are due to the included control. This allows us to compare “apples to apples” by looking at the correlation of the independent and dependent variable **for given (fixed) values of the control variable**.
- We expect annual income to be positively correlated to being a KP member. We also expect annual income to be negatively related to the risk of death from cancer. Therefore, we expect the coefficient on KP member in the short regression to be downward biased. If we include annual income, the coefficient will be larger.

Question 1d: Omitted variable bias II



To confirm your suspicions about the OVB, you run an auxiliary regression of Annual income on the KP member dummy. You get a coefficient of 0.7. You then run the full (long) regression, but your computer has a bug. Can you calculate the coefficient on KP member?

Outcome: Died from cancer	Model 1	Model 2
(Intercept)	**0.020** (0.0010)	**0.030** (0.0011)
KP member	**-0.002** (0.0009)	**XXXXXXXXXXX**
Annual income (in 10,000)		**-0.003** (0.001)
N	10,324	8,925

Robust standard errors in parentheses.

Question 1d: Answer



To calculate this, use the OVB formula:

$$\beta_S = \beta + \pi \times \delta \rightarrow \beta = \beta_S - \pi \times \delta$$

In our case β_S=-0.02, π=0.7 and δ=-0.003.

Therefore, β=0.0001.

Outcome: Died from cancer	Model 1	Model 2
(Intercept)	**0.020** (0.0010)	**0.030** (0.0011)
KP member	**-0.002** (0.0009)	**0.0001**
Annual income (in 10,000)		**-0.003** (0.001)
N	10,324	8,925

Robust standard errors in parentheses.

Question 1e: Regression tables



- Interpret the output of the table for model 2
- Looking at the output, your friend claims: The coefficient on KP member is close to zero. Therefore, being a member of the insurance does have no effect on cancer risk. Give **three reasons** why your friend may be wrong.

Outcome: Died from cancer	Model 1	Model 2
(Intercept)	**0.020** (0.0010)	**0.030** (0.0011)
KP member	**-0.002** (0.0009)	**0.0001** (0.050)
Annual income (in 10,000)		**-0.003** (0.001)
N	10,324	8,925

Robust standard errors in parentheses.

Question 1e: Answer

Died from cancer	Model 2
(Intercept)	**0.030** (0.0011)
KP member	**-0.0001** (0.050)
Income (10,000)	**-0.003** (0.001)
N	8,925

- Intercept: For a hypothetical individual with 0 income that is not a member of KP, we expect their risk of dying from cancer to be 3%.
- Coefficient on KP member: For any given level of income, being a KP member is associated with a 0.01 percentage point lower risk of dying from cancer. This is not significantly different from zero.
- Coefficient on income: Holding KP membership constant, each additional $10,000 income is associated with a 0.3 percentage point lower risk of dying from cancer.
- Reasons why friend is wrong:
  - Coefficient is insignificant, but not zero (although it is very small)
  - Hypothesis testing: We fail to reject the null, but we do not accept it!
  - Controlling for one variable does not fully remove OVB and our estimate is still not causal.

Question 1f: RCT



You managed to get a summer internship with Kaiser Permanente, who are sympathetic to your concerns. They have a lot of money and suggest running an RCT in order to find out what their impact on the risk of dying from cancer actually is.

- How would an RCT solve the issues discussed before?
- **Bonus question**: We cannot simply randomly give insurances to people (they have to sign up, etc.). Therefore, your manager asks you to design an RCT that can help you to identify the causal effect of being a KP member. Give at least one idea how to design such an RCT.

Question 1f: Answer



- RCTs help us solve the fundamental problem of causal identification via random assignment of the treatment. Recall that our problem was that the two groups did not have the same expected outcome in the absence of a treatment. With an RCT, the two groups are, in expectation, not different on average, and so the mean difference or a regression just gives us the causal effect. Note: We need a large enough sample for this to work!
- This is a creative question with no right or wrong. The key is to look for a feasible randomized treatment that affects the probability of joining KP, but not other things. Some ideas:
  - Giving people without any insurance randomly vouchers to join KP for free or at a discount
  - Give people who just sign up for insurance (e.g. incoming international students) an incentive (such as a default option)
  - Randomly give people a cash incentive to change insurance - away from KP towards other insurances
  - Note: These three experiments all would give very different estimates and interpretations!

Exam practice

https://pollev.com/jonathanold608

Midterm review

- Your questions
- Control variables: A 3D explanation.
- Logarithmic specifications
- Practice question
- Your questions

Your questions

Logarithmic specifications

Logarithmic specifications

y	= a + b*x	+ e
y	= a + b*log(x)	+ e
log(y)	= a + b*x	+ e
log(y)	= a + b*log(x)	+ e

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

Practice questions

Practice question 1 (RCT, coefficient interpretation)

You are analyzing the results from a carefully constructed randomized evaluation of the effect of workplace incentives on worker productivity in a large moving company. You regress the log of worker output (measured in kg moved per hour) on a constant and an indicator for whether the individual was randomly selected to receive a financial incentive for high productivity.

A. **(4 marks)** The coefficient on the indicator variable is 0.20 with a standard error of 0.05. Explain, as if to a non-technical audience, what this estimate implies.

B. **(4 marks)** A colleague notices that because of the physical nature of the work, men are likely to be more productive at this job than women, all else equal. As such, your colleague claims, the worker's gender constitutes an omitted variable. You have data on each employee's gender so you rerun the regression including a gender dummy. How do you expect your estimate for the effect of incentives to change and why?

Practice question 1 (RCT, coefficient interpretation)

a. Individuals that were randomly selected to receive the financial incentives had a 0.2 higher log worker output. In other words, they had around 20% more output than individuals that were not selected. This difference is statistically significant at the 5% level.

b. Let us look at the OVB formula. Clearly, gender could be related to output, so $\partial > 0$. If the incentive was really random, then gender is uncorrelated to the incentive, so $\pi = 0$. Hence, $\partial\pi = 0$, so there should be no effect of including gender as control.
Omitted variable bias is not a problem in RCTs.

Practice question 2 (OVB, coefficient interpretation)

We are interested to predict whether an individual was insured. We have data on their log wages and on their age. Interpret the coefficient on log wage in the first model. [**NOT EASY**]

	Model 1	Model 2
(Intercept)	1.84	-18.66 ***
	(1.25)	(0.68)
logwage	-0.18	2.57 ***
	(0.17)	(0.09)
age		-0.69 ***
		(0.02)
nobs	300	300
r.squared	0.00	0.86

Practice question 2 - Answer

a. In increase in wages by 1% is associated with an decrease in the probability of being insured by 0.18 percentage points. However, this difference is not significant at the 5% level.

Practice question 2 (OVB, coefficient interpretation)

b. Interpret the coefficient on age in column 2.
c. Using the OVB formula, find the coefficient in the auxiliary regression. If you don't find the coefficient, tell us what sign the coefficient has.

	Model 1	Model 2
(Intercept)	1.84	-18.66 ***
	(1.25)	(0.68)
logwage	-0.18	2.57 ***
	(0.17)	(0.09)
age		-0.69 ***
		(0.02)
nobs	300	300
r.squared	0.00	0.86

Practice question 2 - Answer

b. Individuals with one additional year of age are 69 percentage points less likely to be insured.

c. In the auxiliary regression, we regress the omitted variable on the included variable, so age on log wages.
Use the OVB formula:

$$\beta_S = \beta + \pi \; x \; \delta \rightarrow \delta = (\beta_S - \beta)/\pi$$

In our case β_S=-0.18, π=-0.69 and β=2.57.

Therefore, δ=3.98.

The coefficient on log wages goes up, so we had downward bias. Therefore, $\pi \; x \; \delta$ must be negative. Since π is negative, δ must be positive.

A systematic way to approach an exam question

1. **Think**: About the question, about the real world
2. Start with the numbers **you see**
   a. One-sentence summary
   b. Direction (positive or negative?)
   c. Statistical significance (significant or insignificant, at what level?)
   d. (Economic) magnitude (big or small?)
3. Then: Establish whether estimated relationship is **causal or not**
   a. What do the results **mean**? Correlation (interesting) or causality (policy-relevant)
   b. Is X-variable randomized? Do we have valid counterfactuals?
   c. If not: Do you expect **bias**? Of which sort (**OVB, reverse causality, bad controls**, …)?
   d. Find a plausible story around bias (using the OVB formula)

There are no traps!

Section 8: Instrumental Variables

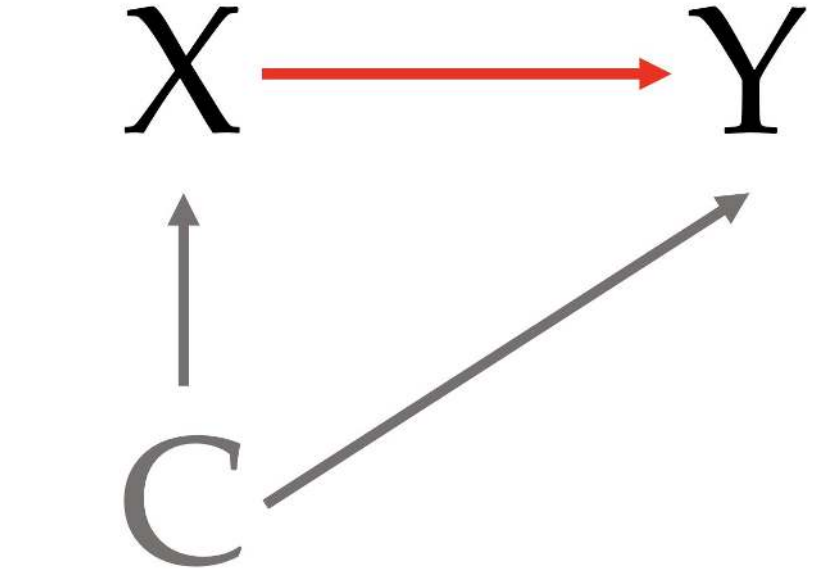
https://pollev.com/jonathanold608

Roadmap for today

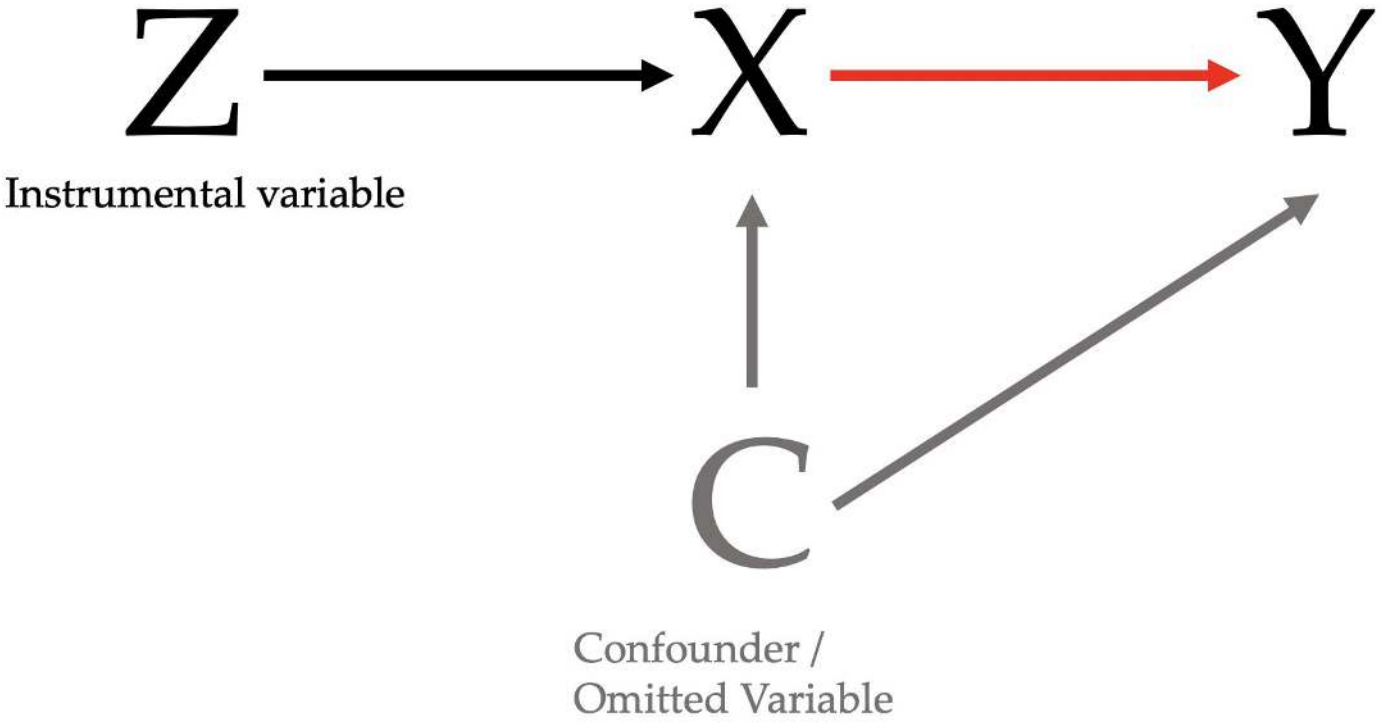
- Instrumental Variables
  - Setup (Causal Diagram)
  - Application: Rainfall, Growth, and Conflict
  - IV, LATE and Wald
- Short-term project

