Additional Empirical Techniques

Take the simple **mincerian equation** (income x schooling), below:

$$Ln\, Y_i = \beta_0 + \beta_1 S_i + \beta_k \gamma_{k,i} + \varepsilon_i \qquad (1)$$

- There are two initial problems in a model which may bias β:

1. **Omitted variable bias:**

Imagine that a true population model is

$$Ln\, Y = \beta_0 + \beta_1 S + \beta_2 S_f + u \qquad (2)$$

Where S_f is the level of schooling of the father. However, if it is not specified, we may derive the algebric relationship for our estimated coefficient ($\widehat{\beta_1}$):

$$\widehat{\beta_1} = \beta_1 + \beta_2 \delta_1 \qquad (3)$$

Where β_1 and β_2 are the true coefficients from model (2), and δ is the slope from the simple regression:

$$S = \delta_0 + \delta_1 S_f + \vartheta \qquad (4)$$

So, if there's a positive relationship between a population schooling and their father's schooling, our estimated $\hat{\beta}_1$ will be biased upwards.

2. **Measurement error bias**

Let's say that people are randomly mistaken (or lying) about their own schooling level. So, our variable is:

$$\check{S} = S + \rho \qquad (5)$$

Imagine that our mincerian equation (1) is not underspecified, so there's no omitted variable bias. That means that our estimated coeficiente is:

$$\widehat{\beta_1} = \frac{\sigma_S}{\sigma_S + \sigma_\rho} \beta_1 \qquad (6)$$

Where σ is the standard deviation.

So, measurement error measurement bias our estimate downwards.

Usually, omitted variable bias and measurement error bias have opposite directions, so when they are jointly playing a role in a regression, they are actually mostly cancelled off. However, if one may not rely on the estimate, there are a few ways in order to address this problem. One of the main ways to do so is to use the "fixed-effects" approach, for panel data.

Fixed-effects regressions

Panel data are those in which you have the same set of observations in more than one period in time. Therefore you can use the "fixed-effects" strategy, where you control all fixed mean differences between people in any observable or unobservable predictors, such as genetic and environmental differences among workers with each level of schooling. Fixed time effects also control mean changes that have occurred for all groups.

Summarizing, the estimated β_1 will capture the changes of schooling and earnings that happened differently for workers. Our fixed-effects mincerian equation is presented below:

$$Y_{i,t} = \beta_0 + \beta_1 S_{i,t} + \beta_k X_{k,i,t} + \sum_{\tau=1}^{T} \delta_\tau 1(t = \tau) + \sum_{\varphi=1}^{I} \gamma_\varphi \, 1(i = \varphi) + \varepsilon_{i.t} \qquad (7)$$

Where $\sum_{\tau=1}^{T} \delta_\tau 1(t = \tau)$ are time fixed-effects and $\sum_{\varphi=1}^{I} \gamma_\varphi \, 1(i = \varphi)$ are individual fixed-effects. Panel data also allows the researcher to control for the different trends in broader groups, such as cities, states and aggregated low/high levels of schooling.

Fixed-effect regressions still may be biased. For instance, omitted variable bias may be present if there's any correlation among the changes of schooling, income and another factor, like other public policies that were implemented along with more provision of education. However, when the observations are on the individual level, fixed-effects may be proper controls for any endogeneity.

Propensity Score Matching (PSM)

Propensity-score matching uses an average of the outcomes of similar subjects who get the other treatment level to impute the missing potential outcome for each subject. The average treatment effect (ATE) is computed by taking the average of the difference between the observed and potential outcomes for each subject.

PSM does not need bias correction, because it matches on a single continuous covariate.

The propensity score was defined by Rosenbaum and Rubin (1983a) to be the probability of treatment assignment conditional on observed baseline covariates. The propensity score is a balancing score: conditional on the propensity score, the distribution of measured baseline covariates is similar between treated and untreated subjects. Thus, in a set of subjects all of whom have the same propensity score, the distribution of observed baseline covariates will be the same between the treated and untreated subjects.

