Empirical Economics with R

3 A deeper dive into linear regression and the estimation of causal effects

Uni Ulm

Prof. Dr. Sebastian Kranz

WiSe 20/21

Overview

- In this chapter we return to linear regressions:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_K x_K + u$$

- While classical machine learning is mainly concerned about out-of-sample predictions $\hat{y}$ of the dependent variables, many empirical economic studies are mainly concerned with the estimated coefficients $\hat{\beta}$.
- Typical questions are whether an estimator $\hat{\beta}_k$ "consistently" estimates a particular causal effect of x_k on y, how precise the estimator is, and how sure we are whether the "true effect" is positive or negative.
- This chapter digs deeper into the corresponding methodological background and introduces econometric concepts, like *standard error*, *coefficients of the best linear predictor*, *consistency*, *bias*, *endogeneity*, *confidence intervals*, *p-values*, *significance* and *source of exogenous variation*.
- Some concepts probably seem fairly abstract when you first encounter them, but I believe working through this chapter helps to understand what many empirical research papers talk about.

Empirical Example: Relationship Between Submitted Homwork and Score in Final Exam

- As one empirical example, we take a data set that I have adapted from the data set `attend` in the R package `wooldridge`.
- The data was originally collected by Professors Ronald Fisher and Carl Liedholm. We have observations of a random subset of all students taking an introductionary microeconomics course at Michigan State University.
- We use the following columns:
  - `exam` points in the final exam. Points range between 10 and 39 in the sample.
  - `homework` number of the 8 homework problem sets the student submitted. The homework did not count towards the points of the final exam.
  - `priGPA` grade point average of the student in all previous semesters. Higher grades are better.

Relationship between number of submitted homeworks and exam score.

- We are interested in the relationship between exam performance and the number of submitted homeworks and estimate the simple linear regression.

$$exam = \beta_0 + \beta_1 homework + u$$

- Calling in R the function `summary` on the regression result, we also get for each estimator a so called *standard error*:

```
            Estimate    Std. Error
(Intercept)  22.7294     0.8420
homework      0.4492     0.1170
```

- The OLS estimate $\hat{\beta}_1 = 0.449$ means that one more submitted homework increases our prediction of the exam score by 0.449 points. The standard error $se(\hat{\beta}_1) = 0.117$ is a measure for the precision of the estimator $\hat{\beta}_1$.

One run of a Monte-Carlo simulation

- To better understand what a standard error actually measures, let us repeatedly simulate data and estimate a regression in R. (We perform a so called *Monte-Carlo simulation*.)
- The code on the right simulates a sample with $n = 20$ observations from the regression model

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

  with $\beta_0 = 100$ and $\beta_1 = 1$.
- The estimated coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ are close to the true coefficients β_0 and β_1 but not identical. The difference is due to the random influences from the error term u.

```
set.seed(1)
n = 20
beta0 = 100
beta1 = 1

u = rnorm(n,0,1)
x = rnorm(n,0,1)
y = beta0 + beta1*x + u

# Regress y on x
beta.hat = coef(lm(y~x))
beta.hat
```

```
(Intercept)           x
100.1890485   0.7720182
```

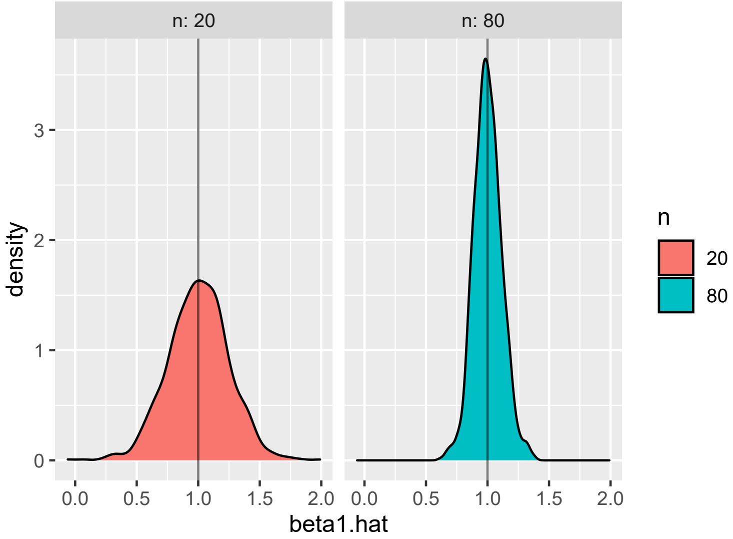
Repeating the Simulation

- If we run the simulation again, we will get in each run somewhat different estimates $\hat{\beta}_0$ and $\hat{\beta}_1$. Here is a table of estimated coefficients from 10 simulation runs:

run	$\hat{\beta}_0$	$\hat{\beta}_1$
1	100.06	0.9
2	99.93	0.93
3	99.72	0.87
4	100.16	0.92
5	99.72	1.36
6	100.05	1.2
7	100.26	1.05
8	100.18	1.01
9	99.86	0.89
10	99.17	0.67

Simulated empirical distribution of $\hat{\beta}_1$

Here are distributions of the simulated $\hat{\beta}_1$ for sample sizes n=20 and n=80 from 1000 runs each:



The OLS estimator can be seen as a random variable

- Given a regression model like

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

the estimator $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$ can be seen as a *random variable* that depends on the random error term u.

- As a random variable, an estimator $\hat{\beta}_k$ has a distribution, an expected value $E\hat{\beta}_k$ and a standard deviation $sd(\hat{\beta}_k)$.
    - The *standard error* of $\hat{\beta}_k$ is its estimated standard deviation.

Formula for standard error in simple linear regression

- Consider a simple linear regression (one explanatory variable)

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

where the u are independently and identically distributed. If we have a single sample of size n, the standard deviation of the OLS estimator $\hat{\beta}_1$ can be *estimated* by

$$se(\hat{\beta}_1) = \hat{sd}(\hat{\beta}_1) = \frac{1}{\sqrt{n}} \frac{sd(\hat{u})}{sd(x)}$$

- We call this estimate of the standard deviation the *standard error* of $\hat{\beta}_1$.
- The standard error gets smaller, i.e. $\hat{\beta}_1$ is more precise if we have...
    - a larger the sample size n
    - a lower standard deviation of the error term u (less noise)
    - more variation in the explanatory variable x (higher standard deviation).

