Course Notes for *Microeconometrics*

Fei Li[*]

[*] Email: fei.li.best@gmail.com.

1 Causality

1.1 Potential Outcome Framework (POF)

- **Population**
- **Observables**: for each unit i, observe the realized value of the following variables: a treatment, denoted by D (e.g. $D \in \{0, 1\}$, and an outcome, denoted by Y.
- **Potential outcomes**: The hypothetical outcome unit i would experience were they given feasible treatment value $\tilde{D}$ is denoted by $Y_i(\tilde{D})$. $Y_i(1)$ or Y_{1i} is i's potential outcome under treatment. $Y_i(0)$ or Y_{0i} is i's potential outcome under no treatment.
- **Individual-level causal effect**: $\Delta_i = Y_i(1) - Y_i(0)$. It is logically impossible to observe the individual-level effect (Fundamental Problem of Causal Inference (FPCI) [Hol86]
- **Average Treatment Effect (ATE)**: expectation of individual-level causal effect over the entire population.

$$\delta^{ATE} = E[\Delta_i] = E[Y_i(1) - Y_i(0)] = E[Y_i(1)] - E[Y_i(0)].$$

For $D \in \{0, 1\}$, this is

$$\begin{aligned} &E[Y_i(1)|D_i = 1] \cdot P(D_i = 1) + E[Y_i(1)|D_i = 0] \cdot P(D_i = 0) \\ -&\big(E[Y_i(0)|D_i = 0] \cdot P(D_i = 0) + E[Y_i(0)|D_i = 1] \cdot P(D_i = 1)\big), \end{aligned}$$

where we have used the law of iterated expectations:

$$E[Y] = \sum_{X=x} E[Y|X = x] \cdot P(X = x).$$

This is the way we calculate δ^{ATE}, since we usually have the following kind of data:

Group	Sample Size	Mean Health Status	Std. Error
Hospital	7,774	3.21	0.014
No Hospital	90,049	3.93	0.003

In this case, $E[Y_i(1)|D_i = 1] = 3.21$, $P(D_i = 1) = 7,774/97,823$, and $E[Y_i(0)|D_i = 0] = 3.93$, $P(D_i = 1) = 90,049/97,823$. We don't know the other two expectations, $E[Y_i(1)|D_i = 0]$ and $E[Y_i(0)|D_i = 1]$, so the naive difference $(3.21 - 3.93)$ is not the δ^{ATE} that we'd like to know, unless

$$E[Y_i(1)|D_i = 1] = E[Y_i(1)|D_i = 0] \quad \text{and} \quad E[Y_i(0)|D_i = 0] = E[Y_i(0)|D_i = 1].$$

- Average Treatment effect on the Treated (ATT) and on the Untreated (ATU):

$$\delta^{ATT} = E[\Delta_i|D_i = 1] = E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 1]$$

$$\delta^{ATU} = E[\Delta_i|D_i = 0] = E[Y_i(1)|D_i = 0] - E[Y_i(0)|D_i = 0]$$

The Naive Estimator (NE) can be decomposed as:

$$\begin{aligned}\text{NE} &= E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 0] \\ &= E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 1] + E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0] \\ &= \delta^{ATT} + E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0] \\ &= \delta^{ATT} + \text{selection bias.}\end{aligned}$$

1.2 Randomization

- Randomly divide the population into two groups, T and C. Group T receives treatment ($D_i = 1$), and group C receives no treatment ($D_i = 0$).
- C and T differ only in the treatment, but no other things. In particular, $E[Y_i(0)]$ is the same for both C and T, and $E[Y_i(1)]$ is the same for both C and T.
- Symbolically,

$$E[Y_i(0)|i \in C] = E[Y_i(0)|i \in T] = E[Y_i(0)],$$

$$E[Y_i(1)|i \in T] = E[Y_i(1)|i \in C] = E[Y_i(1)].$$

- This implies that

$$\begin{aligned}\delta^{ATE} \equiv E[Y_i(1)] - E[Y_i(0)] &= E[Y_i(1)|i \in T] - E[Y_i(0)|i \in C] \\ &= E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 0] \\ &= \text{NE}.\end{aligned}$$

- Similarly, we have $\delta^{ATT} = \delta^{ATE}$.

2 Linear Regressions

2.1 Data Structures

1. Cross-sectional data
2. Time series data
3. Pooled cross sections:
    - Samples in two or more time period. The samples in each time period are not related.
    - Often used to **evaluate policy changes** (e.g. by means of Differences-in-Differences). Example: evaluate effect of change in property taxes on house prices. Compare random sample of 250 house prices in 1993 and random sample of 270 house prices in 1995.
4. Panel data:
    - Follow the same unit in multiple time periods (Same cross-sectional units followed over time)
    - Panel data can be used to **account for time-invariant unobservables** (e.g. by means of Fixed Effects), aiding causal inference. Panel data can also be used to model lagged responses.

2.2 Conditional Expectation Function (CEF)

Definition: $E[y_i|X_i]$. When y is continuous, it is $\int t \cdot f_y(t|X_i = x)dt$.
Properties of CEF:

- **Law of Iterated Expectations (LIE)**:

  $$E[y_i] = E\{E[y_i|X_i]\}.$$

  Example: Wage-schooling: gender is discrete X, then the average earnings in a population of men and women is the average for men times the population proportion of men plus the average for women times the population proportion of women. $E(y) = E(y_i|X_i = \text{male}) \cdot p_1 + E(y_i|X_i = \text{female}) \cdot p_2$, where $p_1 + p_2 = 1$.

- **CEF Decomposition Property**: from LIE, we can write

  $$y_i = E[y_i|X_i] + \varepsilon_i,$$

  where ε_i is mean independent of X_i, i.e. $E[\varepsilon_i|X_i] = 0$. In words, any random variable y_i can be decomposed into a piece that is explained by X_i and a piece left over which is uncorrelated with any function of X_i.

- **CEF Prediction Property**: The CEF solves:

$$\underset{m}{\operatorname{argmin}} E\left\{[y_i - m(X_i)]^2\right\},$$

so it is the mean squared error (MMSE) predictor of y_i given X_i. In words, CEF is the best predictor of y_i given X_i under square loss.

- **CEF ANOVA Theorem**:

$$V(y_i) = V\{E[y_i|X_i]\} + E\{V[y_i|X_i]\}.$$

2.3 Population Linear Regression Model

Population Linear Regression Model:

$$y_i = X_i'\beta + \varepsilon_i,$$

where y_i and ε_i are 1×1, X_i and β are $K \times 1$, and the model holds for every i in the population.

β can be defined by minimizing mean squared error (MMSE):

$$\beta = \underset{b}{\operatorname{argmin}} E[(y_i - X_i'b)^2].$$

First order condition:

$$E[X_i(y_i - X_i'b)] = 0$$

Solution:

$$\beta = E[X_i X_i']^{-1} E[X_i y_i].$$

Note that the expectation is taken over all i's.