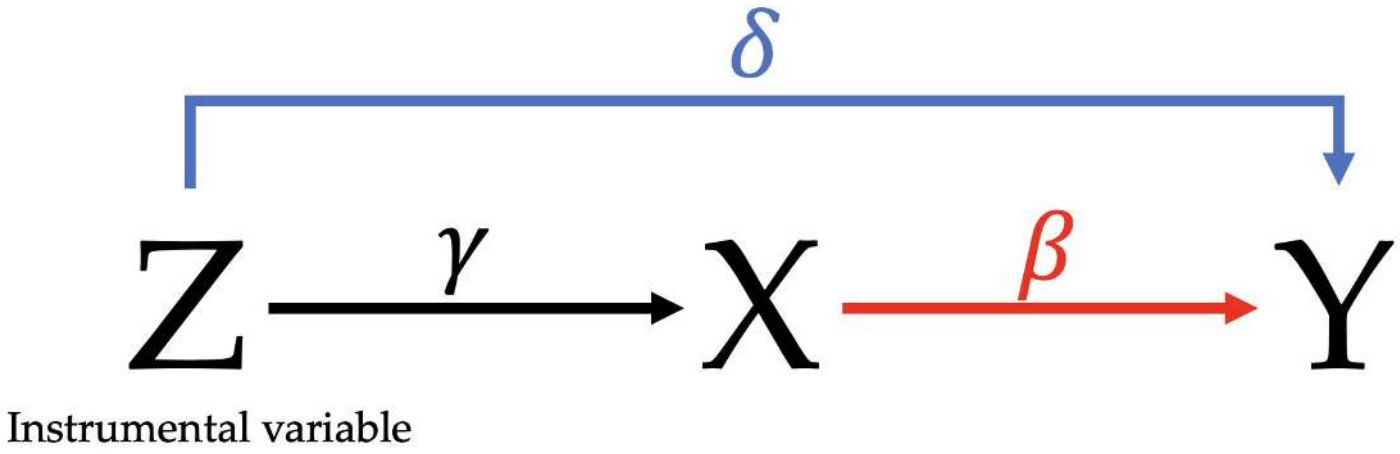
Instrumental Variables

The setup



The setup





How do we calculate β ?

- Assuming Z is a “good” instrument: γ and δ are true causal effects!
- Hence, we can use simple algebra: $\delta=\gamma^*\beta$, so $\beta = \delta/\gamma$
- We can do this more formally on the blackboard

Interpretation of IV: Better LATE than never!

We are interested in the effect of eating chocolate in a given week (D) on happiness (Y).

- Why not estimate: lm(Y ~ D)?

We randomly give some students in the room a voucher to get a chocolate at a 90% discount (Z).

- Why not estimate: lm(Y ~ Z)?

Let us set up the three regressions:

First stage: $D = a + \phi * Z$

Reduced Form: $Y = b + \varrho * Z$

IV: $\lambda = \varrho / \phi$ (Sometimes: Second stage)

IV just “rescales” the effect of the instrument

- How do we interpret ϕ? : The average difference in chocolate consumption between those who got a voucher and those who didn’t
- How do we interpret ϱ? The average difference in happiness between those who got a voucher and those who didn’t

$$\lambda = \frac{\rho}{\phi} = \frac{E\left[Y_i \mid Z_i = 1\right] - E\left[Y_i \mid Z_i = 0\right]}{E\left[D_i \mid Z_i = 1\right] - E\left[D_i \mid Z_i = 0\right]}$$

Better LATE than never!

	Potential outcomes! (unobserved)	***Does not get voucher***	
		Eats chocolate	*Does not eat chocolate*
Gets voucher	*Eats chocolate*	Always-takers: $E(D \mid Z=1)=E(D \mid Z=0)=1$ ➔ $E(Y \mid Z=1)=E(Y \mid Z=0)$	Compliers
	Does not eat chocolate	Defiers	Never-takers: $E(D \mid Z=1)=E(D \mid Z=0)=0$ ➔ $E(Y \mid Z=1)=E(Y \mid Z=0)$

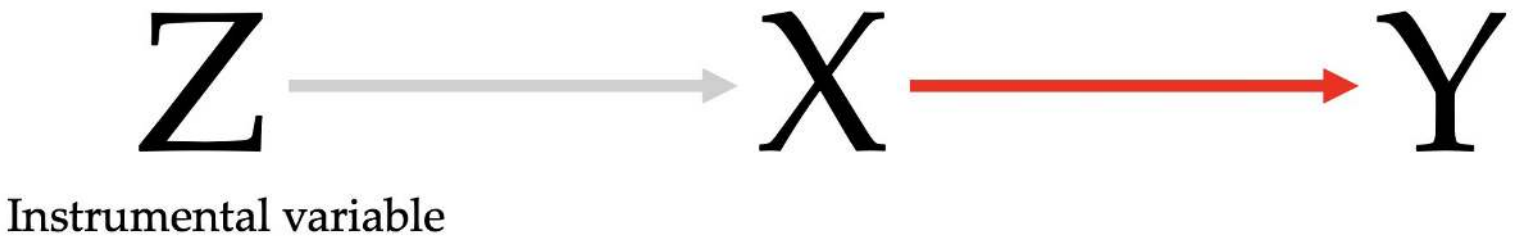
Does this always work?

- No! It only works if we have a valid instrument!
- For this, we need three conditions:

Does this always work?

- No! It only works if we have a valid instrument!
- For this, we need three conditions:

    1. Relevance: Z must truly affect X



Does this always work?

- No! It only works if we have a valid instrument!
- For this, we need three conditions:

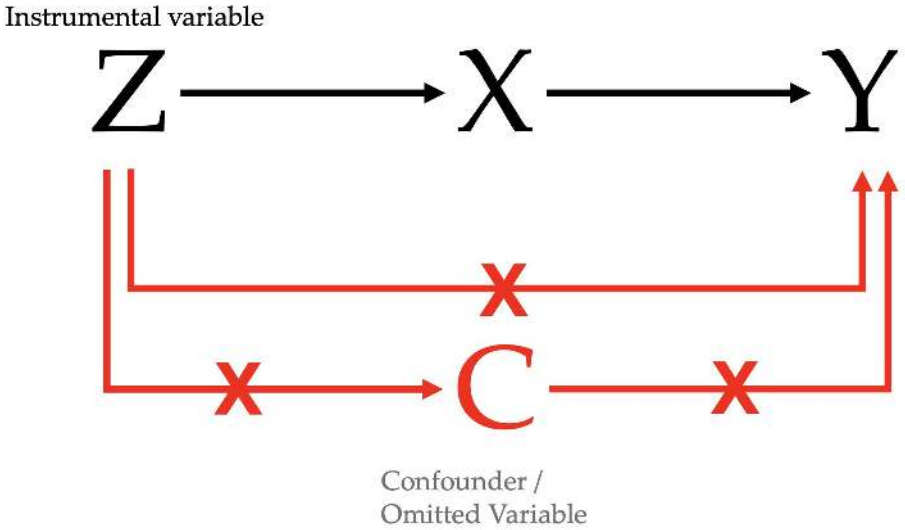
2. Independence: Z is as good as randomly assigned



Does this always work?

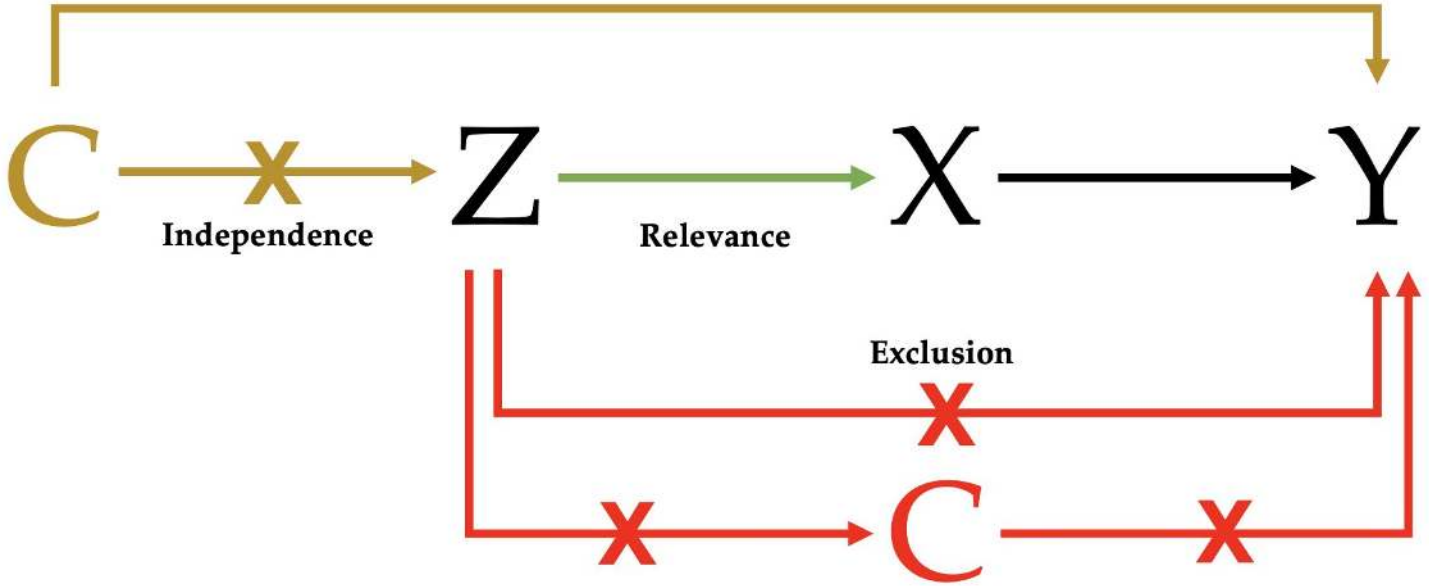
- No! It only works if we have a valid instrument!
- For this, we need three conditions:

3. Exclusion: The **ONLY** way that Z affects Y is via X!



IV Conditions summarized

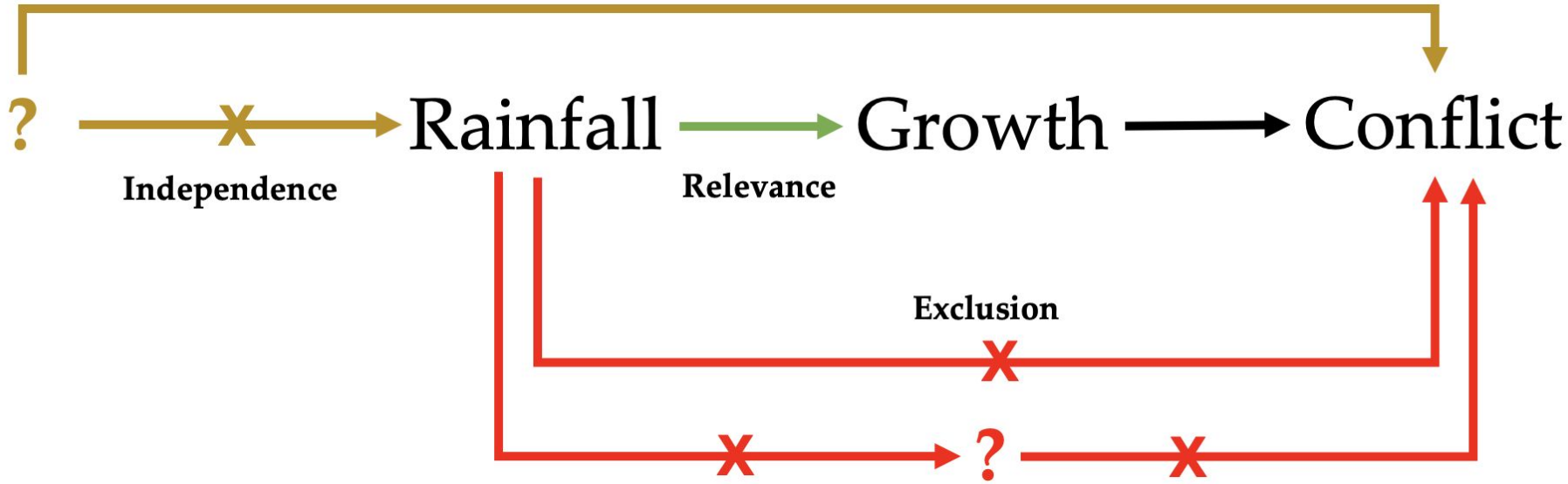
1. Relevance: Z must truly affect X
2. Independence: Z is as good as randomly assigned
3. Exclusion: The **ONLY** way that Z affects Y is via X!



IV: Application



Estimating the Effect of Economic Growth on Conflict



Your job:

A) Find at least one plausible story why this condition **may be violated**.