Convergence of the OLS estimator

- Assume that that all observations i are drawn independently from the same data generating process and that the explanatory variables are essentially not perfectly collinear.
- Then as the sample size n grows large...
  - the standard error of each OLS estimator $\hat{\beta}_k$ converges to zero and
  - each OLS estimator $\hat{\beta}_k$ converges in probability to a coefficient β_k^*, i.e.

$$\underset{n\to\infty}{plim}\hat{\beta}_k = \beta_k^*.$$

  We call β_k^* the corresponding coefficient of the best linear predictor (BLP).

- In our previous simulation study the BLP coefficients β^* are equal to the specified coefficients β where $\beta_1 = 1$ measured the causal effect of x on y. But that is not always the case. We discuss this later.
- Notes on the assumptions:
  - One can relax the assumption that the observations i are drawn independently from each other, but then one has to explain concepts like "ergodic stationary stochastic process".
  - We won't discuss the exact no-collinearity condition in this course. For a basic explanation, you can look e.g. at Wikipedia, but the definition must be adapted for random variables in the limit of large n.

Formal definition of the coefficients of the best linear predictor

- Consider a linear regression with k explanatory variables:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_K x_K + u$$

- Assume the values of the dependent and K explanatory variables for each single observation i are drawn from the random variables Y, X_1 to X_K which can have some joint distribution.
- The coefficients of the best linear predictor (BLP) $\beta^* = (\beta_0^*, ..., \beta_K^*)$ solve the following "population equivalent" of the least squares problem:

$$\beta^* = \arg\min_{\beta} E(Y - \beta_0 - \beta_1 X_1 - ... - \beta_K X_K)^2$$

95% Confidence Interval

- With approximately 95% probability, we find an OLS estimate $\hat{\beta}_k$ such that the interval of plus-minus 2 standard errors around $\hat{\beta}_k$ contains the true BLP coefficient β_k^*.
- This interval

$$[\hat{\beta}_k - 2 \cdot se(\hat{\beta}_k)\ ;\ \hat{\beta}_k + 2 \cdot se(\hat{\beta}_k)]$$

  is called the approximate **95% confidence interval** around $\hat{\beta}_k$.

t-value, p-value and significance levels

- If you call the R command `summary` on the output of a regression model, you see besides the estimated coefficients and their standard errors three addition columns:

```
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  99.7930     0.2303 433.335  < 2e-16 ***
x             0.9820     0.2689   3.652  0.00182 ** 
```

- The `t-value` is simply the estimate divided by its standard error. The absolute value of the t-value thus measures how many standard errors the coefficient is away from 0.
- The other columns show so called `p-values` and `significance stars`.

p-values and confidence intervals

- Assume the `summary` of a regression shows a p-value of size p for an estimate $\hat{\beta}_k$. This means that the $(1-p) \cdot 100\%$ confidence interval starts (if $\hat{\beta}_k > 0$) or ends (if $\hat{\beta}_k < 0$) at zero.
  - Example: Assume we have $\hat{\beta}_1 = 1.5$ and a p-value of $p = 0.02 = 2\%$. This means that the 98% confidence interval around $\hat{\beta}_1$ is $[0, 2 \cdot \hat{\beta}_1] = [0, 3]$.
- Hence the smaller the p-value of $\hat{\beta}_k$, the more confident we are that the true BLP coefficient β_k^* has the same sign as $\hat{\beta}_k$.

Significance stars

- The significance stars in the right-most column roughly indicate the rough size of the p-value. Default in R's `summary` function is
  - `.` means $5\% < p \leq 10\%$
  - `*` means $1\% < p \leq 5\%$
  - `**` means $0.1\% < p \leq 1\%$
  - `***` means $p \leq 0.1\%$
- The `stargazer` package and some empirical papers use other defaults where one star `*` is already given for p-values below 10%.
- A common convention in empirical studies is that one calls an estimated coefficient $\hat{\beta}_k$ significantly positive or negative (depending on its sign) if its p-value is below 5%.
  - If the p-value is between 5% and 10% one sometimes speaks of weak significance.
- It is often criticized that many empirical studies concentrate too narrowly on significance and p-values when discussing their results. See e.g. the "Statement on p-Values" by the American Statistical Association.
  - One recommendation is focus more on economic effect sizes and confidence intervals.

Robust Standard Errors

- If the error term u_i is not identically, independently normal distributed, one should use appropriate *robust* standard errors. Most empirical papers in economics use some form robust standard errors.
  - Robust standard errors should then also be used to compute confidence intervals, t-values and p-values.
- We don't explain robust standard errors further in this course. Just note that in R a convenient way to use robust standard errors is the function `lm_robust` in the package `estimatr` or the function `felm` in the package `lfe` .

The homework example

- Let us come back to our initial example. We regressed the points in the final exam of a microeconomics course on the number of submitted homeworks:

$$exam = \beta_0 + \beta_1 homework + u$$

- The function `tidy` in the package `broom` returns summary statistics as a nice data frame and can also directly show the 95% confidence interval of all estimated coefficients. See code on the right.
- Does that mean we are 95% confident that submitting an additional homework problem set *causes* an increase in the average exam score between 0.22 and 0.679?
  - No. The coefficient β_1^* of the best linear predictor does not necessarily measure a causal effect.

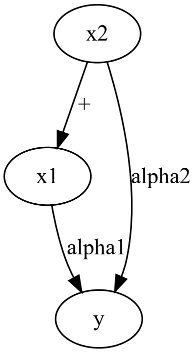
```
reg = lm(exam ~ homework,dat=dat)
broom::tidy(reg, conf.int=TRUE)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	22.7	0.842	27	0	21.1	24.4
homework	0.449	0.117	3.84	0.000134	0.22	0.679

β_k^* does not neccessarily measure a causal effect

Consider the following simulation:

