Fiscal Policy and Inequality
8. Introduction to Regression

Elliott Ash (ashe@ethz.ch)

ETH Zurich

October 8, 2018

Motivation

- RCTs solve the selection problem
  - But with most datasets and research questions, it is not possible to run a controlled experiment
  - Have to rely on observational data

Causality without experiments

- The **identification strategy** or **empirical strategy** is the approach used with observational data (i.e. data not generated by a randomized trial) to approximate a real experiment:
  - Selection based on observables
  - Differences-in-differences
  - Instrumental variables
  - Regression discontinuity design
  - Synthetic control
  - Bunching

Selection on observables

- We usually do not have a controlled experiment:
  - but maybe the treated group and the non-treated group differ only by a set of observable characteristics.
- This is the Conditional Independence Assumption (CIA) assumption:
  - also called" selection on observables"
  - justifies causal interpretation of regression estimates

CIA Example

- Effect of going to school $D_i \in \{0, 1\}$ on lifetime income $Y_i \geq 0$.
  - Potential outcomes Y_{0i}, Y_{1i}
- Recall that the difference in observed outcomes is

$$\mathbb{E}[Y_{1i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 0]$$

$$= \underbrace{\mathbb{E}[Y_{1i} - Y_{0i}|D_i = 1]}_{\text{ATT}} + \underbrace{\mathbb{E}[Y_{0i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 0]}_{\text{Selection Bias}}$$

Observable characteristics

- Say that we observe an IQ test, X_i, for each individual.
- The diference in outcomes, conditional on characteristics, is

$$\mathbb{E}[Y_{1i}|X_i, D_i = 1] - \mathbb{E}[Y_{0i}|X_i, D_i = 0]$$

$$= \underbrace{\mathbb{E}[Y_{1i} - Y_{0i}|X_i, D_i = 1]}_{\text{ATT}} + \underbrace{\mathbb{E}[Y_{0i}|X_i, D_i = 1] - \mathbb{E}[Y_{0i}|X_i, D_i = 0]}_{\text{Selection Bias}}$$

- The Conditional Independence Assumption (CIA) holds when

$$\mathbb{E}[Y_{0i}|X_i, D_i = 1] = \mathbb{E}[Y_{0i}|X_i, D_i = 0]$$

that is, selection bias is zero conditional on observables.

When is selection problem relevant?

- Three possible types of factors that affect the outcome variable:
  1. observable factors
     - not a problem
  2. unobservable factors not correlated with treatment
     - also not a problem
  3. unobservable factors correlated with treatment
     - **this is the problem**
- Questions:
  - what might drive selection in the education/income example?
  - why is this not a problem in an RCT?

Introduction to Regression

- How does schooling affect income?
- Assume a linear model

$$Y_i = \alpha + \beta s_i + \epsilon_i$$

- Y_i is income as a function of s_i, years of education
- α, the "intercept" or "constant", gives the expected income with no schooling ($s_i = 0$)
    - assume $\alpha = 0$ going forward.
- ϵ_i includes all other factors affecting income besides schooling, including randomness
- β is the slope parameter summarizing how wages vary with schooling.

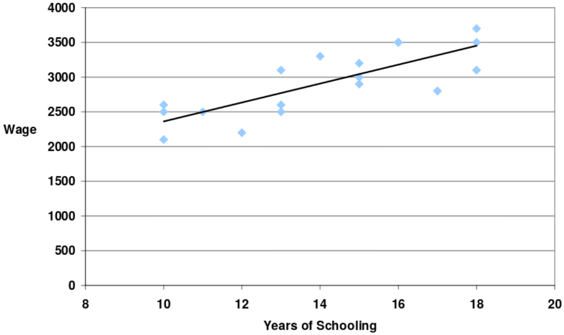
OLS Estimator

$$Y_i = \alpha + \beta s_i + \epsilon_i$$

- The Ordinary Least Squares (OLS) Estimator is the workhorse of applied microeconometrics.
- Assume that s_i is de-meaned and there are n observations. Then the OLS estimator is given by

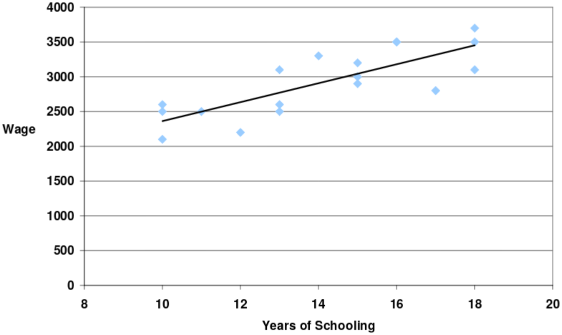
$$\hat{\beta} = \frac{\sum_{i=1}^{n} s_i Y_i}{\sum_{i=1}^{n} s_i^2} = \frac{\text{Cov}[Y_i, s_i]}{\text{Var}[s_i]}$$

