Policy Analysis

2025/2026, Semester 1

Ralf A. Wilke
Copenhagen Business School

Readings:

Wooldridge, J. (2025), Introductory Econometrics, 8th Edition, South-Western, Chapter 3.7, 7.6, 13.2.

Potential Outcomes, Treatment Effects, Policy Analysis

- To estimate the casual effect of for example policy interventions
  - Does job training increase earnings?
  - Do more elective courses increase student outcomes?
- *w* is a binary policy indicator (=1: treated, =0: not treated).
- Potential outcomes: y(0) in absence of treatment, y(1) in presence of treatment
  - In reality we will only observe one of the outcomes.
- If the effect of the intervention is constant, say τ, we have for any observation *i*:

$$y_i(1) = \tau + y_i(0)$$

- More generally, if τ is not constant, we consider the average treatment effect:

$$\tau_{ATE} = E[y_i(1) - y_i(0)]$$

where the expectation is over the entire population.

- For the i'th observation, the observed outcome y_i can be written

$$y_i = (1 - w_i)y_i(0) + w_i y_i(1)$$

- Remark: If we regress y on w, the resulting OLS estimate is only unbiased for τ_{ATE} if we have random assignment of *w*, that is *w* is independent of [y(0),y(1)] conditional on ***x*** *(*if there are further **x** in the model*):*

$$y_i = \tau_{ATE} w_i (+\mathbf{x}_i \beta) + u_i$$

  The point is here that for all values of *w*, it must be unrelated with *u*, so *w* is not permitted to be related with y(0),y(1) (conditional on ***x***).

- Random assignment is rare in economic and business data but can be ensured by doing randomized control trials (RCT).
  - In presence of a RCT, the econometrics becomes easy!
  - By designing policy changes or interventions in a clever way it is easy to learn something about their effects.

$$y_i = \tau_{ATE} w_i + \mathbf{x}_i \beta + u_i$$

- Random assignment conditional on ***x*** is more likely to hold the more variables there are in the data.
  - The more we take out of *u*, the less likely it will contain something that is correlated with *w* conditional on ***x***.
- Random is also called unconfounded assignment or ignorable assignment.
- Unfortunately, in most empirical situation there is no random assignment as the decision to introduce an intervention is related to outcomes.
  - E.g. individuals with poor labour market performance receive treatment. Hopefully, this can be mitigated by using a rich conditioning set.

- We show how random assignment conditional on ***x*** leads to unbiased estimation.
- We use $y = (1-w)y(0) + wy(1)$ and let

$$E[y|w, x] = \alpha + \tau w + \mathbf{x}\gamma$$

- To make the point, we consider models with covariates centered about their means $\eta_j = E(x_j)$, where we consider two separate equations for the two outcomes:

$$\begin{aligned} y(0) &= \phi_0 + (\mathbf{x} - \boldsymbol{\eta})\boldsymbol{\gamma}_0 + u(0) \\ y(1) &= \phi_1 + (\mathbf{x} - \boldsymbol{\eta})\boldsymbol{\gamma}_1 + u(1) \end{aligned}$$

with $\phi_0 = E(y(0))$ and $\phi_1 = E(y(1))$.

- The treatment effect for observation *i* is

$$\begin{aligned} te_i &= y_i(1) - y_i(0) \\ &= (\phi_1 - \phi_0) + (\mathbf{x}_i - \boldsymbol{\eta})(\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_0) + [u_i(1) - u_i(0)] \end{aligned}$$

  which is a function of ***x*** and the error terms.
- The average treatment effect is

$$\begin{aligned} E(te_i) &= \tau_{ATE} \\ &= (\phi_1 - \phi_0) + E\{(\mathbf{x}_i - \boldsymbol{\eta})(\boldsymbol{\gamma}_1) - \boldsymbol{\gamma}_0) + [u_i(1) - u_i(0)]\} \\ &= (\phi_1 - \phi_0) + \mathbf{0}(\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_0) + 0 \\ &= \phi_1 - \phi_0 \end{aligned}$$

- The observed outcome $y_i = (1 - w_i)y_i(0) + w_i y_i(1)$ can be written as

$$\begin{aligned} y_i &= \phi_0 + \tau w_i + (\mathbf{x}_i - \boldsymbol{\eta})\boldsymbol{\gamma}_0 + w_i(\mathbf{x}_i - \boldsymbol{\eta})\boldsymbol{\delta} \\ &\quad + u(0) + w_i[u_i(1) - u_i(0)] \end{aligned}$$

where $\boldsymbol{\delta} = (\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_0)$.

- The realised (but unobserved error) is

$$u_i = u_i(0) + w_i[u_i(1) - u_i(0)]$$

- Random assignment or unconfoundedness implies:

$$\begin{aligned} E(u_i|w_i, \mathbf{x_i}) &= E[u_i(0)|w_i, \mathbf{x}_i] + w_i E\{[u_i(1) - u_i(0)]|w_i, \mathbf{x}_i\} \\ &= E[u_i(0)|\mathbf{x}_i] + w_i E\{[u_i(1) - u_i(0)]|\mathbf{x}_i\} \\ &= 0 \end{aligned}$$

This means the error is conditional zero mean if treatment assignment is random cond. on ***x***.

- In practice, regress:

$$y_i \text{ on } w_i, \mathbf{x}_i, w_i(\mathbf{x}_i - \bar{\mathbf{x}}_i)$$

The coefficient on w_i is $\hat{\tau}_{ATE}$.