```
n = 10000
alpha0 = 0; alpha1 = 1; alpha2 = 1
u = rnorm(n,0,1)
x2 = rnorm(n,0,1)
x1 = x2+rnorm(n,0,1)
y = alpha0 + alpha1*x1 + alpha2*x2 + u
reg = lm(y~x1) # estimate short regression
broom::tidy(reg,conf.int=TRUE)
```

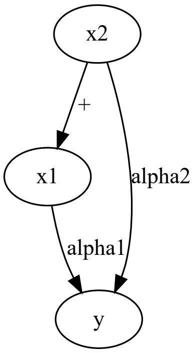


term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-0.01	0.01	-0.73	0.47	-0.03	0.02
x1	1.51	0.01	174.77	0	1.49	1.52

- In the simulation above the causal effect from x_1 on y is $\alpha_1 = 1$. This means if we increase x_1 by one unit, while holding x_2 and u constant, then y increases by α_1 units.
- Yet, we find a quite different OLS estimator $\hat{\beta}_1 = 1.51$ with 95% confidence interval $[1.49; 1.52]$.
- That is because we have a confounder x_2 for which we have not controlled when estimating the short regression

$$y = \beta_0 + \beta_1 x_1 + \varepsilon.$$

- The best linear predictor β_1^* (which can be shown to be 1.5) in this short regression incorporates both the direct causal effect α_1 from x_1 on y and the indirect positive relationship between x_1 and y via x_2.



Consistency of an estimator

- We say that $\hat{\beta}_k$ is a *consistent* estimator of a coefficient β_k if it converges (in probability) against it as the sample size grows large.
  - Otherwise $\hat{\beta}_k$ is an *inconsistent* estimator of β_k.
- This means an OLS estimator $\hat{\beta}_k$ is under weak assumptions a consistent estimator of the coefficient of the best linear predictor β_k^*.
- But often an OLS estimator is an inconsistent estimator of a particular causal effect we might be interested in.

What is β_1?

- Is in the short regression

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

  in our previous example $\hat{\beta}_1$ a consistent estimator of β_1?
- This depends on how we define β_1:
  - if β_1 is defined as the BLP coefficient β_1^* then $\hat{\beta}_1$ is a consistent estimator of β_1.
  - if β_1 is defined as the causal effect of x_1 on y, and if as in our example the causal effect is unequal to β_1^* then $\hat{\beta}_1$ is an inconsistent estimator of β_1.
- Just by looking at the regression equation, we cannot see whether β_1 shall just be the BLP coefficient or whether it is assumed to measure a particular causal effect. This depends on the context and is often verbally stated in research papers and not always perfectly clear.
  - Drawing a causal graph can often be helpful to clear up what β_1 is assumed to measure.

Asymptotic bias (aka the inconsistency)

- The *asymptotic bias*, also called the *inconsistency*, of an estimator $\hat{\beta}_k$ compared to a coefficient β_k is defined as

$$asym.\ bias(\hat{\beta}_k) = \underset{n \to \infty}{plim} \hat{\beta}_k - \beta_k$$

  - A consistent estimator always has a zero asymptotic bias.
- For the OLS estimator of a linear regression the asymptotic bias is the difference between the coefficient of the best linear predictor and β_k:

$$asym.\ bias(\hat{\beta}_k) = \beta_k^* - \beta_k$$

Remark: Bias vs Asymptotic Bias

- There is also a definition of the bias of an estimator given a fixed sample size n. It is defined as:

$$bias(\hat{\beta}_k) = E\hat{\beta}_k - \beta_k.$$

- There are some estimators that have a positive or negative bias for small sample sizes n but still are consistent, i.e. their asymptotic bias is zero.
- If there is a non-zero asymptotic bias, the bias and asymptotic bias will typically have the same sign.
- The difference between the two concepts will not be important for this applied course, but it becomes important once you want to mathematically prove econometric results.

(Asymptotic) Bias Formula

- Consider a simple linear regression ("simple" means one explanatory variable)

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

- One can show that the asymptotic bias for the OLS estimator $\hat{\beta}_1$ satisfies:

$$asym.\, bias(\hat{\beta}_1) = \beta_1^* - \beta_1 = cor(x, u)\frac{sd(u)}{sd(x)}$$

- This means the sign of the (asymptotic) bias of the OLS estimator is the same as the correlation between the explanatory variable x and the error term u.
  - If $cor(x, u) > 0$ then $\hat{\beta}_1$ overestimates the causal effect β_1
  - If $cor(x, u) < 0$ then $\hat{\beta}_1$ underestimates the causal effect β_1
- The OLS estimator $\hat{\beta}_1$ consistently estimates β_1 only if $cor(x, u) = 0$.

Endogeneity and Exogeneity

- Consider a linear model

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_K x_K + u$$

- We say that in the regression an explanatory variable
  - x_k is *exogenous* if it is uncorrelated with the error term $cor(x_k, u) = 0$.
  - x_k is *endogenous* if it is correlated with the error term $cor(x_k, u) \neq 0$.
- The OLS estimator $\hat{\beta} = (\hat{\beta}_0, \ldots, \hat{\beta}_K)$ is consistent only if for all $k = 1, \ldots, K$ every explanatory variable x_k is exogenous.
  - Otherwise one says that there is an *endogeneity problem* that makes the OLS estimator inconsistent.

A simulation example

- On the right we simulate the data generating process:

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

  If we estimate this long regression, both x_1 and x_2 are exogenous and the OLS estimator $\hat{\beta}$ consistently estimates the true β.

- Assume we estimate the short regression

  $$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

  Since the data was generated by the long model, the error term in the short regression is

  $$\varepsilon = \beta_2 x_2 + u$$

  Since $\beta_2 > 0$ and x_2 positively affects x_1, we have $cor(x_1, \varepsilon) > 0$. This means x_1 is endogenous in the short regression and the OLS estimator $\hat{\beta}_1$ of the short regression inconsistently estimates the true causal effect β_1.

```
n = 10000
beta0=0
beta1 = 1
beta2 = 1
u = rnorm(n,0,1)
x2 = rnorm(n,0,1)
x1 = x2+rnorm(n,0,1)
y = beta0 + beta1*x1 + b
eta2*x2 + u
# short regression
coef(lm(y~x1))
```