Let's verify the formula for simple linear regression with a constant. In this case $y = \beta_0 + \beta_1 x + \epsilon$, so $X = \begin{pmatrix} 1 \\ x \end{pmatrix}$ and $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$. We omit the subscript i. We have

$$XX' = \begin{pmatrix} 1 \\ x \end{pmatrix} \begin{pmatrix} 1 & x \end{pmatrix} = \begin{pmatrix} 1 & x \\ x & x^2 \end{pmatrix},$$

so that

$$E[XX'] = \begin{pmatrix} 1 & E[x] \\ E[x] & E[x^2] \end{pmatrix} \Longrightarrow E[XX']^{-1} = \frac{1}{E[x^2] - E[x]^2} \begin{pmatrix} E[x^2] & -E[x] \\ -E[x] & 1 \end{pmatrix}$$

$$= \frac{1}{V(x)} \begin{pmatrix} E[x^2] & -E[x] \\ -E[x] & 1 \end{pmatrix}.$$

$Xy = \begin{pmatrix} 1 \\ x \end{pmatrix} \cdot y = \begin{pmatrix} y \\ xy \end{pmatrix}$. Thus we have

$$\beta = \frac{1}{V(x)} \begin{pmatrix} E[x^2] & -E[x] \\ -E[x] & 1 \end{pmatrix} \begin{pmatrix} E[y] \\ E[xy] \end{pmatrix} = \frac{1}{V(x)} \begin{pmatrix} E[x^2] \cdot E[y] - E[x] \cdot E[xy] \\ -E[x] \cdot E[y] + E[xy] \end{pmatrix}.$$

So we indeed have $\beta_1 = (E[xy] - E[x] \cdot E[y])/V(x) = Cov(y, x)/V(x)$.

We can actually derive β_0 and β_1 in a simple way. The first order conditions are

$$\begin{cases} -2E[y - \beta_0 - \beta_1 x] = 0 \\ -2E[x(y - \beta_0 - \beta_1 x)] = 0 \end{cases}$$

From the first we get $\beta_0 = E[y] - \beta_1 E[x]$. Substitute β_0 into the second we get $\beta_1 = Cov(y, x)/V(x)$.

An even more straightforward way is this: Since $Cov(x, \epsilon) = 0$, we have

$$Cov(y, x) = Cov(\beta_0 + \beta_1 x + \epsilon, x) = \beta_1 Cov(x, x) + Cov(\epsilon, x) = \beta_1 Cov(x, x) \Longrightarrow \beta_1 = Cov(y, x)/V(x).$$

In the multivariate case, we have $\beta_k = Cov(y, \tilde{x}_k)/V(\tilde{x}_k)$, where $\tilde{x_k}$ is the residual from a regression of x_k on all other covariates. In words, each slope coefficient in a multivariate regression is the bivariate slope coefficient for the corresponding regressor after partialing out all the other covariates (Frisch-Waugh Theorem).

This implies that the coefficient of any regressor, say x_1, in a multiple regression model can be obtained in two steps:

1. Regress x_1 on all other regressors x_2, x_3, x_4, etc.
2. Simple linear regression: regress y on the residual from the first regression.

This works because:

1. The residuals from the first regression is the part of x_1 that is uncorrelated with x_2, x_3, x_4 etc.
2. The slope coefficient of the second regression is the isolated effect of x_1 on y, after the effect of x_2, x_3, x_4 etc. on x_1 has been partialed or netted out.

2.4 Linking Mean Linear Regression and CEF

1. Regression Justification I: Linear CEF Theorem

   Theorem 2.1. Suppose the CEF is linear, that is $E[y_i|X_i] = m(X_i) = X_i'b$ for some parameters vector b. Then the population regression function (PRF), $X_i'\beta$, is the CEF: $E[y_i|X_i] = X_i'\beta$.

2. Regression Justification II: Best Linear Predictor Theorem

   Theorem 2.2 (BLP Theorem)**.** The function $X_i'\beta$ is the best linear predictor of y_i given X_i in a MMSE sense.

3. Regression Justification III: Regression CEF Theorem

   Theorem 2.3. The PRF, $X_i'\beta$, provides the MMSE linear approximation to the CEF, $E[y_i|X_i]$:

$$\beta = \underset{b}{\operatorname{argmin}}\, E\left\{(E[y_i|X_i] - X_i'b)^2\right\}$$

The last theorem implies that regression coefficients can be obtained using $E[y_i|X_i]$ as a dependent variable instead of y_i. This is *weighted regression*:

$$\begin{aligned}\beta &= \underset{b}{\operatorname{argmin}}\, E\left\{(E[y_i|X_i] - X_i'b)^2\right\} \\ &= \underset{b}{\operatorname{argmin}} \sum_{v} (E[y_i|X_i = v] - v'b)^2 f_X(v).\end{aligned}$$

2.5 Regression Specifications

- **Level-Level Benchmark**

$$\text{wage} = \beta_0 + \beta_1 \text{educ} + \cdots + \varepsilon.$$

  Two characteristics:

  - *Homogeneity*: Does not allow for differential returns to education across different groups, i.e. β_1 is the same for everybody.
  - *Constant partial effect*: the effect of an additional year of education, β_1, is constant across levels of education.

- **Log-Level Specification**

$$\log(\text{wage}) = \beta_0 + \beta_1 \text{educ} + \cdots + \varepsilon.$$

In this case,

$$\beta_1 = \frac{\Delta \log(\text{wage})}{\Delta \text{educ}} \approx \frac{\Delta \text{wage}/\text{wage}}{\Delta \text{educ}}.$$

β_1 has a **constant percentage** interpretation: wage is increased by $100\beta_1$ percent if education is increased by one year. This allows for **increasing returns** to education.

- **Log-Log Specification**

  $$\log(\text{salary}) = \beta_0 + \beta_1 \log(\text{sales}) + \cdots + \varepsilon.$$

  β_1 has an **elasticity** interpretation:

  $$\beta_1 = \frac{\Delta \log(\text{salary})}{\Delta \log(\text{sales})} \approx \frac{\Delta \text{salary}/\text{salary}}{\Delta \text{sales}/\text{sales}},$$

  or the precentage change of wage if sales increase by one percent.

- **Non-Constant Marginal Effect**
  Consider the following model relating family consumption and income:

  $$\text{cons} = \beta_0 + \beta_1 \text{inc} + \beta_2 \text{inc}^2 + \cdots + \varepsilon.$$

  The marginal effect of income on consumption is $\beta_1 + 2\beta_2\text{inc}$. We expect $\beta_0 > 0$ and $\beta_1 < 0$, so that an increase in income should increase consumption at a **decreasing rate**.

- **Dummy Variables**
  Consider the following model relating wage with gender and education:

  $$\text{wage} = \beta_0 + \delta_0 \text{female} + \beta_1 \text{educ} + \cdots + \varepsilon.$$

  In this case, δ_0 is the difference in mean wage between men and women with the same level of education:

  $$\begin{aligned} \delta_0 &= [\beta_0 + \delta_0 + \beta_1 \text{educ}] - [\beta_0 + \beta_1 \text{educ}] \\ &= E[\text{wage}|\text{female} = 1, \text{educ}] - E[\text{wage}|\text{female} = 0, \text{educ}]. \end{aligned}$$

  Note that δ_0 is an intercept shift. The model does not allow the wage difference between the two groups to depend on education.

- **Dummy Variables With Interaction Effects**
  Consider the following model:

  $$\begin{aligned} \text{wage} &= \beta_0 + \delta_0 \text{female} + \beta_1 \text{educ} + \delta_1 \text{female} \cdot \text{educ} + \cdots + \varepsilon \\ &= (\beta_0 + \delta_0 \text{female}) + (\beta_1 + \delta_1 \text{female}) \cdot \text{educ} + \cdots + \varepsilon. \end{aligned}$$

  The four parameters have the following interpretation:

  1. β_0 is the intercept for men.
  2. β_1 is the slope for men.
  3. $\beta_0 + \delta_0$ is the intercept for women.
  4. $\beta_1 + \delta_1$ is the slope for women.