B) Explain how you would **test** whether this is actually true.

Your turn

We need at least one group for each of the three conditions.

1. Relevance: Z must truly affect X
2. Independence: Z is as good as randomly assigned
3. Exclusion: The **ONLY** way that Z affects Y is via X.

Your job:

A) Find at least one plausible story why this condition **may be violated**.
B) Explain how you would **test** whether this is actually true.

Why IV?

There is a growing body of research that highlights the association between economic conditions and civil conflict (see Sambanis [2001] for a review). However, the existing literature does not adequately address the endogeneity of economic variables to civil war and thus does not convincingly establish a causal relationship. In addition to endogeneity, omitted variables—for example, government institutional quality—may drive both economic outcomes and conflict, producing misleading cross-country estimates.

Relevance

In this paper we use exogenous variation in rainfall as an instrumental variable for income growth in order to estimate the impact of economic growth on civil conflict.[1] Weather shocks are plausible instruments for growth in gross domestic product in economies that largely rely on rain-fed agriculture, that is, neither have extensive irrigation systems nor are heavily industrialized. The instrumental variable method makes it credible to assert that the association between economic conditions and civil war is a causal relationship rather than simply a correlation. As such,

Sub-Saharan Africa is the ideal region for this identification strategy: the World Development Indicator database indicates that only 1 percent of cropland is irrigated in the median African country, and the agricultural sector remains large. We find that weather shocks are in fact closely related to income growth in sub-Saharan Africa (in the first-stage regression). However, our identification strategy is inappropriate for other regions of the world, since weather is not sufficiently closely linked to income growth.[2] Although the analysis is not global, it is likely to be of exceptional interest from both the research and policy perspectives, since the incidence of civil wars in Africa is high and has increased in the past two decades.

Relevance

TABLE 2
RAINFALL AND ECONOMIC GROWTH (First-Stage)
Dependent Variable: Economic Growth Rate, t

EXPLANATORY VARIABLE	ORDINARY LEAST SQUARES				
	(1)	(2)	(3)	(4)	(5)
Growth in rainfall, t	.055***	.053***	.049***	.049***	.053***
	(.016)	(.017)	(.017)	(.018)	(.018)
Growth in rainfall, $t-1$	.034**	.032**	.028**	.028*	.037**
	(.013)	(.014)	(.014)	(.014)	(.015)
Growth in rainfall, $t+1$				.001	
				(.019)	
Growth in terms of trade, t					−.002
					(.023)
Log(GDP per capita), 1979		−.011			
		(.007)			

Exclusion restriction

Edward Miguel

University of California, Berkeley and National Bureau of Economic Research

Shanker Satyanath and Ernest Sergenti

New York University

Estimating the impact of economic conditions on the likelihood of civil conflict is difficult because of endogeneity and omitted variable bias. We use rainfall variation as an instrumental variable for economic growth in 41 African countries during 1981–99. Growth is strongly negatively related to civil conflict: a negative growth shock of five percentage points increases the likelihood of conflict by one-half the following year. We attempt to rule out other channels through which rainfall may affect conflict. Surprisingly, the impact of growth shocks on conflict is *not* significantly different in richer, more democratic, or more ethnically diverse countries.

Exclusion restriction I

A. Potential Violations of the Exclusion Restriction

While it is intuitively plausible that the rainfall instruments are exogenous, they must also satisfy the exclusion restriction: weather shocks should affect civil conflict only through economic growth. In the Introduction above, we acknowledge the possibility that economic channels other than per capita economic growth per se (i.e., income inequality or rural poverty rates) may be key underlying causes of civil conflict in the aftermath of adverse rainfall shocks; unfortunately we do not have reliable cross-country data on these other intermediate channels. There are, however, central government budget figures for approximately half of our sample period from the World Bank, and we find that rainfall growth is not significantly associated with tax revenues (neither total revenues nor revenue as a proportion of national income; results not shown), indicating that changes in governments' fiscal positions are unlikely to be driving our findings.

Exclusion restriction II

A more serious violation of the exclusion restriction is the possibility that high levels of rainfall might directly affect civil conflict independently of economic conditions. For instance, floods may destroy the road network and thus make it more costly for government troops to contain rebel groups. Note that this first possibility is not a serious threat to our estimation strategy, since higher levels of rainfall are empirically associated with significantly *less* conflict in the reduced-form regressions.

Exclusion restriction III

because of more difficult transportation conditions. To explore this possibility, we estimated the impact of rainfall shocks on the extent of the usable road network using World Bank data, and we did not find a statistically significant relationship. In fact, the point estimates on current and lagged rainfall are both positive (e.g., the coefficient on current rainfall growth is 192, standard error 1,025; regression not shown), which argues against the theory above. Another potential violation of the exclusion restriction could occur if low rainfall is associated with heat waves that raise tempers and spontaneously provoke conflict. How-

Results

The main empirical findings are as follows. Using the comprehensive new database of conflicts developed by the International Peace Research Institute of Oslo, Norway, and the University of Uppsala, Sweden, we find that GDP growth is significantly negatively related to the incidence of civil conflict in sub-Saharan Africa during the period 1981–99 across a range of regression specifications, including some with country fixed effects. The relationship between GDP growth and the incidence of civil wars is extremely strong: a five-percentage-point drop in annual economic growth increases the likelihood of a civil conflict (at least 25 deaths per year) in the following year by over 12 percentage points—which amounts to an increase of more than one-half in the likelihood of civil war. Other variables that have gained prominence in the recent

If you want to know more...
...listen to the Freakonomics Podcast!



Other instruments

- We are interested in the effect of being in the army on crime. We instrument being in the army with a lottery (paper)
- We are interested in the effect of protestant religion on economic growth. We instrument protestantism in a region with the distance to Wittenberg (paper)
- We are interested in the effect of air pollution on mortality. We instrument local air pollution with wind direction (paper)

1. Relevance: Z must truly affect X
2. Independence: Z is as good as randomly assigned
3. Exclusion: The **ONLY** way that Z affects Y is via X.

Your job: Discuss the plausibility of the three assumptions

Short term project:
Your questions and ideas

Instrumental Variables and 2SLS

Econ 140, Section 9

Jonathan Old

Roadmap

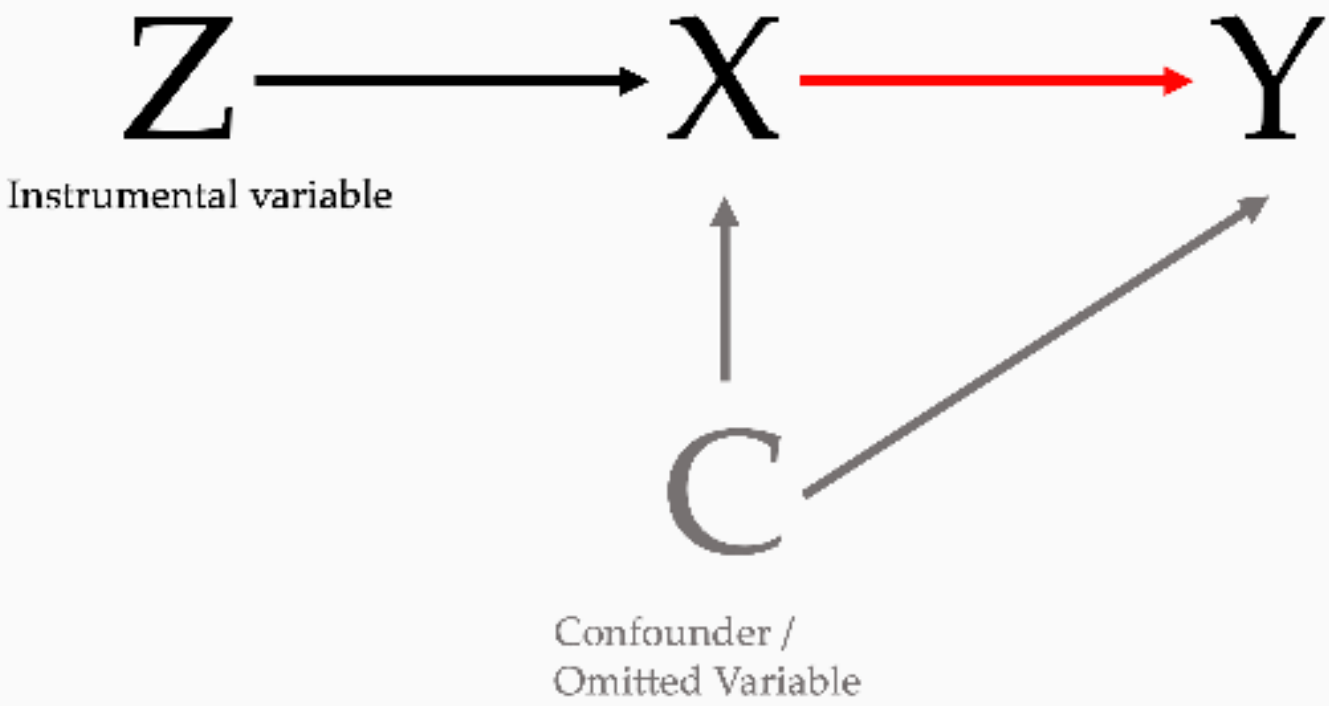
1. Time for your questions
2. Recap: IV
3. Time for your questions
4. Group work
5. Time for your questions
6. Coding

Any questions?

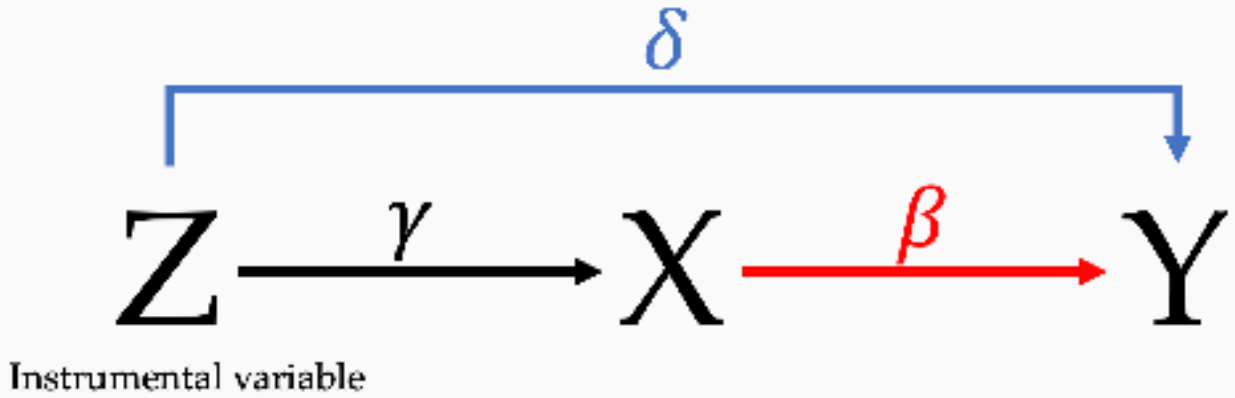
… Remember – Every question is useful!

Recap: IV

Recap: IV



Recap: IV ”rescales” the effect



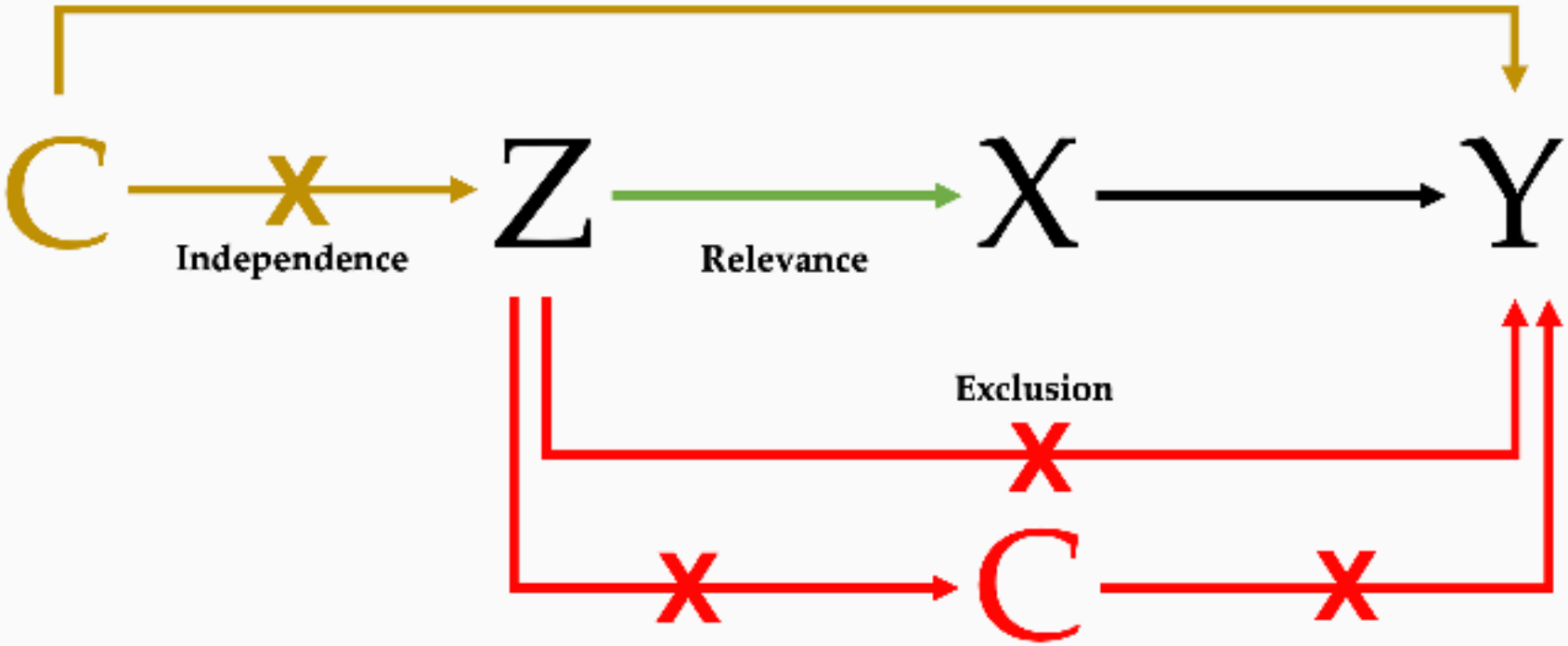
A simple example:

- We want to know the effect of chocolate (X) on happiness (Y), using a randomized voucher as instrument (Z).
- We find: people with voucher were 3 points more happy ($\delta = 3$), and ate 0.5 more chocolates ($\gamma = 0.5$).
- Then, the effect of eating one more chocolate is $\beta = \delta/\gamma = 3/0.5 = 6$.