```
(Intercept)          x1
-0.01312561  1.49382451
```

```
# Error term short regression
eps = beta2*x2+u
# x1 is endogenous
cor(x1,eps)
```

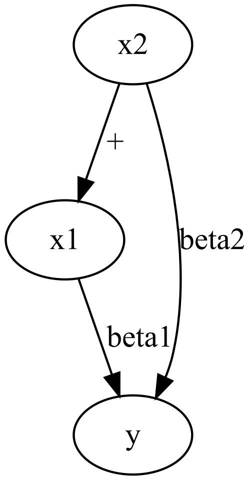
```
[1] 0.4916275
```

Causal graphs vs correlation with error term

- We have now learned two related approaches to think about the (in)consistency and (asymptotic) bias of an OLS estimator β_k.

1. Think about whether there are unobserved confounders and how they bias your estimator by drawing causal graphs, like the one on the right.
2. Think about whether or not x_k is exogenous or endogenous by thinking about its correlation with the error term of the regression. Using the bias formula you may also make a good educated guess for the sign of the (asymptotic) bias.

- Both approaches are usually very similar (and yield the same results), but sometimes one can be more helpful than the other:
  - Causal graphs are particularly useful to distinguish between confounders and channel variables and to make clearer which causal effect we want to measure.
  - Some types of bias are hard to characterize with causal graphs but can be nicely understood with the bias formula. One example is the so called attenuation bias that arises when our explanatory variable x is imprecisely measured.



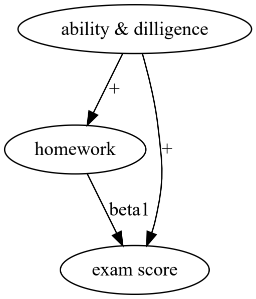
Do we consistently estimate the causal effect in the homework example?

- Let us come back to the homework example.
- Let us define β_1 in the short regression

$$exam = \beta_0 + \beta_1 homework + u$$

as the causal effect of one more solved homework problem set on exam score.

  - This means β_1 shall measure by how much a student's exam score would increase on average if he solved and submitted one more homework, keeping everything else (like student's characteristics) constant.
- Is it reasonable that the OLS estimator $\hat{\beta}_1$ of the short regression consistently estimates this causal effect β_1?
  - No. There are most likely unobserved confounders like ability or diligence (German: "Fleiß") of a student. A more diligent student probably solves more homework problem sets but would also get a better exam score absent homework, because she learns more.



- Assume the causal effects are as in the graph on the right.
- Here $\hat{\beta}_1$ is systematically larger than the causal effect β_1 since it incorporates the indirect positive correlation from the fact that more homeworks are submitted by more diligent students.
- This means $\hat{\beta}_1$ has a positive bias: $E(\hat{\beta}_1) > \beta_1$.
- Equivalently, the graph implies that in the short regression

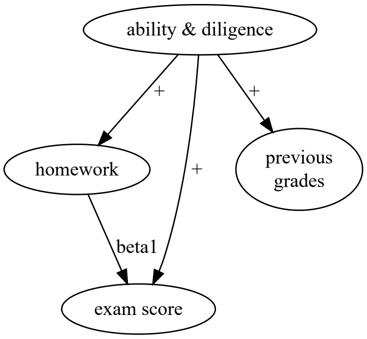
$$exam = \beta_0 + \beta_1 homework + u$$

the error term u positively depends on a student's ability and diligence, which also positively affects `homework`.

- Hence `homework` is endogenous with $cor(homework, u) > 0$.
- From the bias formula we thus also see that $\hat{\beta}_1$ has a positive (asymptotic) bias.

Controlling with a proxy variable

- Factors like diligence and ability can typically not be directly measured.
- Sometimes one tries to control for unobserved confounders by so called proxy variables. These are variables that are correlated (ideally strongly) with the unobserved confounders.
- In our setting the grade point average (GPA) of prior semesters `priGPA` would be such a proxy variable for the unobserved ability and diligence of a student.



- On the right you see the regression results with and without proxy. We used the `stargazer` package to format the results in a table in a similar format as those in economic articles:
- Controlling for prior grades (which proxies factors like ability and diligence) reduces the positive effect of an additional homework on the exam score. Because we now control for part of the indirect positive relationship between homework and exam score that arises from the unobserved heterogeneity in student's ability and diligence.
- Yet, it is still very doubtful that controlling for `priGPA` already yields an estimator $\hat{\beta}_1$ for the causal effect β_1 that only has a small bias. The main problem is that we have no clear source of exogenous variation in `homework` that is uncorrelated to factors that directly affect exam performance.

	Dependent variable:	
	exam	
	(1)	(2)
homework	0.449***	0.110
	(0.117)	(0.115)
priGPA		3.128***
		(0.327)
Constant	22.729***	17.008***
	(0.842)	(0.991)
Observations	674	674
R^2	0.021	0.139
Note:	*p<0.1; **p<0.05; ***p<0.01	

Simulation study about bias, proxy noise and exogenous variation

- Consider the simulation on the right. Controlling for an unobserved confounder x_2 with a proxy variable reduces but not fully eliminates the bias of the OLS estimator $\hat{\beta}_1$ for the causal effect $\beta_1 = 1$ of x_1 on y.
- `sd.proxy.noise` measures how noisy our proxy measures the confounder x_2. How will larger / smaller noise affect the remaining bias? What if we add another noisy proxy? Can that help to reduce the bias? Study in R.
- `sd.exogenous.variation` measures how much exogenous variation we have in x_1 that is not due to variation in the confounder x_2. How does more or less exogenous variation affect the bias and standard errors of $\hat{\beta}_1$? Study in R.

```
n = 10000
u = rnorm(n,0,1)
# unobserved confounder
x2 = rnorm(n,0,1)
sd.exogenous.variation = 1
x1 = x2+rnorm(n,0,sd.exogenous.variation)
beta0=0; beta1 = 1; beta2=1
y = beta0+beta1*x1 + beta2*x2 + u
# short regression -> biased
coef(lm(y~x1))[2]
```

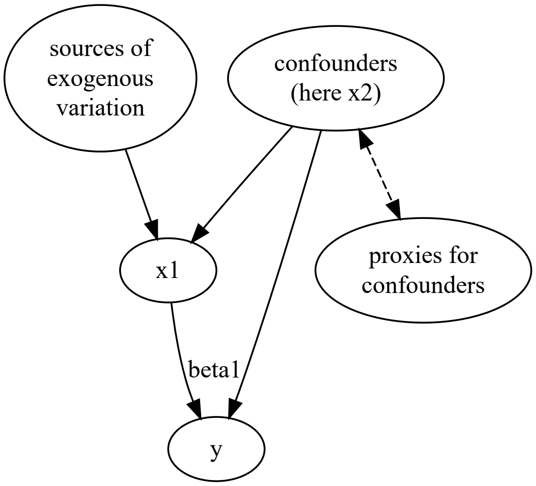
```
      x1
1.517203
```

```
# Add noisy proxy: reduces bias
sd.proxy.noise = 1
proxy = x2+rnorm(n,0,sd.proxy.noise)
coef(lm(y~x1+proxy))[2]
```