2.6 Sample Estimation

Recall that X_i is $K \times 1$. We let X denote the data matrix whose row is X_i'. The problem now is to minimize SSR across sample units:

$$\min_{\beta} \sum_{i=1}^{N} (y_i - X_i'\beta)^2 \quad \text{i.e.} \quad \min_{\beta} \| y - X\beta \|^2.$$

The solution is

$$\hat{\beta}^{OLS} = \left[\sum_{i=1}^{N} X_i X_i' \right]^{-1} \left[\sum_{i=1}^{N} X_i y_i \right] = (X'X)^{-1}(X'y).$$

The k-th element can be obtained from the Frisch-Waugh Theorem as

$$\hat{\beta}_k^{OLS} = \frac{\sum_{i=1}^{N} y_i \hat{r}_{ki}}{\sum_{i=1}^{N} \hat{r}_{ki}^2},$$

where $\hat{r}_{ki} = x_{ki} - (\sum_{j \neq k} \hat{\alpha}_j x_{ji})$.

Some properties:

1. From the first order condition of the intercept β_0, we have

$$\sum_{i=1}^{N} \hat{\varepsilon}_i = 0,$$

where $\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_K x_{Ki})$.

2. From the first order condition on all β's, we have $X'(y - X\beta) = 0$, i.e.

$$\sum_{i=1}^{N} x_{ki} \hat{\varepsilon}_i = 0 \quad \text{for } k = 1, 2, \ldots, K$$

(The first one is a special case of this one, where $x_{1i} = 1$ for all unit i)

3. Sample averages of y and x lie on th regression line:

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + \cdots + \hat{\beta}_K \bar{x}_K.$$

This is derived by dividing the equation in the first property by N.

A common measure of goodness-of-fit is R squared. For each unit i, the difference between predicted value $\hat{y}_i$ and the data y_i is the error $\varepsilon_i = y_i - \hat{y}_i$. The SSR is $\sum_{i=1}^{N} \varepsilon_i^2$. This could be a measure of how well the model fits the data (the smaller the better), but if we want to compare between different models, we need to standardize it. To do so, we use $\bar{y}$ as a benchmark. If

$\varepsilon_i^2 = (y_i - \hat{y}_i)^2 > (y_i - \bar{y})^2$, then the prediction $\hat{y}_i$ is worse than using the average $\bar{y}$ instead, so we expect $(y_i - \hat{y}_i)^2 \leq (y_i - \bar{y})^2$. Thus we can use

$$R^2 = 1 - \frac{(y_1 - \hat{y}_1)^2 + \cdots + (y_N - \hat{y}_N)^2}{(y_1 - \bar{y})^2 + \cdots + (y_N - \bar{y})^2} = 1 - \frac{\sum_{i=1}^N \varepsilon_i^2}{\sum_{i=1}^N (y_i - \bar{y})^2}$$

as a measure of goodness-to-fit.

2.7 Sample Estimation: Gauss-Markov Theorem

Assumptions:

- **(A1) Linear in parameters**.

- **(A2) Random sampling**.

- **(A3) No perfect collinearity**: none of the variables is constant, and there are no exact linear relationships among the variables. In other words, the data matrix X should be full rank ($\text{rank}(X) = K$). This insures that $X'X$ is invertible, so the formula for β would work. Multicollinearity (as long as it is not "perfect") can be present resulting in a less efficient, but still unbiased estimate. The estimates will be less precise and highly sensitive to particular sets of data.

  Recall X is $N \times K$ and $X'X$ is $K \times K$. We give a proof that $\text{rank}(X) = \text{rank}(X'X)$. First, from the fundamental theorem of linear algebra

  $$K = \dim \text{null} X + \dim \text{range} X = \dim \text{null} X + \text{rank}(X)$$

  and

  $$K = \dim \text{null} X'X + \dim \text{range} X'X = \dim \text{null} X'X + \text{rank}(X'X).$$

  We prove $\text{null} X = \text{null} X'X$. If $v \in \text{null} X$, so that $Xv = 0$, then certainly $X'Xv = 0$, so that $v \in \text{null} X'X$. On the other hand, if $X'Xv = 0$, then

  $$\begin{aligned} 0 &= v^T X'Xv \\ &= \|Xv\|^2 \end{aligned}$$

  so that $Xv = 0$. The proof is complete.

- **(A4) Zero conditional mean**: $E(\varepsilon_i | X_i) = 0$.
  Note that *mean independence implies zero covariance, but not the other way around*:

  $$E[X|Y] = E[X] \Rightarrow E[XY] = E[X]E[Y]$$

(e.g. $Cov(X, Y) = E[XY] - E[X]E[Y] = 0$). This can be proved by the law of iterated expectation:

$$E[XY] = E[E[XY|Y]] = E[E[X|Y] \cdot Y] = E[E[X] \cdot Y] = E[X] \cdot E[Y],$$

where we used LIE in the first equality; the second equality holds because $E[X \cdot y|Y = y] = E[X|Y = y] \cdot y$ for every y.

- **(A5) Homoskedasticity**: $V(\varepsilon_i|X_i) = \sigma^2$.

Theorem 2.4 (Unbiasedness of OLS estimator)**.** Under assumptions **(A1) – (A4)**, the OLS estimator is unbiased:

$$E(\hat{\beta}_k^{OLS}) = \beta_k \quad \text{for } k = 0, 1, \ldots, K.$$

Under assumptions **(A1) – (A5)**, the OLS estimator is the best linear unbiased estimator (BLUE):

$$V(\hat{\beta}_k^{OLS}) \leq V(\tilde{\beta}_k) \quad \text{for } k = 0, 1, \ldots, K$$

for all $\tilde{\beta}_k = \sum_{i=1}^N w_{ki} \cdot y_i$ for which $E(\tilde{\beta}_k) = \beta_k$.

The variance of OLS estimator under **(A1)–(A5)** is

$$V(\hat{\beta}_k^{OLS}) = \begin{cases} \frac{\sigma^2}{\sum_{i=1}^N (x_i - \bar{x})^2} & \text{(bivariate case)} \\ \\ \frac{\sigma^2}{SST_k(1-R_k^2)} & k = 1, \ldots, K \text{ (multivariate case)} \end{cases}$$

where $\sigma^2 = V(\varepsilon_i|X_i)$, $SST_k = \sum_{i=1}^N (x_{ik} - \bar{x}_k)^2$, and R_k^2 is the R^2 of a mean linear regression of x_k on all other variables in X.

Several points:

- We want R_k^2 to be small, i.e. there is little correlation between independent variables, so that the variance of $\hat{\beta}_k^{OLS}$ could be small.
- We can see that increase N will increase SST_k, which decreases $V(\hat{\beta}_k^{OLS})$, so large N improves precision.

To estimate the variance, we estimate σ^2 as $\hat{\sigma}^2 = (\sum_{i=1}^N \varepsilon_i^2)/(N - K - 1)$. It can be shown that under (A1)–(A5) it is unbiased: $E(\hat{\sigma}^2) = \sigma^2$. The standard errors for regression coefficients are thus

$$SE(\hat{\beta}_k^{OLS}) = \sqrt{\hat{V}(\hat{\beta}_k^{OLS})} = \begin{cases} \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^N (x_i - \bar{x})^2}} & k = 1 \\ \\ \sqrt{\frac{\hat{\sigma}^2}{SST_k(1-R_k^2)}} & k = 1, \ldots, K \end{cases}$$

We now talk about asymptotic OLS inference.

Theorem 2.5. For reference, we list several results below.