Interpreting OLS Coefficients



- The OLS estimate for β, denoted by $\hat{\beta}$, gives the predicted change in the outcome variable Y_i in response to increasing the explanatory variable s_i by 1.
  - In this case, the average increase in income for taking one more year of school.

OLS for prediction



- Using the estimated constant $\hat{\alpha}$ and estimated slope coefficient $\hat{\beta}$, we obtain a predicted income $\hat{Y}_i$ for any level of schooling s_i:

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} s_i$$

OLS in Python

```
# open data set as pandas dataframe
import pandas as pd
df = pd.read_csv('state-tax-govt-data.csv')

# Run OLS for effect of estate tax on GSP
import statsmodels.formula.api as smf
model = smf.ols('gsp_q ~ death_and_gift_tax',
                data=df)
results = model.fit()
results.params # contains estimated coefficients (alpha and beta)
```

Statistical Significance

- The value for β is interesting because it provides a prediction for the effect of the explanatory variable on the outcome.
  - But if this prediction is very noisy, then it might not be useful for policy analysis.
- The second half of OLS regression is determining statistical significance.
  - This is generally achieved by computing a **standard error** for each coefficient, and then using the standard error to compute a *p*-**value** for statistical significance.

Residuals

- The **residuals** or **errors** from an OLS regression are defined as

$$\begin{aligned}\tilde{\epsilon}_i &= Y_i - \hat{Y}_i \\ &= Y_i - \hat{\alpha} - \hat{\beta} s_i\end{aligned}$$

  - In statsmodels, provided by **`results.resid`**

```
# histogram of residuals
results.resid.hist()
```

Standard Errors

- The **standard error** for the OLS estimate $\hat{\beta}$ is

$$\hat{\sigma}_\beta = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\tilde{\epsilon}_i^2},$$

  the square root of the average of the squared residuals.
  - In statsmodels, contained in `results.bse`.
  - This standard error provides information about the precision of the estimate: a lower standard error is a more precise estimate.
  - On regression tables, usually reported in parentheses right beneath the point estimate.

t-statistics and p-values

- A rule of thumb for statistical significance is to compute the t-statistic:
$$t = \frac{\hat{\beta}}{\hat{\sigma}_\beta}$$
- $t > 2$: there is a statistically significant positive effect
  - $t < 2$: there is a statistically significant negative effect
  - $t \in [-2, 2]$, no effect
- A high t (in absolute value) is associated with a small p-value (e.g., $t = 1.96 \rightarrow p = .05$).
  - Small p-values are often indicated on regression tables with stars to indicate statistical significance.
- Statistical significance $\neq$ economic significance.

Multivariate Regression

- Assume we have n observations and k explanatory variables.
- Let Y be the $n \times 1$ vector for the outcome variable (also called dependent variable or label).
- Let X be the $n \times k$ matrix of explanatory variables (also called independent variables or predictors)
- The $k \times 1$ vector of OLS coefficients (one for reach explanatory variable) is

$$\hat{\beta} = (X'X)^{-1}X'Y$$

with standard errors given by the diagonal entries of

$$\hat{\sigma}\sqrt{(X'X)^{-1}}$$

Multivariate Regression: Python Code

```
model2 = smf.ols('pop_annual ~ alcoholic_beverage_tax + tobacco_tax',
                 data=df)
results2 = model2.fit()
results2.summary()
```

OLS Estimator is unbiased under exogeneity (1)

- Take the OLS estimator

$$\hat{\beta} = \frac{\sum_{i=1}^{n} s_i Y_i}{\sum_{i=1}^{n} s_i^2}$$

and plug in the equation definition for Y_i (setting $\alpha = 0$)

$$\begin{aligned}
\hat{\beta} &= \frac{\sum_{i=1}^{n} s_i(\beta s_i + \epsilon_i)}{\sum_{i=1}^{n} s_i^2} \\
&= (\frac{\sum_{i=1}^{n} s_i^2}{\sum_{i=1}^{n} s_i^2})\beta + \frac{\sum_{i=1}^{n} s_i(\epsilon_i)}{\sum_{i=1}^{n} s_i^2} \\
&= \beta + \frac{\sum_{i=1}^{n} s_i \epsilon_i}{\sum_{i=1}^{n} s_i^2}
\end{aligned}$$