```
      x1
1.338691
```

Simulation study about bias, proxy noise and exogenous variation

- The following factors reduce the bias of $\hat{\beta}_1$
  - Less noise in our proxy for the confounder. Controlling with more than one proxy can also help.
  - More exogenous variation in x_1 (uncorrelated to confounders)
- In non-experimental real world data, we will probably never perfectly control for all confounders. All control variables are typically to some extend noisy proxies.
- If we want to ensure that we estimate a causal effect with little bias using an OLS regression with control variables, it is therefore not only important to think about controlling for all relevant confounders but also to think about whether there are sources that generate sufficient exogenous variation in our explanatory variable of interest.



The homework example: Are there sources of exogenous variation?

- What factors possibly create variation across students in their number of submitted homework? We already thought of students' general ability and diligence and tried to control for these confounders using prior average grades as proxy.
- What other factors can influence the number of submitted homeworks and are not perfectly controlled for by prior grades? To my mind come e.g.
  - How much time has the student in this semester to work on his microeconomic problem sets.
  - How much more does the student like (or is good in) this microeconomics course compared to other courses from previous semesters?
- Are these factors confounders or sources of exogenous variation?
  - All of these factors seem to be confounders. E.g. if a student has less time to work on the problem sets, she probably also has less time to directly learn for the exam and thus gets worse grades.
  - I actually cannot think of any obvious source of exogenous variation in the number of submitted homework problem sets. Any factor that affects the number of submitted problem sets seems likely to be also correlated through other non-controlled channels with how much the student learns for the exam or how good she performs in the exam.

- If there is no clear source of exogenous variation in an explanatory variable x, it is typically impossible to estimate its causal effect on a variable y. Even if we would have proxy variables to control for all confounders, we still would get a large bias if the proxies are imperfect and there is no or little exogenous variation in x.
- Actually, I don't see any plausible empirical strategy to consistently estimate the causal effect of solving one more homework on the exam score using this non-experimental data set.
- Unfortunately, it is the case for many non-experimental data sets that it is very hard or impossible to consistently estimate particular causal effects.

A different example: The effect of education on wages

- The data set `wage2` in the R package `wooldridge` contains information about the monthly wages, years of education and other background variables for 935 men.
- A commonly studied (and tough to answer) question is how additional education affects a person's labor market outcome measured by wages.
- Here the variable `edu` measures the years of formal education.
    - Unfortunately, I could not find how this variable was exactly generated. Consider a German employee who made Abitur (13 years) and a Bachelor that regularly takes 3 years, but for which he needed 4 years. Is `edu` equal to the actual time 13+4 = 17 years or equal to the regular time 13+3 = 16 years?
    - I will assume the latter.

A wage regression

- On the right we estimate two linear regressions:

$$wage = \beta_0 + \beta_1 edu + \varepsilon$$

$$wage = \beta_0 + \beta_1 edu + \beta_2 IQ + \beta_3 age + u$$

- β_1 shall denote the causal effect of one additional year of education on wages.
- The variable IQ is test score of of an IQ test taken at school age. We add it as noisy proxy variables for an employees abilities. We also control for age.
- Similar to the intuition from our homework example, we find that the estimated coefficient in front of `edu` goes down once we control with our ability proxy. But how convincing is it that we estimate with only little bias the causal effect of `edu` on `wages` ?

	Dependent variable:	
	wage	
	(1)	(2)
educ	60.214***	41.623***
	(5.695)	(6.446)
IQ		5.368***
		(0.942)
age		21.887***
		(3.907)
Constant	146.952*	-870.379***
	(77.715)	(160.478)
Observations	935	935
R^2	0.107	0.162
Note:	*p<0.1; **p<0.05; ***p<0.01	

Unobserved confounder in our wage-education regression

- What other factors, we have not yet controlled for, could lead to variation in the level of education? To mind come:
    - Parents' education, income and attitudes towards education.
    - Other characteristics of a student like curiosity, social skills, ...
    - The business cycle
    - ...
- Those factors seem to be confounders. E.g. parent's attitudes might affect a decision to study but also affect the job choice and resulting wage.
- It seems very hard to estimate the causal effect of additional education on wages given so many confounders and no really clear source of exogenous variation in education.
- However, on the next slide we see the abstracts of two studies that use a *natural experiment* that creates exogenous variation in the length of schooling in order to measure the effect of longer schooling on health outcomes.

Abstracts of two empirical studies

Parental Education and Child Health: Evidence from a Natural Experiment in Taiwan (Chou et. al. 2010)

Abstract: In 1968, the Taiwanese government extended compulsory education from 6 to 9 years and opened over 150 new junior high schools at a differential rate among regions. Within each region, we exploit variations across cohorts in new junior high school openings to construct an instrument for schooling, and employ it to estimate the causal effects of mother's or father's schooling on infant birth outcomes in the years 1978-1999. Parents' schooling does cause favorable infant health outcomes. The increase in schooling associated with the reform saved almost 1 infant life in 1,000 live births.

Education and Mortality: Evidence from a Social Experiment (Meghir et. al. 2018)

Abstract: We examine the effects on mortality and health due to a major Swedish educational reform that increased the years of compulsory schooling. Using the gradual phase-in of the reform between 1949 and 1962 across municipalities, we estimate insignificant effects of the reform on mortality in the affected cohort. From the confidence intervals, we can rule out effects larger than 1–1.4 months of increased life expectancy. We find no significant impacts on mortality for individuals of low socioeconomic status backgrounds, on deaths that are more likely to be affected by behavior, on hospitalizations, and consumption of prescribed drugs.

Remarks on the two studies (1)

- Both abstracts first mention their (similar) source of exogenous variation across children's years of schooling. Both exploit a new law that created new schools across the country and allowed more children to go longer to school.
    - Importantly, not all children were affected in the same way by the law. New schools were not build at the same time in all cities and whether you got more years of schooling depended on your exact age when the new school was build. It is assumed that this variation is essentially uncorrelated with the outcome variable of interest and therefore is a source of exogenous variation in the years of schooling across individuals.
- The first study has the term *natural experiment* in its title. This expression is often used if we have some particular event, like a law change that is not implemented simultaneously in all regions, that creates a source exogenous variation that allows us to consistently estimate some causal effect.

Remarks on the two studies (2)

- If you don't come from Sweden or Taiwan, your initial reaction to the studies might be: Who cares about the effects of more schooling in Sweden or Taiwan and for people who went to school more than half a century ago? Why don't economists systematically study the effects of schooling for all countries and also use newer data?
    - The problem is that it is difficult and often impossible to consistently estimate causal effects with field data. In particular you need a clear source of exogenous variation. Often we only have it if we are lucky that a natural experiment took place. This means for many interesting economic questions, we have empirically very clean answers only for sporadic examples across time and space.
    - Of course, economists also try to answer important questions even if there are no natural experiments. But answers are then typically more debatable because they depend on many assumptions which often are hard to test.
- Both studies analyse a relationship between education and health. Very roughly said..
    - The Taiwan study finds a positive effect
    - The Sweden study finds no significant effect and the small confidence intervals rule out large effects.
- The different results illustrate that a lot of factors (place, time, exact question, ...) matter and that the examples do not generalize too broadly. Unfortunately, that seems to be the case for many questions in economics and social sciences.

Remarks on the two studies (3)

- Even if one has a source of exogenous variation for x, one still needs to deal with confounders. But often one does not have data for all confounders. A standard multiple linear regression with control variables only for a subset of confounders is not enough to consistently estimate causal effects.
- If you look in more detail at the studies, you will see that they employ methods called *instrumental variable estimation*, *difference-in-differences* or *regression discontinuity*. Some of these methods we study later in this course.
- Note that all methods to consistently estimate a causal effect require some source of exogenous variation in the explanatory variable whose causal effect we want to measure.

Randomized Experiments

- The scientific gold standard to estimate a causal effect is to run a randomized experiment.
- Consider the data set `apple` in the package `wooldridge`.
- 660 households were asked how many regular apples and how many ecolabelled apples they wanted to buy.
- We are interested in how the price of apples affects the demanded quantity.
- In the data set the prices for both types were randomly chosen for each household independent of any household characteristics.
    - This random choice of prices creates exogenous variation in the explanatory variable.
    - Moreover since prices are independently chosen of household characteristics, we can rule out that there are unobserved confounders that systematically affect prices.
- A short variable description:
    - `ecolbs` The quantity of ecological apples bought.
    - `ecoprc` The stated price of ecological apples.
    - `regprc` The stated price of regular apples, which could be bought alternatively.
    - `male` , `age` , `hhsize` , ... background variables of the household.

Results of 3 regression specifications

Code:

```
library(wooldridge)
data(apple)
reg1=lm(ecolbs ~ ecoprc, data=ap
ple)
reg2=lm(ecolbs ~ ecoprc+regprc,
 data=apple)
reg3=lm(ecolbs ~ ecoprc+regprc+m
ale+hhsize+age+faminc, data=appl
e)

stargazer(reg1,reg2,reg3,type="h
tml", omit.stat=c("f","adj.rsq",
"ser"))
```

	Dependent variable:		
	ecolbs		
	(1)	(2)	(3)
ecoprc	-0.845**	-2.926***	-2.949***
	(0.331)	(0.588)	(0.593)
regprc		3.029***	3.060***
		(0.711)	(0.715)
male			-0.108
			(0.227)
inseason			-0.176
			(0.206)
hhsize			0.053
			(0.069)
age			0.001
			(0.007)
faminc			0.003
			(0.003)
Constant	2.388***	1.965***	1.703***
	(0.372)	(0.380)	(0.591)
Observations	660	660	660
R^2	0.010	0.036	0.041
Note:	*p<0.1; **p<0.05; ***p<0.01		

Exercise questions for apple experiment

Let us train and sharpen our understanding with the following exercise questions about the apple experiment:

a) If we have a well randomized experiment, is then the OLS estimator in our second regression

$$ecolbs = \beta_0 + \beta_1 ecoprc + \beta_2 regprc + u$$

consistent? Are the signs of $\hat{\beta}_1 < 0$ and $\hat{\beta}_2 > 0$ consistent with what we would expect from economic theory?

Solution

In a well randomized experiment, the offered prices `ecoprc` and `regprc` should be uncorrelated to all other factors that may affect demand, like e.g. household characteristics. All these other factors are part of the error term u in our regression.

Hence, in a well randomized experiment, we have $cor(ecoprc, u) = 0$ and $cor(regprc, u) = 0$, i.e. both explanatory variables are exogenous. As we have previously stated, the OLS estimator is consistent if all explanatory variables are exogenous.

The signs of the estimated coefficients also make economic sense. We expect a higher price of the eco apples to reduce their demand, i.e. $\beta_1 < 0$. In contrast, if the price of the regular apples (a subtitute for eco apples) increases, we would expect more people to buy eco apples instead, i.e. $\beta_2 > 0$

b) If we don't add `regprc` does the OLS estimator seem to be biased? If yes, in which direction?

Solution

In column 1 we estimate the short regression, where we omit `ecoprc` . There the estimator is $\hat{\beta}_1 = -0.84$, which is substantially larger (i.e. less negative) than in the long regression, where we had $\hat{\beta}_1 = -2.92$. This suggests a positive omitted variable bias if we don't include `regprc` .

Note that small changes in the coefficient will virtually always occur if we add or remove another control variable even if it the control variable is not a confounder, i.e. even if there is no omitted variable bias. So how can we say that a change is large? One indicator are the standard errors. The coefficient change between -0.84 and -2.92 is more than 6 times the estimated standard error (0.33) of $\hat{\beta}_1$ in the first regression. In terms of standard errors, this is a very large change. Personally, I would consider changes of more than one standard error already relatively large.

In column 1 we estimate just the short regression

$$ecolbs = \beta_0 + \beta_1 ecoprc + \varepsilon$$

We find in the short regression a larger (less negative) estimate $\hat{\beta}_1 = -0.84$ than in the second regression. This means omitting `regprc` yields a positive bias in this short regression.

c) Looking at the regression results derive the likely sign of the correlation between the two prices `ecoprice` and `regprice` in the experiment.

Solution

Detailed solution using bias formula

In column 1 we estimate just the short regression

$$ecolbs = \beta_0 + \beta_1 ecoprc + \varepsilon$$

The bias formula for this short regression is

$$asym.\, bias(\hat{\beta}_1) = cor(ecoprc, \varepsilon)\frac{sd(\varepsilon)}{sd(ecoprc)}.$$

We already established in b) that this bias is positive. Hence, we must have

$$cor(ecoprc, \varepsilon) > 0.$$

Given that the short and long regressions are representations of the same data generating process, the error term in the short regression must be

$$\varepsilon = \beta_2 regprc + u.$$

We thus have

$$cor(ecoprc, \varepsilon) = \beta_2 cor(ecoprc, regprc).$$

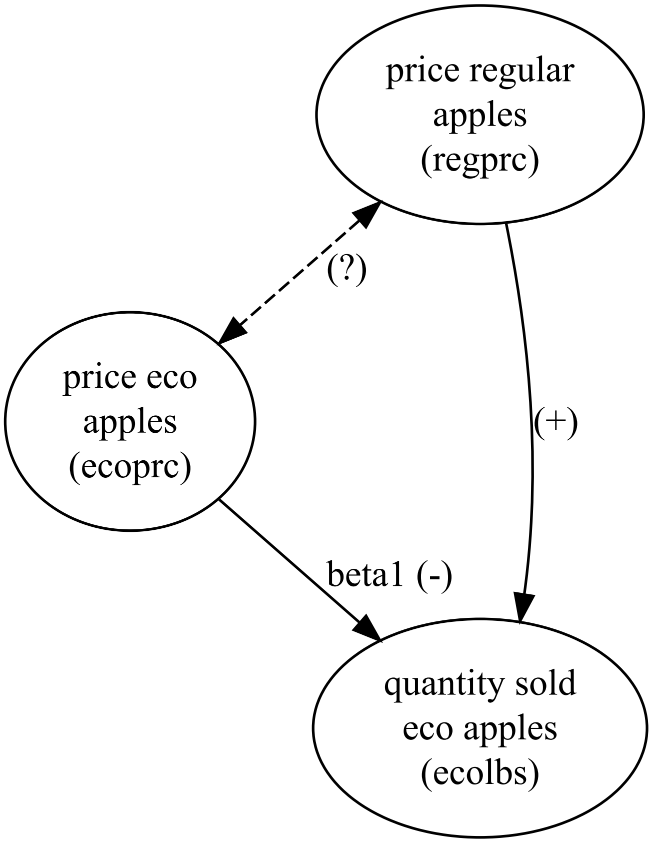
Since $\beta_2 > 0$, we thus also know that

$$cor(ecoprc, regprc) > 0.$$

This means in the experiment higher prices for ecological apples were systematically combined with higher prices for regular apples.

Short solution using a graph

We can come to the same result by drawing a small graph:


price regular apples (regprc)
(?)
price eco apples (ecoprc)
(+)
beta1 (-)
quantity sold eco apples (ecolbs)

We know that the OLS estimator $\hat{\beta}_1$ in the short regression measures the direct effect β_1 from `ecoprc` on demand plus the indirect effect due to the correlation between `ecoprc` and `regprc`. Since our estimator in the short regression is larger (less negative) than in the long regression, this indirect effect must be positive. Hence `ecoprc` and `regprc` must be positively correlated.

Actual correlation

Indeed, we can verify such a strong positive correlation between the two prices in our data set:

```
cor(apple$ecoprc, apple$regprc)
```

```
[1] 0.8307587
```

d) Assume you were not sure whether the prices were indeed correctly randomized over households, i.e. chosen independently of household characteristics. Which of the following results suggest that we indeed had proper randomization?

1. The fact that no estimated coefficients for household characteristic is significant in regression 3.
2. The fact that the coefficient $\hat{\beta}_1$ for `ecoprc` does almost not change between regressions 2 and 3.

Solution

1. That the household characteristics do not significantly affect the demand for eco apples does not tell us anything about the randomization procedure. For a proper randomization, we want household characteristics to be uncorrelated with the prices `ecoprc` and `regprc` , but they may still affect the demand. This result just shows that the demand for eco apples does not seem to depend much on those household characteristics.
2. That the price coefficient does not change between regression 2 and 3 suggests that in regression 2 the household characteristics are not any omitted confounders that generate a bias. And they are no confounders if they are not correlated with the price. This suggests that indeed prices were properly randomized and not systematically correlated with any household characteristics.

Interpreting effect sizes

- In our regression

$$ecolbs = \beta_0 + \beta_1 ecoprc + \beta_2 regprc + \varepsilon$$

we estimated $\hat{\beta}_1 = -2.926$.
- (Causal) Interpretation: A price increase by 1 Dollar per lbs, reduces demand for ecological apples on average by -2.926 lbs. (1 lbs is roughly 454 grams).
- To get a better grasp of effect sizes, it is often helpful to put them in relation to the means of the variables. In our data, the mean of `ecolbs` is `1.47` and the mean of `ecoprc` is `1.08` .
- This means if we increase prices by 10% of the average price (`=0.108 $/lbs`), we estimate to reduce demand by `2.926 * 0.108 = 0.316 lbs` , which is `0.316 / 1.47 = 21.5%` of the average demand.
- We can perform a similar transformation for confidence intervals: With the `confint` function, we find that the 95% confidence interval for $\hat{\beta}_1$ is `[-4.08, -1.77]` . Hence we are 95% confident that a price increase by 10% of the average price, reduces demand between `4.08 * 0.108 / 1.47 = 30%` and `1.77 * 0.108 / 1.47 = 13%` of the average demand.

Non-linear Effects

- In our regressions for the apple experiment, we assumed that the prices affect the demand in a linear fashion:

$$ecolbs = \beta_0 + \beta_1 ecoprc + \beta_2 regprc + u$$

- This means we assume a constant causal effect β_1: one dollar price increase reduces demand by the same amount irrespective of the current price.
- But of course real world demand functions need not to be linear. In chapter 1, we already discussed linear regressions in logs. But also other specifications can be easily estimated via OLS, e.g. a demand function that is quadratic in `ecoprc` and `regprc` :

$$ecolbs = \beta_0 + \beta_1 ecoprc + \beta_2 ecoprc^2 + \beta_3 regprc + \beta_4 regprc^2 + u$$

- Note that this is still called a linear regression, since it is a linear function in the coefficients β and the error term u.
- The causal effect of an increase of `ecoprc` by one dollar on demand is now assumed to be:

$$\frac{\partial ecolbs}{\partial ecoprc} = \beta_1 + 2 \cdot \beta_2 \cdot ecoprc,$$

  i.e. it depends on the current price level.

Comparing linear and quadratic demand function

- We estimate linear and quadratic specifications on the right.
- Standard errors in the quadratic specification are much larger than in the linear specification. This often occurs because the quadratic and linear prices are highly correlated with each other.
- In the quadratic specification it is hard to see the sign of the aggregate effect of `regprc` . Without knowing the range of `regprc` , we don't know whether the negative linear coefficient or the positive quadratic coefficient dominates.
- The linear specification thus can be more precisely estimated and is easier to interpret. Therefore often research papers just estimate linear effects.

	Dependent variable:	
	ecolbs	
	(1)	(2)
ecoprc	-2.926***	-0.742
	(0.588)	(2.633)
ecoprc_sqr		-0.999
		(1.178)
regprc	3.029***	-5.361
	(0.711)	(4.601)
regprc_sqr		4.717*
		(2.555)
Constant	1.965***	4.308**
	(0.380)	(1.753)
Observations	660	660
R^2	0.036	0.041
Note:	*p<0.1; **p<0.05; ***p<0.01	

Code:

```
apple$ecoprc_sqr = apple$ecoprc^2
apple$regprc_sqr = apple$regprc^2
reg4=lm(ecolbs ~ ecoprc+ecoprc_sqr+
      regprc+regprc_sqr,data=apple)
stargazer(reg2,reg4,type="html", omit.st
at=c("f","adj.rsq","ser"))
```

Simulation: Quadratic causal effects

- We can think of a linear function approximating more complex non-linear functions.
- Consider the simulation on the right: x affects y in a quadratic fashion.
- We estimate both the quadratic and linear function.