- Consistency of OLS estimator: Under **(A1) – (A4)**, the OLS estimator is consistent:

$$\hat{\beta}_k^{OLS} \xrightarrow{p} \beta_k \quad \text{for } k = 0, 1, \ldots, K$$

  as $N \to \infty$. In fact, for consistency, (A4) can be replaced by weaker condition (A4'): $E(\varepsilon) = 0$ and $E(x_k \varepsilon) = Cov(x_k, \varepsilon) = 0$ for $k = 1, \ldots, K$.

- Asymptotic normality of OLS estimator: Under **(A1) – (A5)**,

$$\frac{\hat{\beta}_k - \beta_k}{SE(\hat{\beta}_k)} \rightsquigarrow N(0, 1)$$

  as $N \to \infty$.

- Asymptotic efficiency of OLS estimator: Under **(A1) – (A5)**, the OLS estimator has the smallest asymptotic variances:

$$AVar\sqrt{N}(\hat{\beta}_k^{OLS} - \beta_k) \leq AVar\sqrt{N}(\tilde{\beta}_k - \beta_k) \quad \text{for } k = 0, \ldots, K,$$

  where $\tilde{\beta}_k$ solves equations of the form

$$\sum_{i=1}^{N} g_k(X_i)(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_{i1} - \cdots - \tilde{\beta}_k x_{iK}),$$

  where $g_k(X_i)$ is any function of all regressors for observation i. $\hat{\beta}_k^{OLS}$ obtains for $g_0(X_i) = 1$ and $g_k(X_i) = x_{ik}$.

Several points:

- Consistency is a minimal requirement for sensible estimators. If biased, then better be at least consistent.

- Failure of (A5) implies that the earlier variance formulas for OLS estimators are no longer valid, affecting efficiency of OLS as well as testing. It can be addressed using robust standard errors developed by [Whi80]:

$$\sqrt{\hat{V}(\hat{\beta}_k^{OLS})} = \sqrt{\frac{\sum_{i=1}^{N} \hat{r}_{ik} \cdot \hat{\varepsilon}_i^2}{SSR_k^2}}.$$

  Using robust standard error formulas, the usual t statistic is valid asymptotically. The F statistic does not work under heteroskedasticity. Robust versions are available in most softwares.

3 Statistical Inference

3.1 Testing a single parameter – t test

Assumption (A6): $\varepsilon_i \sim N(0, \sigma^2)$ independent of X. A(6) implies (A4) and (A5). (A6) can be replaced by "large N".

Under assumption (A1) – (A6):

1.
$$\hat{\beta}_k \sim N(\beta_k, V(\hat{\beta}_k)), \text{ or equivalently } \frac{\hat{\beta}_k - \beta_k}{SD(\hat{\beta}_k)} \sim N(0,1)$$
where $SD(\hat{\beta}_k) = \sqrt{V(\hat{\beta}_k)}$.

2.
$$t_k = \frac{\hat{\beta}_k - \beta_k}{SE(\hat{\beta}_k)} \sim t_{N-K-1}.$$

Significance level α and **critical value** c (Right-sided):

$$P(t_k > c) = \int_c^\infty f_{t_k}(x)dx = \alpha.$$

Rejection rule: reject $H_0 : \beta_k = 0$ if $\hat{\beta}_k$ is too large, i.e. $t_k > c$.
Significance level α and critical value c (Two-sided):

$$P(|t_k| > c) = \int_{-\infty}^{-c} f_{t_k}(x)dx + \int_c^\infty f_{t_k}(x)dx = \alpha$$

A regressor is said to be "statistically significant" if its regression coefficient is different from zero in a two-sided test. Rule of thumb:

- $|t_k| > 1.64 \longrightarrow$ "statistically significant at 10% level"
- $|t_k| > 1.96 \longrightarrow$ "statistically significant at 5% level"
- $|t_k| > 2.58 \longrightarrow$ "statistically significant at 1% level"

p**-values**:

- p-value in a right-sided test:
$$p = \int_{t_k}^\infty f_{t_k}(x)dx$$
- p-value in a two-sided test:
$$p = \int_{-\infty}^{-|t_k|} f_{t_k}(x)dx + \int_{|t_k|}^\infty f_{t_k}(x)dx$$

The p-value is the smallest significance level at which H_0 is rejected. If p-value is very small, e.g. close to 0, then the coefficient is very significant.

3.2 Testing multiple parameters – F test

Example 3.1. Motivating example:

- Unrestricted model:

  cumgpa $= (\beta_0+\delta_0\text{female})+(\beta_1+\delta_1\text{female})\cdot\text{sat}+(\beta_2+\delta_2\text{female})\cdot\text{hsperc}+(\beta_3+\delta_3\text{female})\cdot\text{tothrs}+\text{error}$

- Restricted model:

  $$\text{cumgpa} = \beta_0 + \beta_1 \cdot \text{sat} + \beta_2 \cdot \text{hsperc} + \beta_3 \cdot \text{tothrs} + \text{error}$$

- Hypothesis: $H_0 : \delta_0 = \delta_1 = \delta_2 = \delta_3 = 0$.
- When moving from the unrestricted to the restricted model, the SSR tends to increase. If the increase is large enough, then we may conclude that the δ_k's are indeed relevant.
- The F-statistic is:

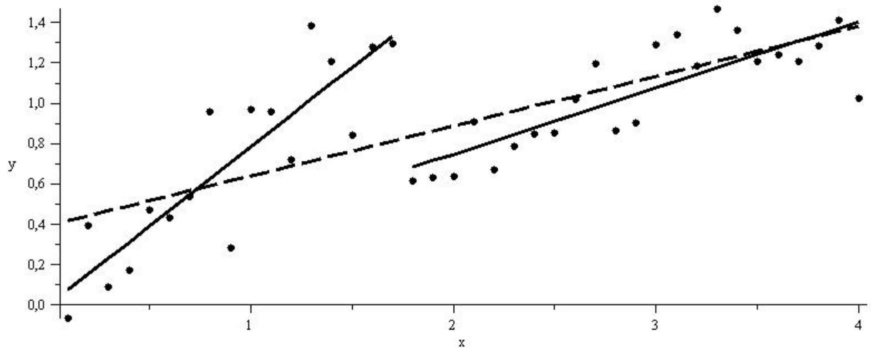
  $$F = \frac{(SSR_R - SSR_U)/q}{SSR_U/(N-K-1)} \sim F_{q,N-K-1}$$

  where q is the number of restrictions in the restricted model, and $(K+1)$ is the number of parameters in the unrestricted model.

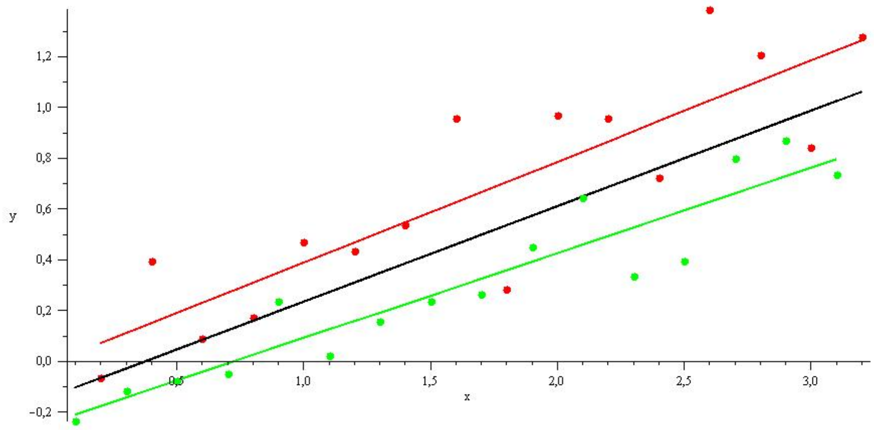
- Chow test: test whether the true coefficients in two linear regressions on different data sets are equal. Say our model is $y = a + bx + \varepsilon$, and each unit is either male or female. We run three regressions: one using the whole dataset ($\Rightarrow SSR$), one only on males ($\Rightarrow SSR_m$), and one only on females ($\Rightarrow SSR_f$). Then perform the test using

  $$F = \frac{[SSR - (SSR_m + SSR_f)]/(K+1)}{(SSR_m + SSR_f)/[N-2(K+1)]} \sim F_{(K+1),[N-2(K+1)]}$$

  Below are two illustrations[1] on motivations for using Chow test. The first figure shows that the two groups have different slopes, while the second shows that the two groups have different intercepts.