OLS Estimator is unbiased under exogeneity (2)

- Taking expectations gives

$$\begin{aligned}\mathbb{E}[\hat{\beta}] &= \beta + \mathbb{E}[\frac{\sum_{i=1}^{n} s_i \epsilon_i}{\sum_{i=1}^{n} s_i^2}] \\ &= \beta + \frac{\text{Cov}[s_i, \epsilon_i]}{\text{Var}[s_i]} \\ &= \beta\end{aligned}$$

  - The last line follows from the exogeneity assumption $\mathbb{E}[\epsilon_i | s_i] = 0$, which implies $\text{Cov}[s_i, \epsilon_i] = 0$.

Endogeneity

- When the conditional independence assumption is not satisfied, we say that "s is endogenous":
  - That is, an explanatory variable s_i is said to be **endogenous** if it is correlated with unobservable factors that are also correlated with the outcome variable.
  - this is why it is called "omitted variable bias"
- Since the error term ϵ_i includes all unobserved factors affecting the outcome, we can define endogeneity as correlation between an explanatory variable and the error term:

$$\text{Cov}[s_i, \epsilon_i] \neq 0$$

Omitted variable bias

- Assume that the "true" model states that income is affected by schooling and ability
$$Y_i = \beta s_i + \gamma a_i + \eta_i \tag{1}$$
where η_i is random (exogenous), but we cannot measure ability a_i.
- We can only estimate
$$Y_i = \beta s_i + \epsilon_i \tag{2}$$
- The OLS estimates for β from (1) and (2) will be different unless: (1) $\gamma = 0$, or (2) $\mathrm{Cov}(s_i, a_i) = 0$.

Understanding omitted variable bias

- Recall the formula for the OLS estimator

$$\hat{\beta} = \frac{\sum_{i=1}^{n} s_i Y_i}{\sum_{i=1}^{n} s_i^2}$$

and plug in the new equation definition for Y_i

$$\begin{aligned}\hat{\beta} &= \frac{\sum_{i=1}^{n} s_i(\beta s_i + \gamma a_i + \eta_i)}{\sum_{i=1}^{n} s_i^2} \\ &= \beta + \frac{\sum_{i=1}^{n} s_i(\gamma a_i)}{\sum_{i=1}^{n} s_i^2} + \frac{\sum_{i=1}^{n} s_i \epsilon_i}{\sum_{i=1}^{n} s_i^2}\end{aligned}$$

- Taking expectations gives

$$\mathbb{E}[\hat{\beta}] = \beta + \underbrace{\gamma \frac{\mathrm{Cov}[s_i, \epsilon_i]}{\mathrm{Var}[s_i]}}_{\text{Omitted variable bias}} + \underbrace{\frac{\mathrm{Cov}[s_i, \epsilon_i]}{\mathrm{Var}[s_i]}}_{=0 \text{ by assumption}}$$

What happens if we omit a variable

		Correlation of omitted variable with explanatory variable	
		Corr[s, a] > 0	Corr[s, a] < 0
Correlation of omitted variable with outcome	$\gamma > 0$	$\hat{\beta} > \beta$	$\hat{\beta} < \beta$
	$\gamma < 0$	$\hat{\beta} < \beta$	$\hat{\beta} > \beta$

- How does the example of ability/schooling/income fit in this table?

Is adding controls always a good idea?

- The short answer is no.
  - With a good identification strategy, you don't need controls.
  - "Bad controls" are variables that are jointly determined along with the outcome.
    - for example, controlling for occupation in the effect of education on income: education affects both occupation and income.
    - these variables could add bias to your estimates.

Russia Elections Paper: Regression Estimates

Table 1. Spillovers

	Vote share of		
Sample	United Russia	Just Russia	LDPR
Observers present	−0.130*** (0.013)	0.029*** (0.004)	0.027*** (0.003)
Observers present in a neighboring polling station	−0.052*** (0.014)	0.014*** (0.004)	0.022*** (0.004)
Constant	0.452*** (0.010)	0.125*** (0.003)	0.097*** (0.002)
Observations	3,164	3,164	3,164
r^2	0.03	0.02	0.03

SEs clustered by electoral district are in parentheses. *$P < 0.1$, **$P < 0.05$, ***$P < 0.01$.