Recap: IV summary

We need the following three assumptions for IV to work:

1. **Relevance**: Z must truly affect X
2. **Independence**: Z is as good as randomly assigned
3. **Exclusion Restriction**: The **only** way that Z affects Y is via X.



Any questions?

… Remember – Every question is useful!

Group work

Group work

Group 1: We are interested in the effect of being in the army on crime. We instrument being in the army with a lottery (paper)

Group 2: We are interested in the effect of protestant religion on economic growth. We instrument protestantism in a region with the distance to Wittenberg (paper)

Group 3: We are interested in the effect of air pollution on mortality. We instrument local air pollution with wind direction (paper)

1. Relevance: Z must truly affect X
2. Independence: Z is as good as randomly assigned
3. Exclusion restriction: The **only** way that Z affects Y is via X

Your job: Discuss whether these assumptions hold!

Any questions?

… Remember – Every question is useful!

Coding

See you in datahub!

We will meet here.

Instrumental Variables and 2SLS

Today, we are running some code that will help us understand the basics of instrumental variables. We will analyze the relationship between eating chocolate and happiness. Clearly, we cannot run a simple regression: There may be many omitted variables (e.g., people with lactose intolerance are happier, but they also consume more chocolate) or even reverse causality (e.g., when your GSI is stressed and unhappy, they consume tons of chocolate).

Thankfully, we have two potential instrumental variables at hand: 1) We randomly assigned people a voucher that gives them free chocolate, and 2) we know how far they live away from a grocery store.

We will run through the mechanics of the IV estimation. To understand what exactly is going on, we also show you **how the data is generated**, i.e., what the actual **truth** is. This is a trick we can use when we want to check whether a method performs well: We simulate some data, and because we simulated it, we know the truth. Then we can just check whether running a regression with the method we want will give us the correct result.

Setting up the data

We first load the required packages and set the number of observations (3,000 individuals) and a "seed" - this allows us to use random numbers and get exactly the same numbers every time we run the code.

In [1]:
```
install.packages("ivreg")
install.packages("huxtable")
install.packages("jtools")

library('ivreg')
library('huxtable')
library('jtools')


set.seed(12345)
n=3000
```

```
Installing package into ‘/opt/r’
(as ‘lib’ is unspecified)

Installing package into ‘/opt/r’
(as ‘lib’ is unspecified)

Installing package into ‘/opt/r’
(as ‘lib’ is unspecified)
```

Next, we generate a data frame and fill it with some observations. The two instruments (voucher and distance) are random variables (one is a "binomial" random variable and will be a dummy, the other a uniform random variable).

In [6]:
```
data_iv = data.frame(seq(1, n))
colnames(data_iv)="n"

# The first instrument is a dummy variable: A lottery whether you received a vouche
data_iv$voucher = rbinom(n,1,0.5)
```

```
# The second instrument is a continuous variable: The distance to the closest super
data_iv$distance = runif(n,0,1)
```

Next, we generate some other variables: `unobserved_unhappiness` is how unhappy the respondent was *before* buying any chocolate. We do not observe this and this will generate omitted variable bias (strictly speaking, this is reverse causality). We also generate a truly random error that is unrelated to anything else in the data. And we also have data on whether or not a person is lactose intolerant.

In [7]:
```
# These are some other variables: Being unhappy on a given day, an unobserved error
data_iv$unobserved_unhappiness = rnorm(n,0,1)
data_iv$yerror = rnorm(n,0,1)
data_iv$lactose_intolerant = rbinom(n,1,0.5)
```

Finally, we know exactly what determines the consumption of chocolate, and what determines happiness. This is often called the "data-generating process".

In [8]:
```
# This is the "data-generating process" for chocolate consumption:
  # - People who got the voucher eat more chocolate
  # People who live further away from supermarket eat less chocolate, people who ar
data_iv$chocolate = 0.8*data_iv$voucher - data_iv$distance - data_iv$lactose_intole

# This is the DGP for happiness: Eating chocolate makes you happier, being lactose
data_iv$happiness = 1*data_iv$chocolate + data_iv$lactose_intolerant - data_iv$unob
```

Questions for you

- Can you see from the DGP: What is the true effect of chocolate on happiness? What would you want to see as regression result?
- Can you guess: If we run the OLS regression (pretending we do not know unobserved unhappiness), if there will be OVB?
- Are distance and voucher valid instruments in this framework (i.e., do they satisfy the relevance, independence, and exclusion restriction)?

Running OLS

In [11]:
```
# We immediately see that OLS is biased: unobserved_unhappiness is correlated with
summary(lm(happiness ~ chocolate + lactose_intolerant , data=data_iv))
```

```
Call:
lm(formula = happiness ~ chocolate + lactose_intolerant, data = data_iv)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5643 -0.7370 -0.0031  0.7222  4.0859

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        -0.04349    0.02788  -1.560  0.11885
chocolate           0.18416    0.01744  10.558  < 2e-16 ***
lactose_intolerant  0.11486    0.04276   2.686  0.00727 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.073 on 2997 degrees of freedom
Multiple R-squared:  0.03673,   Adjusted R-squared:  0.03609
F-statistic: 57.14 on 2 and 2997 DF,  p-value: < 2.2e-16
```

IV estimation

We can use the `ivreg` package to use the `voucher` as an instrument for `chocolate` consumption.

We can also verify that in this simple setup (where the instrument is a dummy variable), we can simply calculate four averages in the data and get **exactly** the same result - so we don't even need to run a regression!

Cheeky question: Can you come up with at least two reasons why we would still want to run a regression?

In [13]:
```
# Runnnig the IV regression
summary(ivreg(happiness ~ chocolate  | voucher , data=data_iv))



# Implementing the Wald estimator
a = mean(data_iv$happiness[data_iv$voucher==1])
  print(a)
b = mean(data_iv$happiness[data_iv$voucher==0])
  print(b)

c = mean(data_iv$chocolate[data_iv$voucher==1])
  print(c)
d = mean(data_iv$chocolate[data_iv$voucher==0])
  print(d)

wald_estimator = (a-b)/(c-d)
  print(wald_estimator)
```

```
Call:
ivreg(formula = happiness ~ chocolate | voucher, data = data_iv)

Residuals:
     Min       1Q   Median       3Q      Max 
-5.30086 -0.98411 -0.01969  0.99256  5.50574 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.52713    0.05003   10.54   <2e-16 ***
chocolate    1.02429    0.06850   14.95   <2e-16 ***

Diagnostic tests:
                 df1  df2 statistic p-value    
Weak instruments   1 2998     359.2  <2e-16 ***
Wu-Hausman         1 2997     390.0  <2e-16 ***
Sargan             0   NA        NA      NA    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.504 on 2998 degrees of freedom
Multiple R-Squared: -0.8934,    Adjusted R-squared: -0.894 
Wald test: 223.6 on 1 and 2998 DF,  p-value: < 2.2e-16 
[1] 0.3186439
[1] -0.5027664
[1] -0.2035451
[1] -1.005476
[1] 1.024291
```

Two stage least squares

We have seen in class that we can also get the estimate from running two separate regressions and then getting the result as the ration between two OLS coefficients:

In [15]:
```
# Two-Stage least squares
reduced_form = summary(lm(happiness ~ voucher , data=data_iv))
print(reduced_form)
first_stage  = summary(lm(chocolate ~ voucher , data=data_iv))
print(first_stage)


tsls = reduced_form$coefficients[2,1] / first_stage$coefficients[2,1]
print(tsls)
```

```
Call:
lm(formula = happiness ~ voucher, data = data_iv)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6808 -0.6865  0.0049  0.6869  3.5050

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.50277    0.02597  -19.36   <2e-16 ***
voucher      0.82141    0.03700   22.20   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.013 on 2998 degrees of freedom
Multiple R-squared:  0.1412,    Adjusted R-squared:  0.1409
F-statistic: 492.9 on 1 and 2998 DF,  p-value: < 2.2e-16


Call:
lm(formula = chocolate ~ voucher, data = data_iv)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6845 -0.7832 -0.0153  0.8050  3.8856

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.00548    0.02970  -33.86   <2e-16 ***
voucher      0.80193    0.04231   18.95   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.159 on 2998 degrees of freedom
Multiple R-squared:  0.107,     Adjusted R-squared:  0.1067
F-statistic: 359.2 on 1 and 2998 DF,  p-value: < 2.2e-16

[1] 1.024291
```

Advantages of 2SLS

2SLS gives us several advantages:

- We can use **two instruments at the same time**: distance and voucher. This can help us get more precise estimates because we use more information on what determines chocolate consumption
- We can also **control for additional variables** that are important - such as, in our case, lactose intolerance
- We can directly **test whether instruments are relevant**. This is particularly useful if we have multiple instruments (how would we even do it otherwise?). The way we test this is by looking at the so-called "First stage F-statistic" or here, at the test for "Weak instruments".

In [16]:
```
summary(a <- ivreg(happiness ~ chocolate  | voucher , data=data_iv))
```

```r
# Including Distance as instrument
summary(b <- ivreg(happiness ~ chocolate  | distance , data=data_iv))
# Including both instrument
summary(c <- ivreg(happiness ~ chocolate  | voucher + distance , data=data_iv))
# Including control
summary(d <- ivreg(happiness ~ lactose_intolerant + chocolate   | voucher + dista

export_summs(a,b,c,d)
```

```
Call:
ivreg(formula = happiness ~ chocolate | voucher, data = data_iv)

Residuals:
     Min       1Q   Median       3Q      Max 
-5.30086 -0.98411 -0.01969  0.99256  5.50574 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.52713    0.05003   10.54   <2e-16 ***
chocolate    1.02429    0.06850   14.95   <2e-16 ***

Diagnostic tests:
                  df1  df2 statistic p-value    
Weak instruments    1 2998     359.2  <2e-16 ***
Wu-Hausman          1 2997     390.0  <2e-16 ***
Sargan              0   NA        NA      NA    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.504 on 2998 degrees of freedom
Multiple R-Squared: -0.8934,    Adjusted R-squared: -0.894 
Wald test: 223.6 on 1 and 2998 DF,  p-value: < 2.2e-16 
Call:
ivreg(formula = happiness ~ chocolate | distance, data = data_iv)

Residuals:
     Min       1Q   Median       3Q      Max 
-5.43300 -1.01154 -0.02154  1.01202  5.56888 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.55089    0.07253   7.595 4.07e-14 ***
chocolate    1.06321    0.10956   9.705  < 2e-16 ***

Diagnostic tests:
                  df1  df2 statistic p-value    
Weak instruments    1 2998     137.1  <2e-16 ***
Wu-Hausman          1 2997     151.2  <2e-16 ***
Sargan              0   NA        NA      NA    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.538 on 2998 degrees of freedom
Multiple R-Squared: -0.9794,    Adjusted R-squared: -0.9801 
Wald test: 94.18 on 1 and 2998 DF,  p-value: < 2.2e-16 
```

```
Call:
ivreg(formula = happiness ~ chocolate | voucher + distance, data = data_iv)

Residuals:
     Min       1Q   Median       3Q      Max
-5.33925 -0.98975 -0.02054  0.99758  5.52409

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.53404    0.04493   11.89   <2e-16 ***
chocolate    1.03560    0.05803   17.85   <2e-16 ***

Diagnostic tests:
                 df1  df2 statistic p-value
Weak instruments   2 2997   266.611  <2e-16 ***
Wu-Hausman         1 2997   637.897  <2e-16 ***
Sargan             1   NA     0.092   0.761
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.514 on 2998 degrees of freedom
Multiple R-Squared: -0.918,     Adjusted R-squared: -0.9186
Wald test: 318.5 on 1 and 2998 DF,  p-value: < 2.2e-16
Call:
ivreg(formula = happiness ~ lactose_intolerant + chocolate |
    voucher + distance + lactose_intolerant, data = data_iv)

Residuals:
     Min       1Q   Median       3Q      Max
-5.08936 -0.96582 -0.03077  0.94779  5.04358

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         0.05495    0.03764    1.46    0.144
lactose_intolerant  0.94259    0.07475   12.61   <2e-16 ***
chocolate           1.02798    0.05445   18.88   <2e-16 ***

Diagnostic tests:
                 df1  df2 statistic p-value
Weak instruments   2 2996   334.962  <2e-16 ***
Wu-Hausman         1 2996   633.774  <2e-16 ***
Sargan             1   NA     0.056   0.813
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.432 on 2997 degrees of freedom
Multiple R-Squared: -0.7154,    Adjusted R-squared: -0.7166
Wald test:   179 on 2 and 2997 DF,  p-value: < 2.2e-16
Registered S3 methods overwritten by 'broom':
  method            from
  tidy.glht         jtools
  tidy.summary.glht jtools
```