[1] taken from the Internet.



- Testing overall significance: model is $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K + \varepsilon$, and hypothesis is $H_0 : \beta_0 = \cdots = \beta_K = 0$. The F statistic is

$$F = \frac{(SSR_H - SSR)/K}{SSR/(N-K-1)} = \frac{R^2/K}{(1-R^2)/(N-K-1)} \sim F_{K,(N-K-1)},$$

where R^2 is the R^2 of the model.

4 Linear Regression and Causality

- (Mean) Regression Specification of *Potential* Outcomes:

  $$Y_i(D_i) = E[Y_i(D_i)] + U_i(D_i),$$

  where $E[U_i(1)] = E[U_i(0)] = 0$.

  - Causal effect for i:

    $$\Delta_i = Y_i(1) - Y_i(0) = E[\Delta_i] + U_i(1) - U_i(0)$$

  - ATE and ATT

    $$\begin{cases} ATE = E[\Delta_i] \\ ATT = E[\Delta_i] + E[U_i(1) - U_i(0)|D_i = 1] \end{cases}$$

- (Mean) Regression Specification of *Realized* Outcomes:

  $$\begin{aligned} Y_i &= D_i \cdot Y_i(1) + (1 - D_i) \cdot Y_i(0) \\ &= D_i \cdot (E[Y_i(1)] + U_i(1)) + (1 - D_i) \cdot (E[Y_i(0)] + U_i(0)) \\ &= E[Y_i(0)] + \Delta_i \cdot D_i + U_i(0) \\ &= E[Y_i(0)] + (E[\Delta_i] + U_i(1) - U_i(0)) \cdot D_i + U_i(0) \\ &= E[Y_i(0)] + E[\Delta_i] \cdot D_i + \varepsilon_i, \\ &= \beta_0 + \beta_1 \cdot D_i + \varepsilon_i \end{aligned}$$

  where $\varepsilon_i = (U_i(1) - U_i(0)) \cdot D_i + U_i(0)$ is a linear function of D_i.

  $\beta_1 \equiv E[\Delta_i]$ is the ATE. But ε_i and D_i are in general correlated, so $E[\varepsilon_i D_i] = 0$ and hence (A4) $E[\varepsilon_i|D_i] = 0$ are in general violated. The OLS estimator would give a biased estimate of ATE, unless we have ZCM:

  $$E[\varepsilon_i|D_i] = E[\varepsilon_i] = 0 \Leftrightarrow E[\varepsilon_i|D_i = 1] = E[U_i(1)|D_i = 1] = 0 = E[U_i(0)|D_i = 0] = E[\varepsilon_i|D_i = 0].$$

  This can be achieved by two means:

  - Random assignment experiment:

    $$\begin{cases} E[U_i(1)|D_i = 1] = E[U_i(1)|D_i = 0] = E[U_i(1)] = 0, \\ E[U_i(0)|D_i = 0] = E[U_i(0)|D_i = 1] = E[U_i(0)] = 0. \end{cases}$$

- Mean independence assumption in observational settings:

$$\begin{cases} E[U_i(1)|D_i = 1] = E[U_i(1)] = 0, \\ E[U_i(0)|D_i = 0] = E[U_i(0)] = 0. \end{cases}$$

- **Conditional Independence Assumption (CIA)**

  In observational settings, ruling out selection by simply invoking ZCM and insisting on using the naive comparison is likely to yield biased estimates of the causal parameters. An approach is to assume that *conditional* on a set of variables X_i, the treatment is as good as random assignment:

$$f(Y(0), Y(1)|D, X) = f(Y(0), Y(1)|X).$$

  This implies

$$\begin{cases} E[U_i(1)|D_i = 1, X_i] = E[U_i(1)|X_i] = 0 \\ E[U_i(0)|D_i = 0, X_i] = E[U_i(0)|X_i] = 0 \end{cases}$$

  Hence, the naive comparison conditional on X identifies the ATE conditional on X:

$$E[Y_i|D_i = 1, X_i] - E[Y_i|D_i = 0, X_i] = E[Y_i(1) - Y_i(0)|X_i] \equiv ATE^X.$$

  In general, our treatment variable T_i can take on multiple values. In terms of regression, our model is

$$g_i(t) = \beta_0 + \beta_1 t + \epsilon_i \quad \text{with } E(\epsilon_i) = 0$$

  where $g_i(t)$ is an individual-level response function mapping values of t_i to potential outcomes. Substituting the observed values of t_i yields

$$y_i = \beta_0 + \beta_1 t_i + \epsilon_i \quad \text{with } E(\epsilon_i) = 0$$

  The concern here is that the treatment t_i may be correlated with the error ϵ_i $(E[\epsilon_i|t_i] \neq 0)$, and hence the outcome y_i. We solve this problem by assuming that

$$E[\epsilon_i|t_i, X_i] = E[\epsilon_i|X_i] = X_i'\gamma.$$

  In other word, we assume that $\epsilon_i = E[\epsilon_i|X_i] + \varepsilon_i = X_i\gamma + \varepsilon_i$ with $E(\varepsilon_i) = 0$, so the error has nothing to do with t_i. The regression equation then becomes

$$y_i = \beta_0 + \beta_1 t_i + X_i'\gamma + \varepsilon_i.$$

  Now it should be clear that ZCM holds (e.g. $E[\varepsilon_i|t_i, X_i] = E[\varepsilon_i|X_i] = 0$):

$$\begin{aligned} E[\varepsilon_i|t_i, X_i] &= E[\epsilon_i - E[\epsilon_i|X_i]|t_i, X_i] \\ &= E[\epsilon_i - X_i'\gamma|t_i, X_i] \\ &= E[\epsilon_i|t_i, X_i] - E[X_i'\gamma|t_i, X_i] \\ &= E[\epsilon_i|X_i] - X_i'\gamma \\ &= 0 \end{aligned}$$

5 Difference in Difference (DID)

If we also observe outcomes of units *before* the treatment, then we may take the difference

$$A - B = E[Y_{i,t+1}(1)|D_i = 1] - E[Y_{i,t}(0)|D_i = 1] = \bar{Y}^T_{t+1} - \bar{Y}^T_t.$$

This is the Before-After (B-A) comparison. Because we are comparing mean outcomes, we can use both pooled cross sections and panel data. Recall that the ATT is

$$\delta^{ATT} = E[Y_{i,t+1}(1)|D_i = 1] - E[Y_{i,t+1}(0)|D_i = 1],$$

so we are substituting $E[Y_{i,t+1}(0)|D_i = 1]$ by $E[Y_{i,t}(0)|D_i = 1]$. The problem is that $Y_{i,s}(0)$ might change for different time period s. The mean outcomes before and after the treatment is composed of both the treatment effect, and a time trend. For example, if health status improved after treatment, it may be partially due to the treatment, but it may also be the case that all people becomes healthier than before, so the $A - B$ comparison would exaggerate the treatment effect.