- The $\hat{\tau}_{ATE}$ obtained is typically different from the one obtained from the restricted regression of

$$y_i \text{ on } w_i, x_i$$

- The latter approach is called restricted regression adjustment (RRA), while the unrestricted model is unrestricted regression adjustment (URA).
- Without showing this, $\hat{\tau}_{ATE}$ for the unrestricted model can be obtained from two separate regressions, which is typically more convenient:
  - Split the sample by $w_i = 0$ and $w_i = 1$. Run the regression $y_i \text{ on } \mathbf{x}_i$ for each sample.
  - Compute fitted values $\hat{y}_i^{(1)}$ and $\hat{y}_i^{(0)}$ and $\hat{\tau}_{ATE} = n^{-1} \sum_i [\hat{y}_i^{(1)} - \hat{y}_i^{(0)}]$.
- Example: tau_ATE.R

Policy Analysis with Pooled Cross Sections

- If cross sectional data is collected before and after an event, it can be used to determine the effect on economic outcomes.
- Example: Effect of Garbage Incinerator's Location on House Prices
  - New incinerator in North Andover, Massachusetts
  - Construction started in 1981, operation in 1985
  - Rumors that it will be built began in 1978
  - Data about prices of houses sold in 1978 and 1981 (KIELMC.dta), prices in 1978.
  - The hypothesis is that house prices near the incinerator will drop relative to prices of more distant houses.

- <u>Example</u>: cont.
  - A naïve approach to estimate the effect of the incinerator would be:

$$rprice = \gamma_0 + \gamma_1 nearinc + u$$

    where *nearinc* is 1 if the house is within 3 miles from the incinerator.
  - When we estimate it for the 1981 data:

$$\begin{aligned} \widehat{rprice} &= 101,307.5 - 30,688.27 nearinc \\ &\quad (3,093.0) \quad (5,827.71) \\ &\quad n = 142, \quad R^2 = 0.165 \end{aligned}$$

  - The intercept is the average selling price for homes not near the incinerator
  - Average selling price for home near the incinerator is $30,688 lower. Statistically significant.
  - Is this already evidence for an effect of the incinerator?

- Not really! When we repeat the analysis for the 1978 data, we obtain:

$$\begin{aligned} \widehat{rprice} &= 82,517.23 - 18,824.37 nearinc \\ & \quad (2,653.79) \quad (5,827.71) \\ & \quad n = 179, \quad R^2 = 0.082 \end{aligned}$$

- Home prices were already statistically significant lower around the future location of the incinerator before the planning process was started.
- How can we then assess whether there was a decrease in response to the construction of the incinerator?
- Simply take the difference in the two coefficients on *nearinc*:

$$\begin{aligned} \hat{\delta}_1 &= -30,688.27 - (-18,824.37) = -11,863.9 \\ &= (\overline{rprice}_{81,n} - \overline{rprice}_{81,f}) - (\overline{rprice}_{78,n} - \overline{rprice}_{78,f}) \end{aligned}$$

is an estimate of the effect of the incinerator on values of homes, where _*n* is near and _*f* is farther away from the site.

$$\hat{\delta}_1 = (\overline{rprice}_{81,n} - \overline{rprice}_{81,f}) - (\overline{rprice}_{78,n} - \overline{rprice}_{78,f})$$

- This is known as the difference-in-differences (DID) estimator.
- It is the difference over time in the differences of housing prices in the two locations.
- We can directly estimate $\hat{\delta}_1$ and its standard error by using regression analysis:

$$rprice = \beta_0 + \delta_0 y81 + \beta_1 nearinc + \delta_1 y81 * nearinc + u$$

- β_0 is the average home price not near the site in 1978.
- δ_0 captures changes in prices not near between 1978 and 1981
- β_1 measures the location effect that is not due to the incinerator.
- δ_1 measures the decline in home prices due to the new incinerator, provided we assume everything else equal.

- Example (cont.) Estimation results for three sets of regressors:

Dependent Variable: rprice			
Independent Variable	*(1)*	*(2)*	*(3)*
Constant	*82,517.23* *(2,726.91)*	*89,116.54* *(2,406.05)*	13,807.67 (11,166.59)
y81	*18,790.29* *(4,050.07)*	21,321.04 (3,443.63)	13,928.48 (2,798.75)
nearinc	-18,824.37 (4,875.32)	*9,397.94* *(4,812.22)*	3,780.34 (4,453.42)
*y81*nearinc*	-11,863.90 (7,456.65)	-21,920.27 (6,359.75)	-14,177.93 (4,987.27)
Other controls	No	Age, age^2	Full Set
Observations *R-squared*	321 0.174	321 0.414	321 0.660

  - DID estimator only marginally significant in (1), magnitude depends on model specification.

- In the previous example it would have been better to use the log of price as dependent variable. Then the DID estimator tells us the percentage change in prices in response to the new site.
- A DID estimator can be applied if the data arises from a natural experiment (or quasi experiment):
  - An exogenous event (e.g. change of policy) changes the environment or labour market outcome.
  - A natural experiment has always a control group (C) and always a treatment group (T).
  - We need (at least) two periods of data, one before (T0) and one after the policy change (T1).
  - Denote d2 as the dummy for the second period and dT as the dummy for the treatment group, then the following regression directly estimates the DID treatment effect:

$$y = \beta_0 + \delta_0 d2 + \beta_1 dT + \delta_1 d2 * dT + other\ factors$$

  - Also called the average treatment effect as it measures the effect on the average outcome of y.

Summary

- Multiple regression analysis is frequently used as a tool to analyse the effect of a policy change or intervention on an outcome.
- Under some restrictions it is possible to estimate the average treatment effect.
- Knowledge about how to estimate treatment effects is also useful for designing the treatment or change in a way that it is possible to empirically assess its implications.