A huxtable: 17 × 5

	names	Model 1	Model 2	Model 3	
	<chr>	<chr>	<chr>	<chr>	
		Model 1	Model 2	Model 3	
1	(Intercept)	0.527133275522715 ***	0.550892307613025 ***	0.53403668341017 ***	0.05495
2		(0.0500264760211853)	(0.0725296359271811)	(0.0449305453865909)	(0.03763
3	chocolate	1.02429063664533 ***	1.06321489358638 ***	1.03560044174847 ***	1.02798
4		(0.0685026617240189)	(0.109557616518632)	(0.0580305178439754)	(0.05445
5	lactose_intolerant				0.9425
6					(0.07475
1.1	nobs	3000	3000	3000	
2.1	r.squared	-0.89341100244053	-0.979414098772623	-0.91800718472674	-0.7154
3.1	adj.r.squared	-0.894042560480037	-0.980074343635456	-0.91864694696314	-0.7165
4.1	sigma	1.50427874523487	1.538063268442	1.51401781632436	1.432
5.1	statistic	223.579258814953	94.1796240401501	318.472224853309	178.9
6.1	p.value	8.24775765906999e-49	6.03885534040896e-22	8.86500481357345e-68	3.697497
7	df	2	2	2	
8	df.residual	2998	2998	2998	
9	nobs.1	3000	3000	3000	
.1	*** p < 0.001; ** p < 0.01; * p < 0.05.	*** p < 0.001; ** p < 0.01; * p < 0.05.	*** p < 0.001; ** p < 0.01; * p < 0.05.	*** p < 0.001; ** p < 0.01; * p < 0.05.	*** p < 0.

In [ ]:

Differences in Differences

Econ 140, Section 10

Jonathan Old

Roadmap

1. Motivation

Motivation

Motivation

- In the first part of the course, we talked about comparison between groups: Treatment vs. Control
- Last week, we talked about comparisons over time: Time series
- Both of these have very big and obvious problems, but we can use them **together** in a powerful tool: Differences-in-differences!

How to get the causal effect of a treatment

In 2021, UC Berkeley offered free mental health coachings to students with pre-existing mental health issues. We want to evaluate the effect of this policy and we collect a depression score (0-10) for all students at UC Berkeley.

	2020	2022
Free Mental Health: Treated	6	6
No Free Mental Health: Untreated	4	5

We can do several comparisons:

How to get the causal effect of a treatment II

	2020	2022
Free Mental Health: Treated		6
No Free Mental Health: Untreated		5

We can do several comparisons:

- Comparison 1: Compare treated and non-treated group after the intervention

How to get the causal effect of a treatment II

	2020	2022
Free Mental Health: Treated		6
No Free Mental Health: Untreated		5

We can do several comparisons:

- Comparison 1: Compare treated and non-treated group after the intervention
- Estimated treatment effect? $6 - 5 = 1$

How to get the causal effect of a treatment II

	2020	2022
Free Mental Health: Treated		6
No Free Mental Health: Untreated		5

We can do several comparisons:

- Comparison 1: Compare treated and non-treated group after the intervention
- Estimated treatment effect? $6 - 5 = 1$
- Problems?

How to get the causal effect of a treatment II

	2020	2022
Free Mental Health: Treated		6
No Free Mental Health: Untreated		5

We can do several comparisons:

- Comparison 1: Compare treated and non-treated group after the intervention
- Estimated treatment effect? $6 - 5 = 1$
- Problems?
- Identifying assumption: The average difference between groups is due to the treatment only. Without the treatment, the average outcome of the treated group would have been equal to the average outcome of the control group.

How to get the causal effect of a treatment III

	2020	2022
Free Mental Health: Treated	6	6
No Free Mental Health: Untreated		

We can do several comparisons:

- Comparison 2: Compare treated group before and after the intervention

How to get the causal effect of a treatment III

	2020	2022
Free Mental Health: Treated	6	6
No Free Mental Health: Untreated		

We can do several comparisons:

- Comparison 2: Compare treated group before and after the intervention
- Estimated treatment effect? $6 - 6 = 0$

How to get the causal effect of a treatment III

	2020	2022
Free Mental Health: Treated	6	6
No Free Mental Health: Untreated		

We can do several comparisons:

- Comparison 2: Compare treated group before and after the intervention
- Estimated treatment effect? $6 - 6 = 0$
- Problems?

How to get the causal effect of a treatment III

	2020	2022
Free Mental Health: Treated	6	6
No Free Mental Health: Untreated		

We can do several comparisons:

- Comparison 2: Compare treated group before and after the intervention
- Estimated treatment effect? $6 - 6 = 0$
- Problems?
- Identifying assumption: The average difference across time is due to the treatment only. Without the treatment, the average outcome of the treated group would not have changed.

How to get the causal effect of a treatment: DiD

	2020	2022
Free Mental Health: Treated	6	6
No Free Mental Health: Untreated	4	5

We can do several comparisons:

- Comparison 3: Compare treated to untreated group, before and after the intervention. **Differences in differences!**

How to get the causal effect of a treatment: DiD

	2020	2022
Free Mental Health: Treated	6	6
No Free Mental Health: Untreated	4	5

We can do several comparisons:

- Comparison 3: Compare treated to untreated group, before and after the intervention. **Differences in differences!**
- Estimated treatment effect?
  $(6 - 6) - (5 - 4) = (6 - 5) - (6 - 4) = -1$

How to get the causal effect of a treatment: DiD

	2020	2022
Free Mental Health: Treated	6	6
No Free Mental Health: Untreated	4	5

We can do several comparisons:

- Comparison 3: Compare treated to untreated group, before and after the intervention. **Differences in differences!**
- Estimated treatment effect?
  $(6 - 6) - (5 - 4) = (6 - 5) - (6 - 4) = -1$
- Identifying assumption: Parallel trends: Without the treatment, the average increase in the outcome of the treated would have been the same as the average increase in the outcome of the untreated.

Sidenote

... We can NOT observe the identifying assumption! We can find evidence for or against it, but we can never be sure!

How to get the causal effect of a treatment: DiD

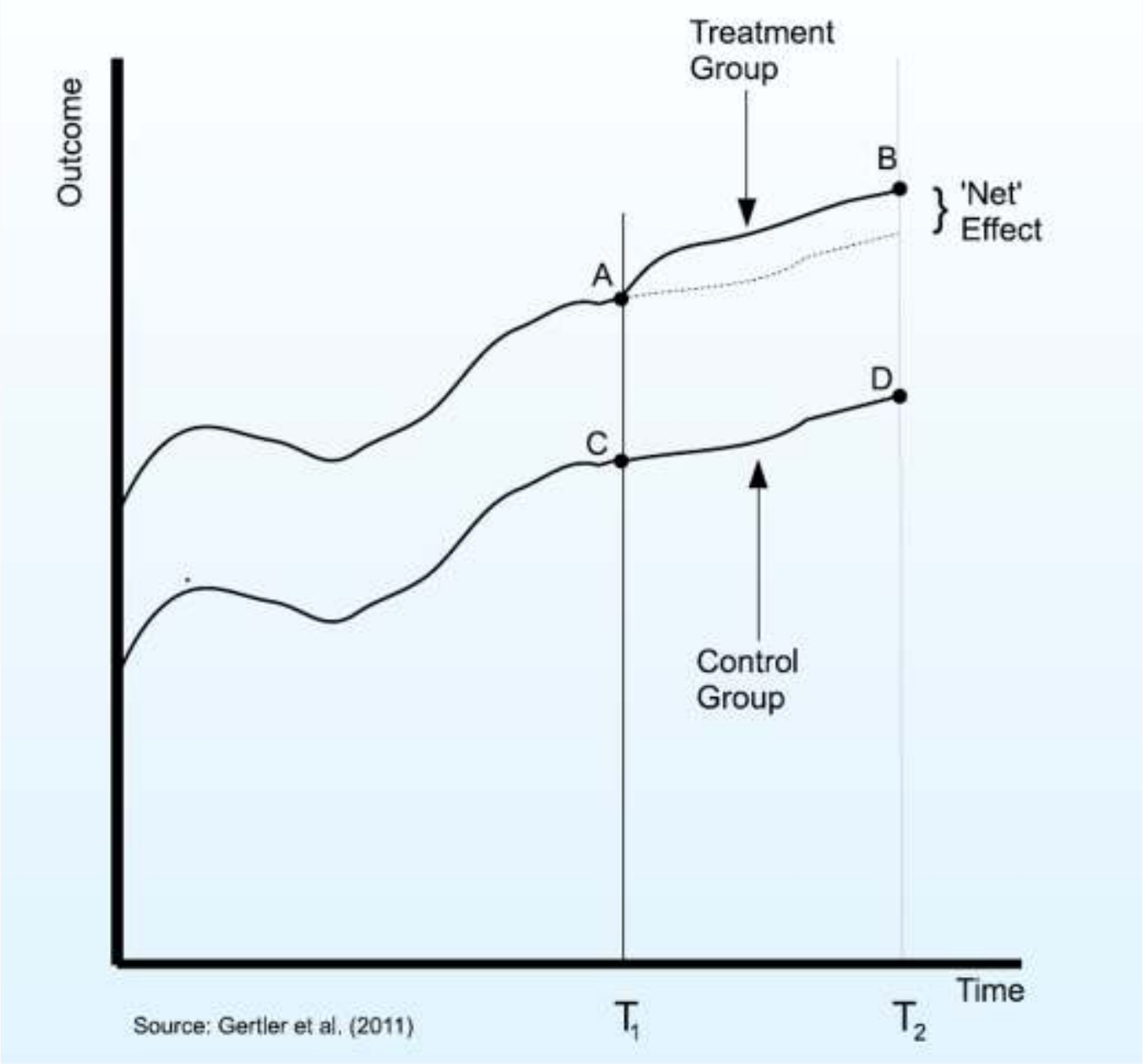
	2020	2022
Free Mental Health: Treated	6	6
No Free Mental Health: Untreated	4	5

We can calculate the DiD estimate in two ways:

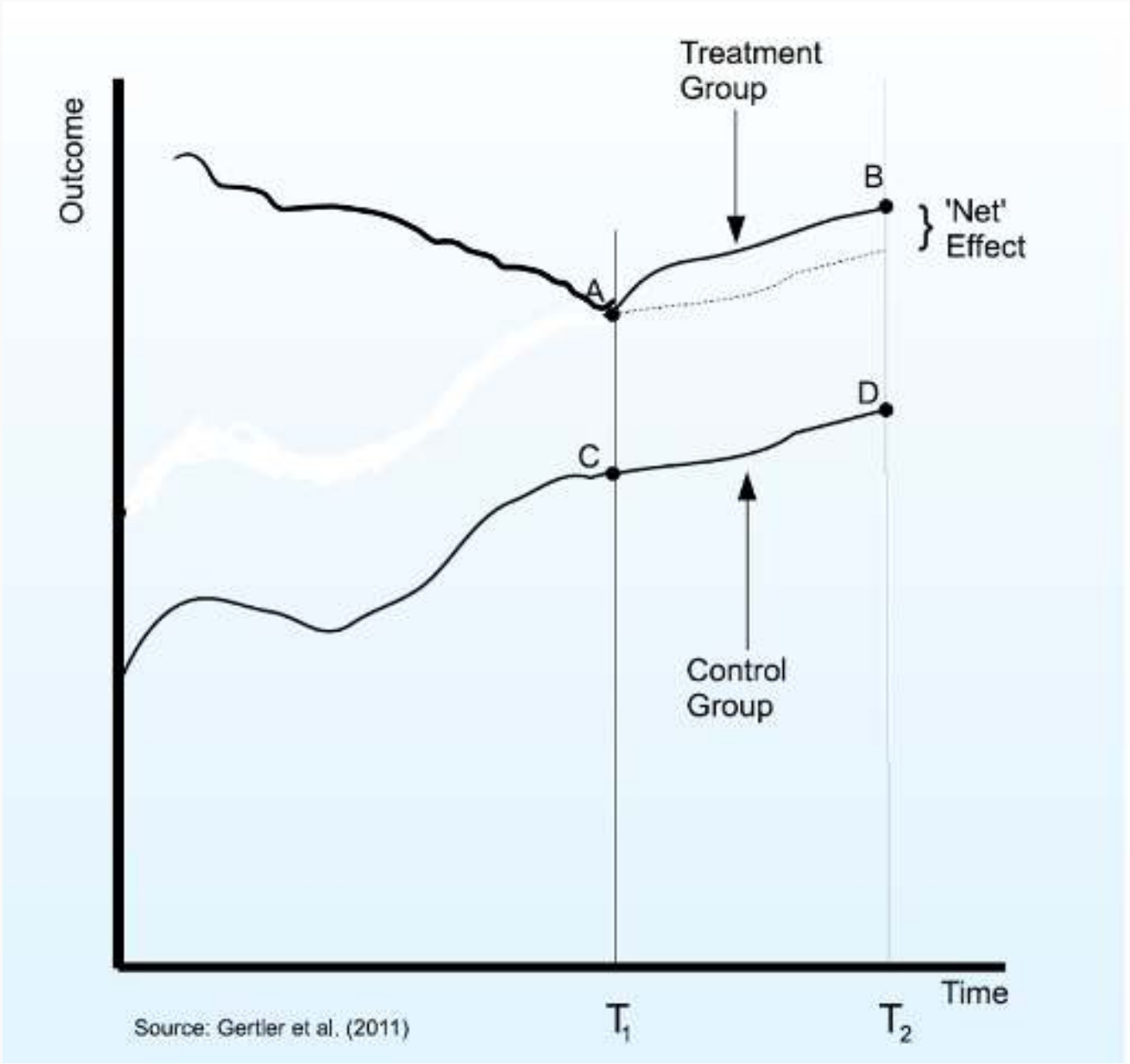
$$DD = E[\ \underbrace{\left(Y_{i1}^{T} - Y_{i0}^{T}\right)}_{\text{Gain over time for treated}} - \underbrace{\left(Y_{i1}^{C} - Y_{i0}^{C}\right)}_{\text{Gain over time for untreated}}\ ]$$

$$= E[\ \underbrace{\left(Y_{i1}^{T} - Y_{i1}^{C}\right)}_{\text{Post-treatment diff treated}} - \underbrace{\left(Y_{i0}^{T} - Y_{i0}^{C}\right)}_{\text{Pre-treatment diff}}\ ]$$

Parallel trends assumption



Parallel trends assumption II



Estimating DiD with regressions

We can set up a simple linear regression model to get the DiD estimate:

$$Y_{it} = \alpha + \beta \text{ Treated }_i + \gamma \text{ Post }_t + \delta \text{ Treated }_i \cdot \text{ Post }_t + u_{it}$$

	2020	2022
Free Mental Health: Treated	$\alpha + \beta$	$\alpha + \beta + \gamma + \delta$
No Free Mental Health: Untreated	α	$\alpha + \gamma$

Estimating DiD with regressions

We can set up a simple linear regression model to get the DiD estimate:

$$Y_{it} = \alpha + \beta \text{ Treated }_i + \gamma \text{ Post }_t + \delta \text{ Treated }_i \cdot \text{ Post }_t + u_{it}$$

	2020	2022
Free Mental Health: Treated	6	6
No Free Mental Health: Untreated	4	5

With our data, we would get:

- $\alpha = 4$
- $\beta = 2$
- $\gamma = 1$
- $\delta = -1$

Interpreting Regression Tables: A (detailed) guide

Interpreting Regression tables

- Standard OLS regression table
  - "Above" the table
  - Main table
  - "Below" the table
  - Your turn
- Special cases
  - Independent variable is dummy
  - Dependent variable is dummy
  - Percentages
  - Logarithmic specifications
  - Quadratic terms
  - Interactions
- …

Standard OLS Regression Table

Back to TOC

The standard OLS regression table (in R)

```
lm(formula = wage_monthly ~ education + male + videogames_childhood,
    data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-361.40  -67.91   -0.56   70.07  364.62

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          2001.1694     9.1823  217.94   <2e-16 ***
education              99.8057     0.6789  147.02   <2e-16 ***
male                   54.8529     3.7536   14.61   <2e-16 ***
videogames_childhood    0.4836     0.8064    0.60    0.549
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 101.8 on 2996 degrees of freedom
Multiple R-squared:  0.8834,    Adjusted R-squared:  0.8833
F-statistic:  7564 on 3 and 2996 DF,  p-value: < 2.2e-16
```

Interpreting Regression tables:
Above the table - Call

```
lm(formula = wage_monthly ~ education + male + videogames_childhood,
    data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-361.40  -67.91   -0.56   70.07  364.62

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          2001.1694     9.1823  217.94   <2e-16 ***
education              99.8057     0.6789  147.02   <2e-16 ***
male                   54.8529     3.7536   14.61   <2e-16 ***
videogames_childhood    0.4836     0.8064    0.60    0.549
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 101.8 on 2996 degrees of freedom
Multiple R-squared:  0.8834,    Adjusted R-squared:  0.8833
F-statistic:  7564 on 3 and 2996 DF,  p-value: < 2.2e-16
```

This section shows the “Call” – the model we gave R to estimate

Here, we ran a regression of monthly wages (in $) on education (in years), a dummy (0-1) variable whether a respondent is male, and the number of hours spent playing videogames in one’s childhood.

Interpreting Regression tables:
Above the table - Call

This section shows the "Call" – the model we gave R to estimate

Here, we ran a regression using data collected at the level of individual respondents. We ran a regression of monthly wages (in $) on education (in years), a dummy (0-1) variable whether a respondent is male, and the number of hours spent playing videogames in one's childhood.

$$\text{Wage}_i = \alpha + \beta_1 \cdot \text{education}_i + \beta_2 \cdot \text{male}_i + \beta_3 \cdot \text{videogames}_i + \epsilon_i$$

This linear regression can serve serveral purposes:

1. To **predict** an individual's wage, based on their information on other variables
2. To **test** whether any of the right-hand side variables is related to wages
3. To **quantify** the direction and strength of the relation between RHS variables and wages
4. To – cautiously – **find a causal estimate of the effect** of a RHS variable on wages.

Interpreting Regression tables:

Above the table – Residuals

```
lm(formula = wage_monthly ~ education + male + videogames_childhood,
    data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-361.40  -67.91   -0.56   70.07  364.62

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          2001.1694     9.1823  217.94   <2e-16 ***
education              99.8057     0.6789  147.02   <2e-16 ***
male                   54.8529     3.7536   14.61   <2e-16 ***
videogames_childhood    0.4836     0.8064    0.60    0.549
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

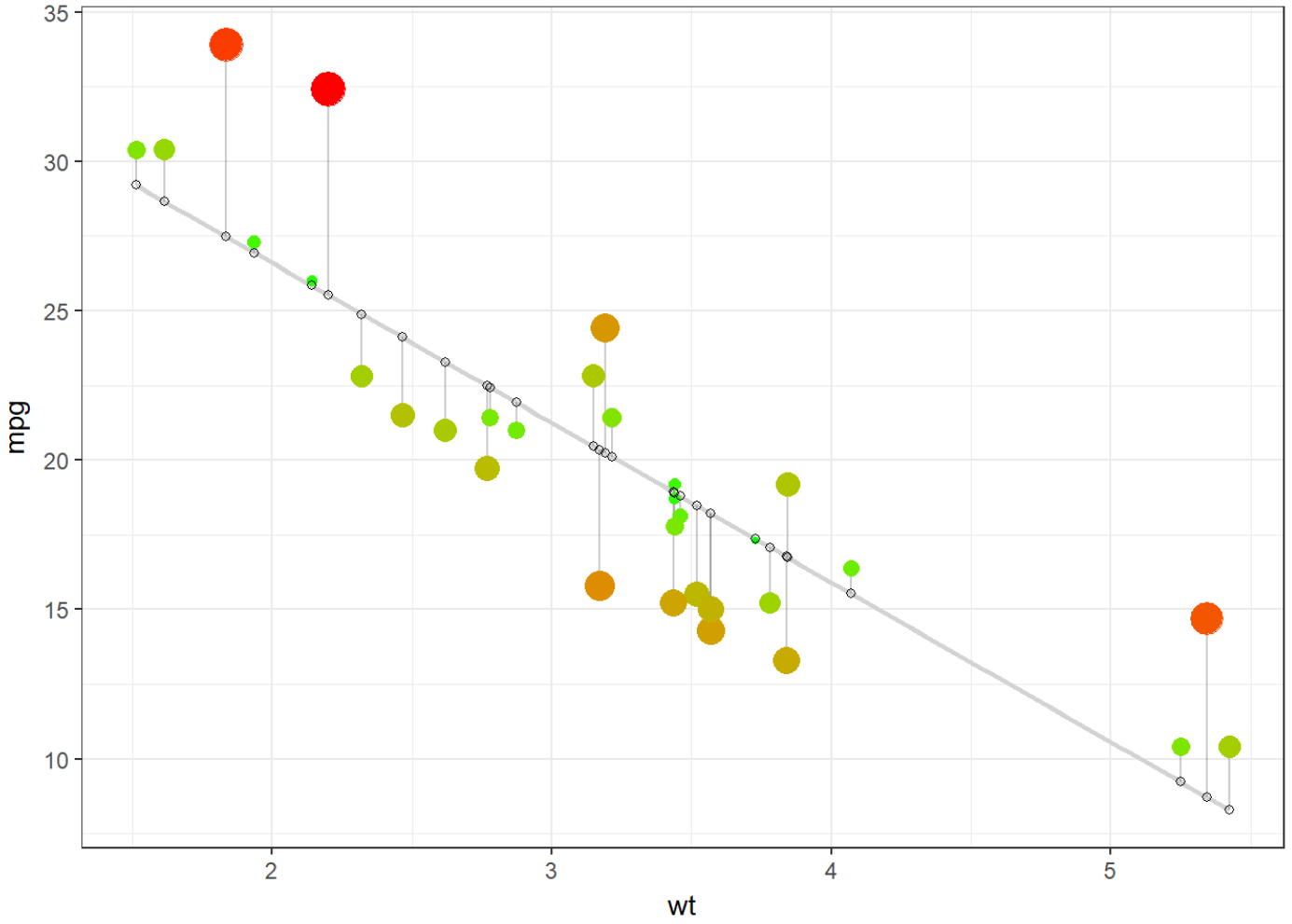
Residual standard error: 101.8 on 2996 degrees of freedom
Multiple R-squared:  0.8834,    Adjusted R-squared:  0.8833
F-statistic:  7564 on 3 and 2996 DF,  p-value: < 2.2e-16
```

This section gives you information about the residuals. Residuals are the difference between your estimated value of y from the model, and the actual data.

Here, the smallest residual is -361.40: There is one individual for whom we predicted (based on their education, gender, and video game history) a wage that was 361$ more than their actual wage.