In general, we can decompose $A - B$ as

$$A - B = \mathit{treatment\_effect} + \mathit{trend},$$

where treatment effect and the trend can be in any direction. Then the true treatment effect is the $A - B$ comparison minus the trend:

$$\mathit{treatment\_effect} = (A - B) - \mathit{trend}.$$

We approximate the $trend$ for the treated by trend for the untreated. Since for the untreated group $\mathit{treatment\_effect} = 0$, we have $\mathit{trend}_{un} = (A - B)_{un}$, so the trend for the untreated is the difference of their mean outcomes before and after:

$$DID = \mathit{treatment\_effect} = (A - B) - (A - B)_{un}.$$

This is DID. A regression specification is

$$y_{it} = (\beta_0 + \delta_0 \cdot dAfter_t) + (\beta_1 + \delta_1 \cdot dAfter_t) \cdot dTreated_i + \epsilon_{it}.$$

In this case $DID = \delta_1 = \delta^{ATT}$:

	Before	After	After - Before
Control	β_0	$\beta_0 + \delta_0$	δ_0
Treatment	$\beta_0 + \beta_1$	$\beta_0 + \delta_0 + \beta_1 + \delta_1$	$\delta_0 + \delta_1$
Treatment-Control			δ_1

Robustness checks:

- "Placebo" test: generate "placebo" treatments and compute DID. We should expect zero effect.
- Sensitivity analysis: re-estimate the same DID regression in different sub-samples of units and/or by adding covariates. Hope for stability of estimated effects across such perturbations.
- Adding covariates: account for factors that may capture different trends for treated and untreated units over time.

6 Instrumental Variables (IV)

Setup:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 s_i + \epsilon_i \\ &= \beta_0 + \beta_1 s_i + [X_i'\gamma + \varepsilon_i], \end{aligned}$$

where some variables in X_i are omitted because unobservable and thus $E[\epsilon_i|s_i] \neq 0$. An instrumental variable is a variable z_i that satisfies two conditions:

1. Relevance: $Cov(z_i, s_i) \neq 0$
   This condition is testable: estimate $s_i = \pi_0 + \pi_1 z_i + \text{error}_i$ and test the hypothesis that $\pi_1 = 0$ using a t test.
2. Validity: $Cov(z_i, \epsilon_i) = 0$
   In words: the IV does not feature the same problem/failure of the treatment variable to be instrumented. Hardly verifiable.

The intuition behind IV is that we use something that is irrelevant of the error ϵ_i to "nudge" the treatment variable s_i, who then induces changes to outcome y_i. This way we can infer the causal effect of the treatment without bias.

We can recover the casual parameter β_1 as

$$Cov(y_i, z_i) = Cov(\beta_0 + \beta_1 s_i + \epsilon_i, z_i) = \beta_1 Cov(s_i, z_i) + Cov(\epsilon_i, z_i) = \beta_1 Cov(s_i, z_i) + 0 = \beta_1 Cov(s_i, z_i)$$

$$\Downarrow$$

$$\beta_1 = \frac{Cov(y_i, z_i)}{Cov(s_i, z_i)} = \frac{Cov(y_i, z_i)/V(z_i)}{Cov(s_i, z_i)/V(z_i)}$$

6.1 IV as LATE

IV can be viewed as a method for estimating the local average treatment effect. Suppose we have a random assignment of treatment. ((A2) random assignment) We let Z_i denote the assignment status for unit i. $Z_i = 1$ if unit i is assigned to treatment and 0 otherwise. The problem here is that compliance is not perfect and there are some units in the treatment group who do not get the treatment, and there may also be units in the control group who get treated in the end. Ideally we would like $D_i(0) = 0$ and $D_i(1) = 1$, but it may be the case that $D_i(1) = 0$ or $D_i(0) = 1$ for some unit i. We can classify the units into four groups (A1):

	$D_i(0) = 0$	$D_i(0) = 1$
$D_i(1) = 0$	*Never-taker*	*Defier*
$D_i(1) = 1$	*Complier*	*Always-taker*

1. We maintain the assumption that the probability of treatment in the two assignment groups must be different, otherwise there is no way to distinguish the two. In other words, Z and D must at least be correlated: $Cov(Z_i, D_i) \neq 0$. ((A3) nonzero ATE of Z on D)
2. We also have to maintain the assumption that outcomes are only affected by treatment; they are not directly related to assignments, i.e. $Cov(Y_i, Z_i) = 0$. This implies in particular that the treatment effect for never-takers is always zero, so they do not contaminate the observed results. This assumption is however not testable, since we can not observe both assignment status for one unit i and compare the two. ((A4) exclusion restriction)
3. Remember our goal here is to identify the effect of treatment for the compliers. We maintain the assumption that all treated units that we observed in the control group are always-takers. In other words, there are no defiers in the units. This way we can pin down the number of compliers using our observed data:

   $$\text{compliers} = (\textit{treatment group}) - (\textit{never-takers}) - (\textit{always-takers}),$$

   where *always-takers* is the number of units in the control group that picked up the treatment, and *never-takers* is the number of units in the treatment group that are not treated in the end. ((A5) monotonicity)

The observed average outcome for the treatment group (i.e.$\{i : Z_i = 1\}$) is composed of

$$\begin{aligned} E\{Y_i[1, D_i(1)]\} &= \alpha_1 \cdot E[Y(1,1) \mid \textit{compliers}\,] \\ &+ \alpha_2 \cdot E[Y(1,1) \mid \textit{always-takers}\ ] \\ &+ \alpha_3 \cdot E[Y(1,0) \mid \textit{never-takers}\,], \end{aligned}$$

where α_j is the proportion of the corresponding group. Similarly, the average outcome for the control group ($\{i : Z_i = 0\}$) can be decomposed into

$$\begin{aligned} E\{Y_i[0, D_i(0)]\} &= \alpha_1 \cdot E[Y(0,0) \mid \textit{compliers}\,] \\ &+ \alpha_2 \cdot E[Y(0,1) \mid \textit{always-takers}\ ] \\ &+ \alpha_3 \cdot E[Y(0,0) \mid \textit{never-takers}\,], \end{aligned}$$

Thus the difference is

$$\begin{aligned} E\{Y_i[1, D_i(1)]\} - E\{Y_i[0, D_i(0)]\} &= \alpha_1 \cdot \big(E[Y(1,1) \mid \textit{compliers}\,] - E[Y(0,0) \mid \textit{compliers}\,]\big) \\ &+ \alpha_2 \cdot \big(E[Y(1,1) \mid \textit{always-takers}\ ] - E[Y(0,1) \mid \textit{always-takers}\,]\big) \\ &+ \alpha_3 \cdot \big(E[Y(1,0) \mid \textit{never-takers}\,] - E[Y(0,0) \mid \textit{never-takers}\,]\big) \\ &= \alpha_1 \cdot LATE + 0 + 0 \\ &= \alpha_1 \cdot LATE \end{aligned}$$

We thus have

$$LATE = \frac{E\{Y_i[1, D_i(1)]\} - E\{Y_i[0, D_i(0)]\}}{\text{proportion of compliers}}.$$