```
n = 10000
alpha1 = 1; alpha2 = 1
u = rnorm(n,0,1)
x = runif(n,0,10)
y = alpha1*x + alpha2*x^2 + u
x2 = x^2 # quadratic term
coef(lm(y~x+x2)) # quadratic
```

```
(Intercept)           x          x2
-0.02501774  1.00792462  0.99993912
```

```
coef(lm(y~x)) # linear
```

```
(Intercept)           x
  -16.38409    10.95043
```

- For our true model

$$y = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + u$$

  the causal effect of a marginal increase in x is simply the derivative

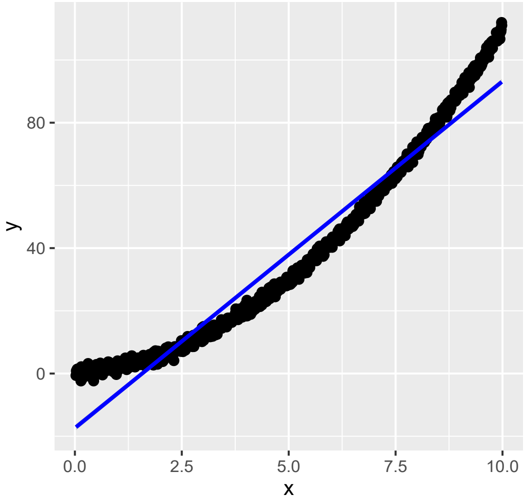
$$\frac{dy}{dx} = \alpha_1 + 2\alpha_2 x = 1 + 2x$$

  - For our range of x between 0 and 10 it varies between 1 and 21.
- In our short regression

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

  we estimated a linear causal effect of $\hat{\beta}_1 \approx 11$ which is approximately the average of the true causal effects over the range of x.

Graphical illustration how the linear function approximates a quadratic relationship.



Heterogeneous Effects

- In our experiment the causal effect of the price on demand for apples possibly differs between customers. E.g. a price increase may affect demand of rich households less strongly than demand of poorer households.
  - We say then that our explanatory variable of interest (here price) has *heterogeneous effects*.
- One way to estimate heterogeneous effects is to run separate regressions for subgroups.

- On the right, we run separate regressions for richer households with income above the median income and the poorer half of households.
  - Note that in the moment we don't add `regprc` as control for didactic purposes.
- We estimate that for the richer households a price increase reduces demand by less than for poorer households (-0.237 vs -1.28)

```
# Create dummy for richer half of households
apple = mutate(apple,richer = 1L*(faminc > median(faminc)))

# richer households
reg.r = lm(ecolbs ~ ecoprc,
  data=filter(apple, richer==1))
# poorer households
reg.p = lm(ecolbs ~ ecoprc,
  data=filter(apple, richer==0))

rbind(richer=coef(reg.r), poorer=coef(reg.p))
```

```
       (Intercept)    ecoprc
richer    1.857381 -0.237818
poorer    2.777275 -1.279665
```

Regression with interaction terms

- We can also estimate heterogeneous treatment effects with the following regression on the whole data set:

$$ecolbs = \beta_0 + \beta_1 ecoprc + \beta_2 richer \\ + \beta_3 \; richer \cdot ecoprc + \varepsilon$$

- The product `richer * ecoprc` is called an interaction effect between our dummy variable `richer` and the price `ecoprc` .

Looking at the results (column (3)), we see that

- $\hat{\beta}_1 = -1.28$ is identical to the price effect for the poorer housholds.
- $\hat{\beta}_1 + \hat{\beta}_3 = -1.28 + 1.042 = -0.238$ is the price effect for the richer households.
- $\hat{\beta}_3 = 1.042$ is the difference of the price effect between the richer and poorer households. While relatively large in absolutely terms, it is not statistically significantly different from 0.

	Dependent variable:		
		ecolbs	
	richer	poorer	all
	(1)	(2)	(3)
ecoprc	-0.238	-1.280***	-1.280***
	(0.622)	(0.340)	(0.442)
richer			-0.920
			(0.747)
ecoprc:richer			1.042
			(0.669)
Constant	1.857***	2.777***	2.777***
	(0.685)	(0.387)	(0.502)
Observations	286	374	660
R^2	0.001	0.037	0.015
Note:	*p<0.1; **p<0.05; ***p<0.01		

Code:

```
reg.inter = lm(ecolbs ~ ecoprc+richer+ecoprc:richer, data=apple)
stargazer(reg.r, reg.p, reg.inter,type = "html", column.labels=c("richer","poorer","all"), omit.stat = c( "f","adj.rsq","ser"))
```

- On the right are regressions were we also control for the confounder `regprc` (price of regular apples). Now richer households react more strongly to higher prices. But again the difference is not significant, see regression (4).
- Only the 4th regression that also adds the interaction effect between `regprc` and `richer` gets now equivalent coefficients than the separate regressions for richer and poorer households.
- The 3rd regression is now more restrictive, since it is assumed that the effect of regular prices is the same for richer and poorer households. That also changes the estimated coefficients for the prices of ecological apples.
- The last column (5) uses all observations without any interaction terms. The estimated coefficient `-2.93` for `ecoprc` in this simple specification is a mix of the effect for richer and poorer households. Often one just runs such simple regressions while keeping in mind that the estimated effect is likely some average of heterogeneous effects from different subgroups.

	Dependent variable:				
	ecolbs				
	poorer	richer	all	all	all
	(1)	(2)	(3)	(4)	(5)
ecoprc	-2.775***	-3.216***	-3.399***	-2.775***	-2.926***
	(0.596)	(1.118)	(0.657)	(0.771)	(0.588)
regprc	2.160***	4.385***	3.062***	2.160**	3.029***
	(0.711)	(1.377)	(0.710)	(0.920)	(0.711)
richer			-0.945	-1.260*	
			(0.737)	(0.764)	
ecoprc:richer			1.082	-0.441	
			(0.660)	(1.189)	
regprc:richer				2.225	
				(1.445)	
Constant	2.482***	1.223*	2.359***	2.482***	1.965***
	(0.394)	(0.703)	(0.504)	(0.510)	(0.380)
Observations	374	286	660	660	660
R^2	0.060	0.035	0.042	0.046	0.036
Note:			*p<0.1; **p<0.05; ***p<0.01		

References

- Chou, Shin-Yi, Jin-Tan Liu, Michael Grossman, and Ted Joyce. 2010. "Parental Education and Child Health: Evidence from a Natural Experiment in Taiwan." American Economic Journal: Applied Economics, 2 (1): 33-61.
- Meghir, Costas, Mårten Palme, and Emilia Simeonova. 2018. "Education and Mortality: Evidence from a Social Experiment." American Economic Journal: Applied Economics, 10 (2): 234-56.