Interpreting Regression tables:

Above the table – Residuals

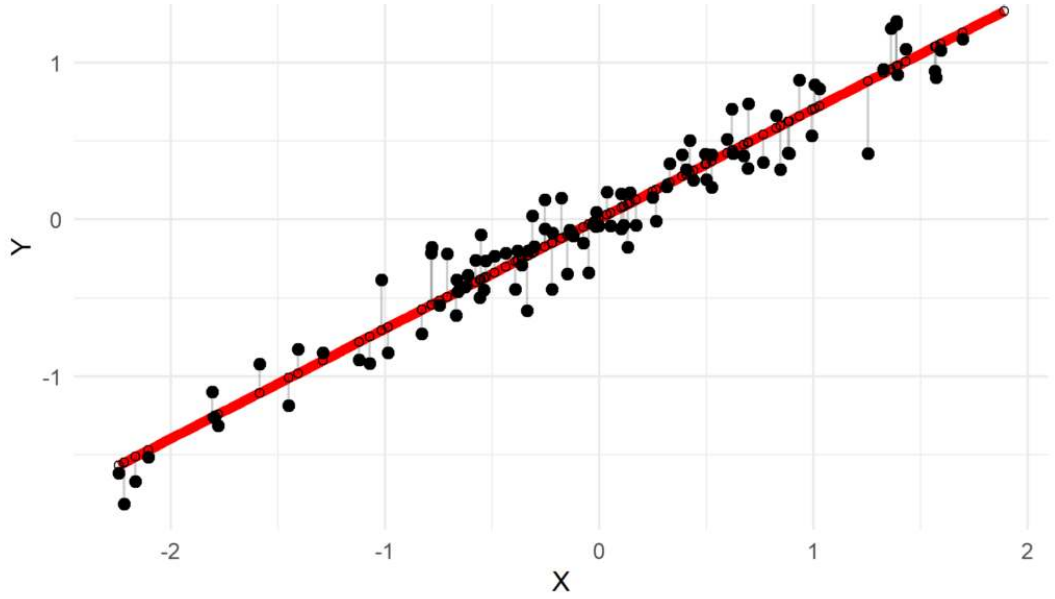


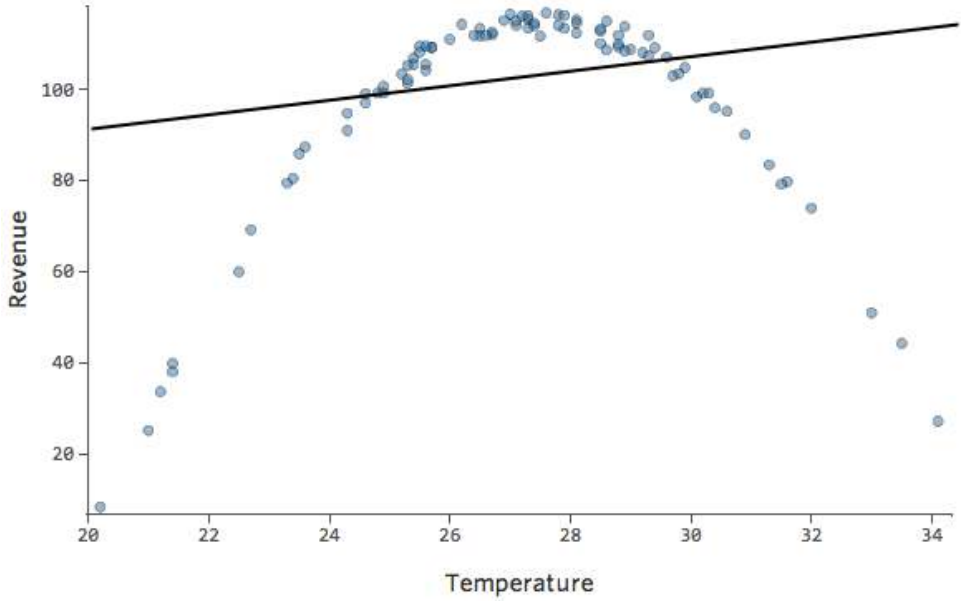
Residuals are simply the difference between the predicted y-value from the model to the actual data – measured in y-direction.

You may remember that OLS is a method that minimizes the sum of the squared residuals!

Interpreting Regression tables:

Above the table – Residuals





"Good" residuals: The residuals are randomly distributed around the estimated regression line

"Bad" residuals: The residuals are systematically related to each other: First negative, then positive, then negative. We have chosen the wrong model to fit the data!

Interpreting Regression tables:
Above the table – Residuals

```
lm(formula = wage_monthly ~ education + male + videogames_childhood,
    data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-361.40  -67.91   -0.56   70.07  364.62

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          2001.1694     9.1823  217.94   <2e-16 ***
education              99.8057     0.6789  147.02   <2e-16 ***
male                   54.8529     3.7536   14.61   <2e-16 ***
videogames_childhood    0.4836     0.8064    0.60    0.549
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 101.8 on 2996 degrees of freedom
Multiple R-squared:  0.8834,    Adjusted R-squared:  0.8833
F-statistic:  7564 on 3 and 2996 DF,  p-value: < 2.2e-16
```

This section can give you a first idea whether your residuals are “ok”.

If the first quartile (Q1) and the third quartile (Q#) of the residuals are far apart, this provides some evidence that the residuals are not “randomly” distributed around 0, but may have a systematic relationship. Similarly, if the median is far away from zero, this may indicate that your model is not appropriate.

Interpreting Regression tables:
Main table

```
lm(formula = wage_monthly ~ education + male + videogames_childhood,
    data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-361.40  -67.91   -0.56   70.07  364.62

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          2001.1694     9.1823  217.94   <2e-16 ***
education              99.8057     0.6789  147.02   <2e-16 ***
male                   54.8529     3.7536   14.61   <2e-16 ***
videogames_childhood    0.4836     0.8064    0.60    0.549
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 101.8 on 2996 degrees of freedom
Multiple R-squared:  0.8834,    Adjusted R-squared:  0.8833
F-statistic:  7564 on 3 and 2996 DF,  p-value: < 2.2e-16
```

This section is the most important one. It gives you information on the coefficients you estimated with the regression. It also gives you information on the precision of the estimate, and information on whether the estimated coefficient is significantly different from zero.

Interpreting Regression tables:
Main table – Intercept

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2001.1694     9.1823  217.94   <2e-16 ***
```

We start with the intercept.
First, let's note that we can write the regression model as:

$$\text{Wage}_i = \alpha + \beta_1 \cdot \text{education}_i + \beta_2 \cdot \text{male}_i + \beta_3 \cdot \text{videogames}_i + \epsilon_i$$

↔️ $\boldsymbol{E}[\textbf{Wage}_i|\textbf{education}_i,\textbf{male}_i,\textbf{videogames}_i\,] = \alpha + \beta_1 \cdot \text{education}_i + \beta_2 \cdot \text{male}_i + \beta_3 \cdot \text{videogames}_i$

So then, the intercept α is simply the expectation of wages, given that all right-hand side (independent) variables are zero.
In other words: **What wage level would we expect for a (hypothetical) individual with 0 years of education, who is female, and who played 0 hours of videogames per week during their childhood?**
Looking at the "estimate" column, the answer is: We would expect a monthly wage of around $2,000.

Interpreting Regression tables:
Main table – Standard error

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2001.1694     9.1823  217.94   <2e-16 ***
```

What about the other numbers? Standard error
The second column gives the standard error for the estimate of the intercept. This is a measure of precision of the estimate: The smaller the standard error, the more precisely did we estimate the intercept.

The standard error on any coefficient answers the question: Given the kind of sample you collected, if you were to collect a different sample (from the same population) many, many times, how far away from the coefficient would your new coefficients be, on average?

Interpreting Regression tables:
Main table – t-statistic

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2001.1694     9.1823  217.94   <2e-16 ***
```

How to use the standard error - t-statistic

If we divide the estimated coefficient by its standard error, we get the corresponding t-statistic. Here, t=2001/9.18=218.

We use the t-statistic to test the null hypothesis that the coefficient is equal to zero. If the null hypothesis is true, the estimator divided by its standard error is distributed according to a t-distribution (similar to a normal distribution for large enough sample sizes), and it would be rare to observe very small or very large values of the t-statistic.

A rule-of-thumb critical value for the t-statistic is 2. If the t-statistic (in absolute value) is greater than 2, the coefficient is significantly different from zero at the 95% confidence level.

Interpreting Regression tables:
Main table – t-statistic

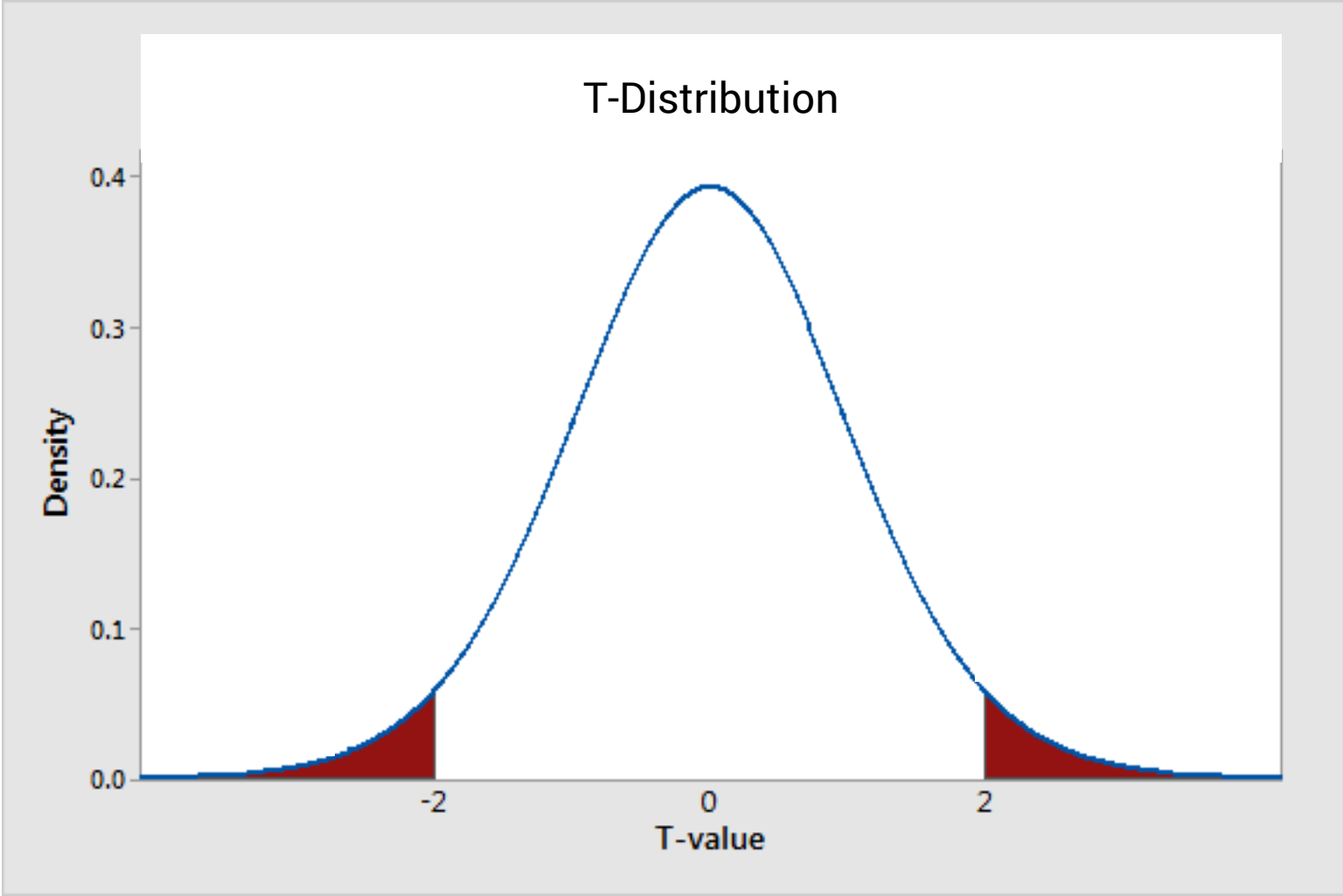
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2001.1694     9.1823  217.94   <2e-16 ***
```

t-statistic - Interpretation
If we divide the estimated coefficient by its standard error, we get the corresponding t-statistic. Here, t=2001/9.18=218.

Since |t|>2, we reject the nul hypothesis. We say: **The estimated intercept is significantly different from zero at the 95% level.**

Interpreting Regression tables:

Main table – t-statistic (and p-value)



Interpreting Regression tables:
Main table – Confidence intervals

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2001.1694     9.1823  217.94   <2e-16 ***
```

How to use the standard error – confidence intervals and other tests
We could also use the standard error to create a (95%) confidence interval for the estimated intercept.

The 95% CI is simply $[\hat{\alpha} - 2 * SE(\widehat{\alpha}),\ \hat{\alpha} + 2 * SE(\widehat{\alpha})]$. In our case, this is [2001- 2*9.18, 2001+2*9.18] = [1983, 2018]. At the 95% confidence level, we can reject the null hypothesis that the intercept is equal to 1982.

A 95% confidence interval means: If we were to draw many, many samples, we would see that 95% of these intervals contain the **TRUE** parameter of interest.

Note: this is **not** the probability that the true parameter of interest is contained in the confidence interval!

Interpreting Regression tables:
Main table – p-value

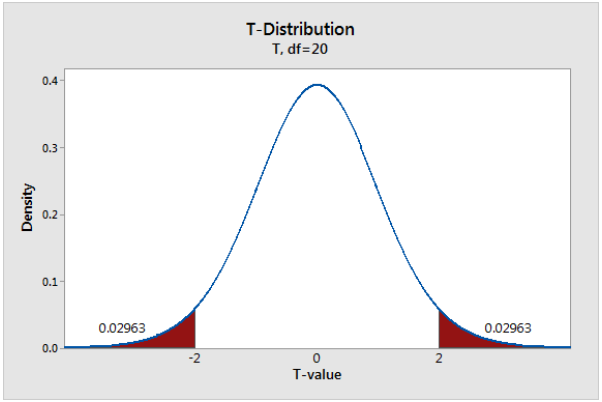
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2001.1694     9.1823  217.94   <2e-16 ***
```

p-value
The p-value is directly related to the t-statistic. For a given value of the t-statistic, if gives the area of the t-distribution that is further to the extreme of that value – in the example to the right, for a t-statistic of 2, it gives the red area.

The p-value is the probability of observing our estimate or an estimate more extreme than ours, given that the null hypothesis is actually true.

In the case here, if the true intercept were 0, the probability of observing (and estimating) an intercept of 2001.17 is less than 2*10^-16, i.e. extremely low.