If we use Z_i as an IV for D_i, we would obtain $\beta = Cov(Y_i, Z_i)/Cov(D_i, Z_i)$. This actually equals the local average treatment effect (LATE). First note that if Z is a binary random variable such that $Z = 1$ with probability p, then

$$\begin{aligned} Cov(Y, Z) &= E[YZ] - E[Y]E[Z] \\ &= E[Y \mid Z = 1] \cdot p - \big(E[Y \mid Z = 1] \cdot p + E[Y \mid Z = 0] \cdot (1 - p)\big) \cdot p \\ &= p(1 - p) \cdot E[Y \mid Z = 1] - p(1 - p) \cdot E[Y \mid Z = 0] \\ &= p(1 - p) \cdot \big(E[Y \mid Z = 1] - E[Y|Z = 0]\big). \end{aligned}$$

So we have

$$IV = \frac{Cov(Y_i, Z_i)}{Cov(D_i, Z_i)} = \frac{E[Y_i \mid Z_i = 1] - E[Y_i \mid Z_i = 0]}{E[D_i \mid Z_i = 1] - E[D_i \mid Z_i = 0]}$$

$$= \frac{E[Y_i(1)] - E[Y_i(0)]}{P[D_i(1) = 1] - P[D_i(0) = 1]} = \frac{E[Y_i(1) - Y_i(0)]}{E[D_i(1) - D_i(0)]}$$

$$\begin{aligned} &= \frac{E\{Y_i[1, D_i(1)]\} - E\{Y_i[0, D_i(0)]\}}{P[D_i(1) - D_i(0) = 1]} \\ &= LATE \end{aligned}$$

In summary, if we use a binary variable Z as an instrument for treatment D, then the IV estimate gives us the *local average treatment effect*, under assumptions (A1)–(A5).

7 Regression Discontinuity Design (RDD)

7.1 Sharp RDD

Setup:
Our *forcing variable* is X_i. If X_i is larger or equal to a threshold, say c, then unit i will get treated; if $X_i < c$, then unit i will not get treated. In other words, $D_i = 1\{X_i \geq c\}$. A classical example is the effect of merit award on earnings. For a student i, if his score X_i is larger than a threshold c then he gets a scholarship, and no scholarship otherwise. The forcing variable is his score X_i and the treatment variable D_i is scholarship. Y_i is his subsequent earnings. $Y_i(0)$ is his potential earnings without scholarship, and $Y_i(1)$ is his potential earnings with the scholarship. The problem here is that $Y_i(0)$ and $Y_i(1)$ are correlated with X_i: students with higher scores obtain higher earnings. So if we want to know the effect of merit award on earnings, then comparing $E[Y_i(1)]$ with $E[Y_i(0)]$ naively would get a biased result that incorporates not only the treatment effect, but also the effect of test scores X (and hence other traits).

The idea of RDD is that we compare outcomes just around the (arbitrary) cutoff. Students just above and just below the cutoff may have similar characteristics, so near the cutoff the assignment of scholarships is as if random.

We would like to identify

$$E[Y_i(1) \mid c^+] - E[Y_i(0) \mid c^-]$$

as the local average treatment effect. Here we are using $E[Y_i(0) \mid c^-]$ as a counterfactual for unobserved $E[Y_i(0) \mid c^+]$, i.e. the average outcome for those who get the awards if their awards were deprived. So we are assuming that $E[Y_i(0) \mid c^+] = E[Y_i(0) \mid c^-]$, in other words the function $g(x) = E[Y_i(0) \mid X = x]$ is *continuous* at the cutoff point c.

7.1.1 Parametric Estimation

To estimate the treatment effect using parametric methods, we define a new variable $\tilde{X}_i = X_i - c$. Our model for $E[Y_i \mid \tilde{X}_i]$ can be:

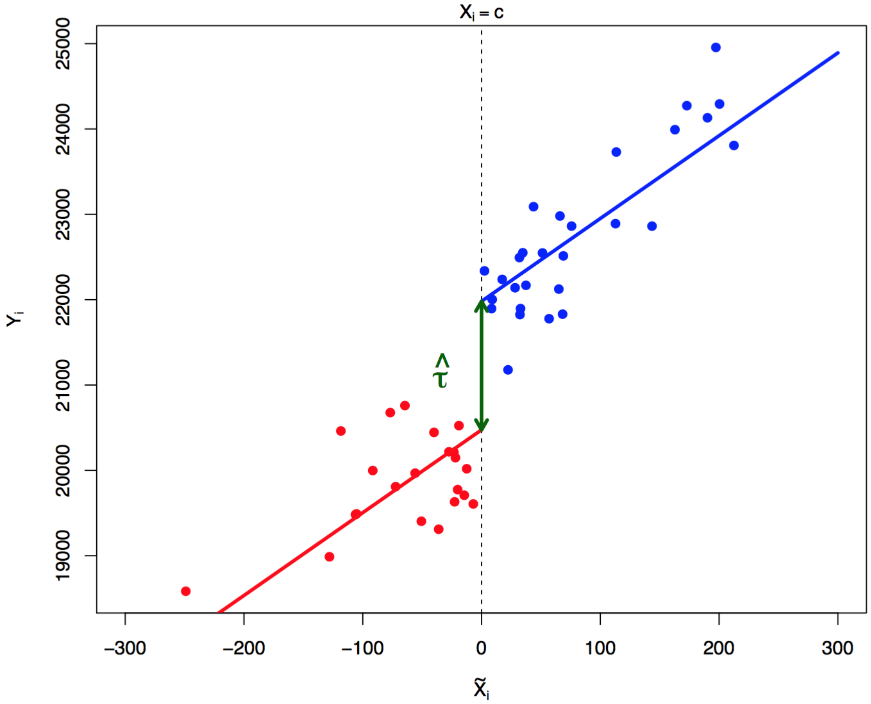
- *Linear, and common slope for $E[Y_i \mid \tilde{X}_i < 0]$ and $E[Y_i \mid \tilde{X}_i > 0]$*
  Specifically, we assume $E[Y_i(0) \mid X_i = x]$ is linear in x, and the treatment effect τ does not depend on X_i, i.e.

$$E[Y_i(0) \mid X_i] = \alpha + \beta X_i \quad \text{and} \quad E[Y_i(1) - Y_i(0) \mid X_i] = \tau,$$

which implies $E[Y_i(1) \mid X_i] = \tau + \alpha + \beta X_i$. Therefore, the model for observed outcome is

$$\begin{aligned} E[Y_i \mid X_i, D_i] &= D_i \cdot E[Y_i(1) \mid X_i] + (1 - D_i) \cdot E[Y_i(0) \mid X_i] \\ &= \alpha + \tau D_i + \beta X_i \\ &= \tilde{\alpha} + \tau D_i + \beta \tilde{X}_i. \end{aligned}$$

So we just regress observed outcome on D_i and $\tilde{X}_i$.



- *Linear, different slopes*

  We now allow treatment effect to depend on X_i. We specify that

$$E[Y_i(0) \mid X_i] = \alpha_0 + \beta_0 X_i \quad \text{and} \quad E[Y_i(1) \mid X_i] = \alpha_1 + \beta_1 X_i,$$

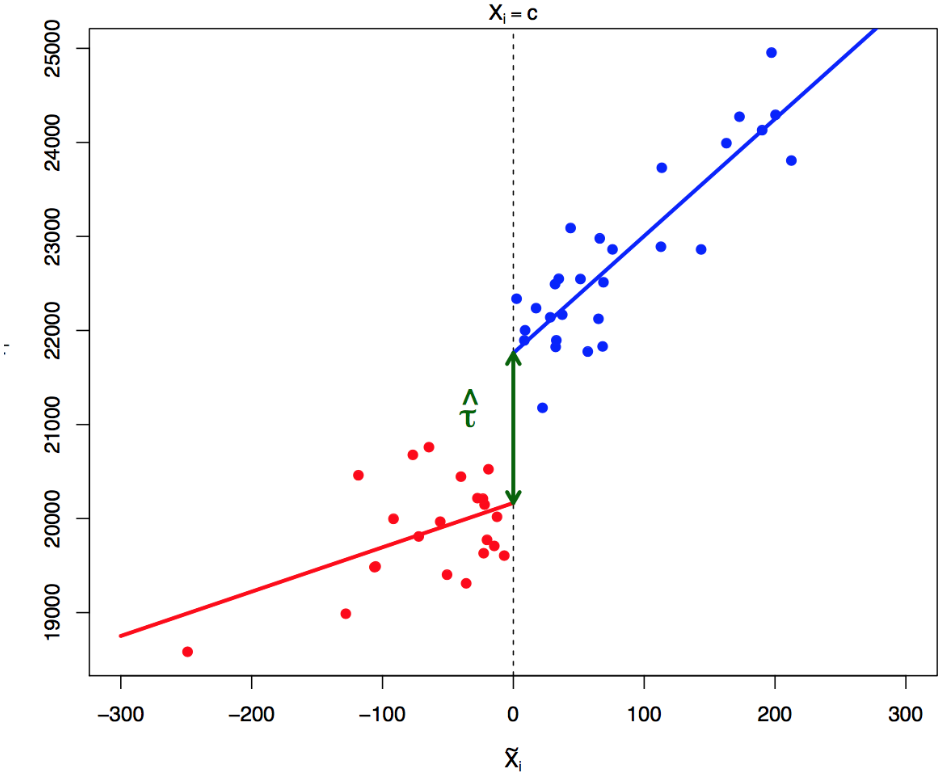
  so that the treatment effect is

$$E[Y_i(1) - Y_i(0) \mid X_i] = (\alpha_1 - \alpha_0) + (\beta_1 - \beta_0) X_i.$$

  The observed outcome model is therefore

$$\begin{aligned} E[Y_i \mid X_i, D_i] &= D_i \cdot E[Y_i(1) \mid X_i] + (1 - D_i) \cdot E[Y_i(0) \mid X_i] \\ &= \alpha_0 + \beta_0 X_i + (\alpha_1 - \alpha_0) D_i + (\beta_1 - \beta_0) D_i X_i \\ &= (\alpha_0 + \beta_0 c) + \beta_0 \tilde{X}_i + \{(\alpha_1 - \alpha_0) + (\beta_1 - \beta_0) c\} D_i + (\beta_1 - \beta_0) D_i \tilde{X}_i \\ &= \tilde{\alpha} + \tau D_i + (\beta_0 + \tilde{\beta} D_i) \tilde{X}_i. \end{aligned}$$

Note that $\tau = E[Y_i(1) - Y_i(0) \mid X_i = c]$, the LATE at the threshold. So in this case we just regress Y_i on $\tilde{X}_i$, D_i and the interaction $D_i\tilde{X}_i$.

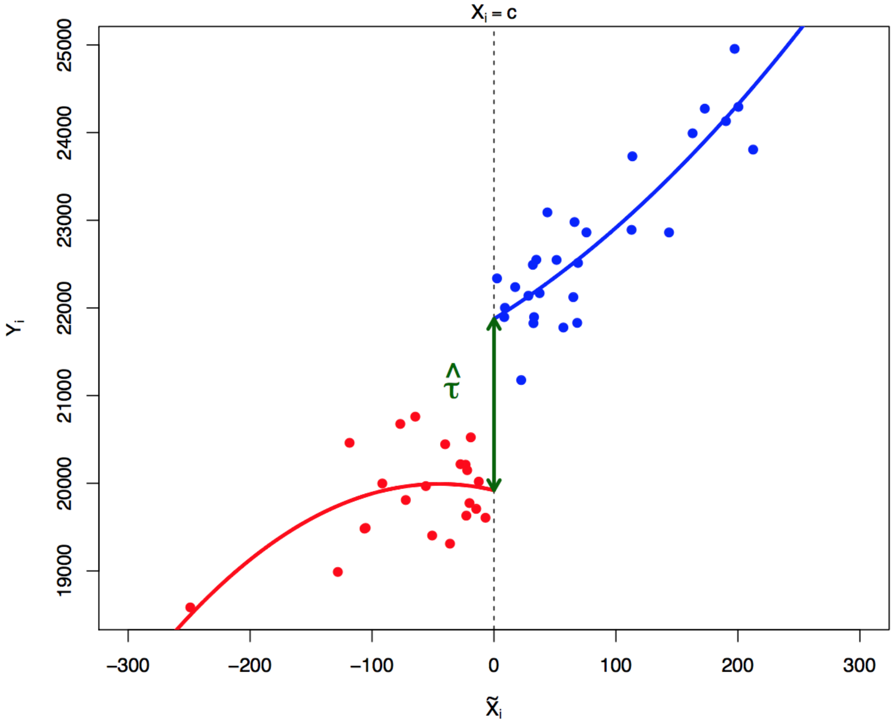


- *non-linear*

  $E[Y_i(0) \mid X_i = x]$ and $E[Y_i(1) \mid X_i = x]$ can be non-linear in X_i, and treatment effect can also vary across X_i. E.g. quadratic specification:

$$E[Y_i \mid X_i, D_i] = (\gamma_0 + \tau D_i) + (\gamma_1 + \alpha_1 D_i)\tilde{X}_i + (\gamma_2 + \alpha_2 D_i)\tilde{X}_i^2.$$

In this case, τ is still LATE at the threshold: $\tau = E[Y_i(1) - Y_i(0) \mid X_i = c]$.



7.2 Fuzzy RDD

In this scenario, compliance is not perfect. The outcome is determined through

$$X_i \longrightarrow Z_i \longrightarrow D_i \longrightarrow Y_i$$

where X_i is unit i's score, Z_i is the encouragement for treatment, i.e. $Z_i = 1\{X_i \geq c\}$, and D_i is the treatment. To identify the treatment effect, we need to assume

1. *Both* $E[D_i(z) \mid X_i = x]$ (potential treatment) and $E[Y_i(z) \mid X_i = x]$ (potential outcome) are continuous in x around $X_i = c$, for $z = 0, 1$.
2. Monotonicity, exclusive restriction, and relevance of Z_i.

Our estimand is

$$\tau_F = E[Y_i(1) - Y_i(0) \mid \text{ unit } i \text{ is complier and } X_i = c].$$

It is identified as

$$\tau_F = \frac{E[Y_i(1) \mid c^+] - E[Y_i(0) \mid c^-]}{E[D_i(1) \mid c^+] - E[D_i(0) \mid c^-]}.$$

7.2.1 Parametric Estimation

We can estimate τ_F using 2SLS:

$$Y_i = \beta_0 + \tau \hat{D}_i + (\beta_1 + \beta_2 Z_i) \cdot \tilde{X}_i + \varepsilon_i,$$

where $\hat{D}_i$ is instrumented by Z_i.

References

Holland, P. W. (1986). “Statistics and Causal Inference”. In: *Journal of the American Statistical Association* 81.396, pp. 945–960.

White, H. (1980). “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity”. In: *Econometrica* 48.4, pp. 817–838.