If the p-value is below our significance level (by convention often 0.05), we say that our estimate is significantly different from zero. This is the case here.



Interpreting Regression tables:
Main table – Slope coefficients (continuous variables)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2001.1694     9.1823  217.94   <2e-16 ***
education     99.8057     0.6789  147.02   <2e-16 ***
```

What about the other coefficients? – Education
All coefficients can be interpreted as **partial** derivatives. To see this, note that if

$$\boldsymbol{E}[\mathbf{Wage}_i|\mathbf{education}_i,\mathbf{male}_i,\mathbf{videogames}_i\,] \quad = \alpha + \beta_1 \cdot \mathrm{education}_i + \beta_2 \cdot \mathrm{male}_i + \beta_3 \cdot \mathrm{videogames}_i$$

Then $\frac{\partial(\mathrm{E}[\mathrm{Wage}_i])}{\partial(\mathrm{education}_i)} = \beta_1$

We can then interpret our estimate of β_1, (denoted $\widehat{\boldsymbol{\beta_1}}$) as follows:
Keeping all other factors constant, every increase in education by one unit (here: one year) is associated with an increase in expected monthly wages by $99.8.

The interpretation of standard error, t-value and p-value is the same. The coefficient on education is significantly different from zero (at the 5% level).

Interpreting Regression tables:

Main table – slope coefficients (Dummy variables)

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         2001.1694     9.1823  217.94   <2e-16 ***
education             99.8057     0.6789  147.02   <2e-16 ***
male                  54.8529     3.7536   14.61   <2e-16 ***
videogames_childhood   0.4836     0.8064    0.60    0.549
```

What about the other coefficients? – Male

Male is a dummy variable. Although the (mathematical) interpretation as partial derivative is the same, it is not very elegant to say we increase male by one unit.

For dummy variables, we can instead say:

Keeping all other factors constant, being male instead of female is associated with an $54.8 higher monthly wages.

The interpretation of standard error, t-value and p-value is the same. The coefficient on male (i.e. the conditional average difference between wages for males and female) is significantly different from zero (at the 5% level).

Interpreting Regression tables:
Main table – insignificant variables

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       2001.1694     9.1823  217.94   <2e-16 ***
education           99.8057     0.6789  147.02   <2e-16 ***
male                54.8529     3.7536   14.61   <2e-16 ***
videogames_childhood   0.4836     0.8064    0.60    0.549
```

What about the other coefficients? – Videogames
The coefficient on videogames is not significantly different from zero.

We say: Keeping everything else constant, every additional hour of videogames played is associated with $0.48 higher monthly wages. However, this association is not significantly different from zero.

We do not say: There is no association between videogames and wages. The effect of videogames on wages is zero.

Why? This is related to statistical testing. If the p-value is >0.05, we **fail to reject** the null hypothesis – but we cannot "accept" the null hypothesis.

Interpreting Regression tables:
Below the table: Residual S.E. and DF

```
lm(formula = wage_monthly ~ education + male + videogames_childhood,
    data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-361.40  -67.91   -0.56   70.07  364.62

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          2001.1694     9.1823  217.94   <2e-16 ***
education              99.8057     0.6789  147.02   <2e-16 ***
male                   54.8529     3.7536   14.61   <2e-16 ***
videogames_childhood    0.4836     0.8064    0.60    0.549
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 101.8 on 2996 degrees of freedom
Multiple R-squared:  0.8834,    Adjusted R-squared:  0.8833
F-statistic:  7564 on 3 and 2996 DF,  p-value: < 2.2e-16
```

This section gives us additional information on our model and how appropriate our model is.

The **residual standard error** gives the standard deviation of the residuals – the smaller the better.

Degrees of freedom is the number of observations used in the regression (3,000) minus the number of coefficients estimated (4). We generally want this to be at least 30.

Interpreting Regression tables:
Below the table: R-squared

```
lm(formula = wage_monthly ~ education + male + videogames_childhood,
    data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-361.40  -67.91   -0.56   70.07  364.62

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          2001.1694     9.1823  217.94   <2e-16 ***
education              99.8057     0.6789  147.02   <2e-16 ***
male                   54.8529     3.7536   14.61   <2e-16 ***
videogames_childhood    0.4836     0.8064    0.60    0.549
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 101.8 on 2996 degrees of freedom
Multiple R-squared:  0.8834,    Adjusted R-squared:  0.8833
F-statistic:  7564 on 3 and 2996 DF,  p-value: < 2.2e-16
```

This section gives us additional information on our model and how appropriate our model is.

The **multiple R-squared** measures what share of the total variation in the outcome variable is explained by our model with our independet variables.

In most cases, it lies between 0 and 1.

Interpreting Regression tables:
Below the table: Adjusted R-squared

```
lm(formula = wage_monthly ~ education + male + videogames_childhood,
    data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-361.40  -67.91   -0.56   70.07  364.62

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          2001.1694     9.1823  217.94   <2e-16 ***
education              99.8057     0.6789  147.02   <2e-16 ***
male                   54.8529     3.7536   14.61   <2e-16 ***
videogames_childhood    0.4836     0.8064    0.60    0.549
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 101.8 on 2996 degrees of freedom
Multiple R-squared:  0.8834,    Adjusted R-squared:  0.8833
F-statistic:  7564 on 3 and 2996 DF,  p-value: < 2.2e-16
```

This section gives us additional information on our model and how appropriate our model is.

The **adjusted R-squared** adjusts R^2 for the number of independent variables in the model. The more independent variables we include, the more it is adjusted downwards.

It is can never be bigger than the multiple R-squared value, and it can also be negative.

Interpreting Regression tables:
Below the table: F-statistic

```
lm(formula = wage_monthly ~ education + male + videogames_childhood,
    data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-361.40  -67.91   -0.56   70.07  364.62

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          2001.1694     9.1823  217.94   <2e-16 ***
education              99.8057     0.6789  147.02   <2e-16 ***
male                   54.8529     3.7536   14.61   <2e-16 ***
videogames_childhood    0.4836     0.8064    0.60    0.549
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 101.8 on 2996 degrees of freedom
Multiple R-squared:  0.8834,    Adjusted R-squared:  0.8833
F-statistic:  7564 on 3 and 2996 DF,  p-value: < 2.2e-16
```

This section gives us additional information on our model and how appropriate our model is.

The **F-statistic** is used to test the null hypothesis that all coefficients on the independent variables are equal to zero.

Under the null hypothesis, it is distributed like an F-distribution, with [number of independent variables] and [regression df] degrees of freedom.

Interpreting Regression tables:
Below the table: F-statistic and joint significance

```
lm(formula = wage_monthly ~ education + male + videogames_childhood,
    data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-361.40  -67.91   -0.56   70.07  364.62

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          2001.1694     9.1823  217.94   <2e-16 ***
education              99.8057     0.6789  147.02   <2e-16 ***
male                   54.8529     3.7536   14.61   <2e-16 ***
videogames_childhood    0.4836     0.8064    0.60    0.549
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 101.8 on 2996 degrees of freedom
Multiple R-squared:  0.8834,    Adjusted R-squared:  0.8833
F-statistic:  7564 on 3 and 2996 DF,  p-value: < 2.2e-16
```

This section gives us additional information on our model and how appropriate our model is.

Here, the **F-statistic** is very large, and the p-value is less than 2.2*10^-16.

Therefore, we reject the null that all coefficients are equal to zero.

We also say that the coefficients in our model are “jointly significant”.

Your turn

Before, we said that we can use regressions:

1. To **predict** an individual's wage, based on their information on other variables
2. To **test** whether any of the right-hand side variables is related to wages
3. To **quantify** the direction and strength of the relation between RHS variables and wages
4. To – cautiously – **find a causal estimate of the effect** of a RHS variable on wages.

Question 1

Predict (give your best guess for) the monthly wages of:

1. A female respondent with 17 years of education who played videogames for 5 hours per week
2. A male respondent with 12 hours of education who played videogames for 1 hour per week.

Your turn

Before, we said that we can use regressions:

1. To **predict** an individual's wage, based on their information on other variables
2. To **test** whether any of the right-hand side variables is related to wages
3. To **quantify** the direction and strength of the relation between RHS variables and wages
4. To – cautiously – **find a causal estimate of the effect** of a RHS variable on wages.

Question 2

Test whether education is (significantly) related to wages.

Note: Our test is always conditional on the other variables in the model.

Your turn

Before, we said that we can use regressions:

1. To **predict** an individual's wage, based on their information on other variables
2. To **test** whether any of the right-hand side variables is related to wages
3. To **quantify** the direction and strength of the relation between RHS variables and wages
4. To – cautiously – **find a causal estimate of the effect** of a RHS variable on wages.

Question 3

How much do we expect monthly wages to differ between:

1. An individual with 15 (undergrad) vs. 12 (high school) years of education?
2. Males and females?

Your turn

Before, we said that we can use regressions:

1. To **predict** an individual's wage, based on their information on other variables
2. To **test** whether any of the right-hand side variables is related to wages
3. To **quantify** the direction and strength of the relation between RHS variables and wages
4. To – cautiously – **find a causal estimate of the effect** of a RHS variable on wages.

Question 4

We found that each additional year of education is associated with $100 higher monthly earnings. **Can we interpret this as the causal effect of one additional year of education?**

Hint: Do we still have any selection bias if we control for gender and videogames? Are there any additional omitted variables? Are there other sources of bias (model misspecification, reverse causality, …)

Back to TOC

Special Cases

Back to TOC

Case 1: Dummy independent variable

```
Call:
lm(formula = wage_monthly ~ male, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-755.30 -209.43  -26.06  201.42 1381.96

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3200.333      7.262  440.68   <2e-16 ***
male         121.810     10.677   11.41   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 291.6 on 2998 degrees of freedom
Multiple R-squared:  0.04161,   Adjusted R-squared:  0.04129
F-statistic: 130.2 on 1 and 2998 DF,  p-value: < 2.2e-16
```

In this model, we regress an outcome variable on a dummy (0-1) independent variable.

Here, we ran a regression of monthly wages (in $) on a dummy (0-1) variable whether a respondent is male.

Back to TOC

Case 1: Dummy independent variable

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3200.333      7.262  440.68   <2e-16 ***
male         121.810     10.677   11.41   <2e-16 ***
```

In cases like this, we can interpret intercepts and coefficients together!

As before, the intercept is the expected value of the dependent variable – wages – if the independent variables – male – are zero. So it is simply the expected value of wages for females in our sample. The expected value for a specific group is just the mean, so the **intercept shows the average wages** for females in our sample - $3,200.

Case 1: Dummy independent variable

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3200.333      7.262  440.68   <2e-16 ***
male         121.810     10.677   11.41   <2e-16 ***
```

In cases like this, we can interpret intercepts and coefficients together!

The coefficient gives the difference between means between the two groups. On average, men earn $121.8 more than women. We can easily calculate the mean wages of men in the sample – they are $3,200+$121.8=$3,321.8.

Case 1: Dummy independent variable

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3200.333      7.262  440.68   <2e-16 ***
male         121.810     10.677   11.41   <2e-16 ***
```

Question 4

How would the regression output change if we ran a regression of wages on a dummy variable that is one if the respondent is female?

Can we run a regression that includes both a male dummy and a female dummy?

Back to TOC

Case 2: Dummy dependent variable

```
Call:
lm(formula = male ~ education, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5656 -0.4560 -0.3682  0.5440  0.6537

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.192806   0.041338   4.664 3.23e-06 ***
education   0.021929   0.003278   6.690 2.65e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4951 on 2998 degrees of freedom
Multiple R-squared:  0.01471,   Adjusted R-squared:  0.01438
F-statistic: 44.76 on 1 and 2998 DF,  p-value: 2.651e-11
```

In this model, we regress a binary/dummy outcome variable on continuous independent variable.

Here, we ran a regression of a dummy (0-1) variable whether a respondent is male, on years of education.

Back to TOC

Case 2: Dummy dependent variable

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.192806   0.041338   4.664 3.23e-06 ***
education   0.021929   0.003278   6.690 2.65e-11 ***
```

In cases like this, we interpret the outcome as probability!

The intercept can be interpreted as follows: For respondents with zero years of education, we would expect the probability of them being male to be 0.19, or 19%. In other words: Among respondents with zero years of education, we would expect a share of 19% to be male.

Case 2: Dummy dependent variable

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.192806   0.041338   4.664 3.23e-06 ***
education   0.021929   0.003278   6.690 2.65e-11 ***
```

In cases like this, we interpret the outcome as probability!

The coefficient on education can be interpreted as follows: For each additional year of education, we expect the probability to observe a male respondent to increase by 0.022, or 2.2 **percentage points**.

For example: If we observe an individual with 10 years of education, we expect them to be male with probability 0.19+0.22*10 = 0.41. Among individuals with 10 years of education, the share of males is expected to be 41%.

Case 2: Dummy dependent variable

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.192806   0.041338   4.664 3.23e-06 ***
education   0.021929   0.003278   6.690 2.65e-11 ***
```

Question 5

How would the regression output change if we ran a regression of a female dummy on education?

Back to TOC