Therefore, the "propensity score" is the conditional probability of receiving treatment given the variables observed X before treatment:

$$p(X) = \Pr\{D = 1|X\} = E\{D|X\} \qquad (8)$$

Lemma 1: Balance the pretreatment variables, X, given the propensity score (Rosenbaum and Rubin, 1983)

Let p(X) be the "propensity score"

$$X \perp D \mid p(X) \tag{9}$$

or

$$F(X|D=1,P(X))=F(X|D=0,P(X)) \tag{10}$$

Motto 2: Not biased, given the "propensity score" (Rosenbaum and Rubin, 1983)

Assuming there is conditional independence:

$$Y(1), Y(0) \perp D \mid X \tag{11}$$

So the treatment assignment is not biased given the "propensity score"

$$Y(1), Y(0) \perp D \mid p(X) \tag{12}$$

Using the "propensity score" we can homogenize treatments and controls based on this, instead of the multidimensional vector X

$$E\{Yi(0)|Di = 0, p(Xi)\} = E\{Yi(0)|Di = 1, p(Xi)\} = E\{Yi(0)|p(Xi)\} \tag{13}$$
$$E\{Yi(1)|Di = 0, p(Xi)\} = E\{Yi(1)|Di = 1, p(Xi)\} = E\{Yi(1)|p(Xi)\} \tag{14}$$

Using these expressions, we can define for each cell defined by p(X):

$$\delta p(x) \equiv E\{\Delta i|p(Xi)\} \tag{15}$$

$$\equiv E\{Yi(1)|p(Xi)\} - E\{Yi(0)|p(Xi)\} \tag{16}$$

$$= E\{Yi|Di = 1, p(Xi)\} - E\{Yi|Di = 0, p(Xi)\}. \tag{17}$$

Parametric estimation; most usual example: Logit (or probit)

$$\Pr\{D_i|X_i\} = \frac{e^{\lambda h(X_i)}}{1+e^{\lambda h(X_i)}} \tag{18}$$

Define the sample to estimate the effect in such a way that the propensity score of the treatment and control group satisfies:

$$0<a<P(X)<b<1 \tag{19}$$

Maximum area for feasible values of "a" and "b"

a = max {min (treated support, counterfactual support)}

b = min {max (treated support, counterfactual support)}

The average of the result variable for the treated group is:

$$\bar{Y}(1) = \frac{1}{N^T} \sum_{i \in treated} Y_i \quad (20)$$

It is necessary to estimate this average for the counterfactual group:

- Closest neighbor (homogenization one by one or more than one) with or without substitution

- Kernel

Steps to follow:

1. For each propensity score value between the treaties compute the following equation using the control sample

$$\bar{Y}^C(\hat{P}(X_i)) = \frac{1}{\sum_{j=1}^{N_C} \omega_j} \sum_{k=1}^{N_C} Y_k \omega\left(\hat{P}(X_i) - P(X_k)\right) \quad (21)$$

Where W is a function of weights that decreases with distance. K denotes a kernel function:

$$W\ (P(X)) = \frac{K(\frac{P(X_i) - P(X_k)}{h})}{\sum_{\{D=0\}} K(\frac{P(X_i) - P(X_k)}{h})} \quad (22)$$

Another possibility is to use what became known as Propensity Score Weghting. This method mixes PSM with regression, which can increase the efficiency of the estimators because you use the covariates when calculating the ToT. In this case, the model is:

$$Y_i = \beta_0 + \beta_1 T_i + \beta_{k,i} \gamma_k + \varepsilon_i \quad (23)$$

Weights are 1 for treated individuals. For individuals in the control group, the weights are:

$$\widehat{P(X)} \Big/ 1 - \widehat{P(X)} \quad (24)$$

Heckman Adjustment

Considering $\mathrm{Y_i} = \beta_0 + \beta_1 \mathrm{S_i} + \beta_k \gamma_{k,i} + \varepsilon_i$, but where not all $\mathrm{Y_i}$ is observed, but rather there is a selection bias, where:

$$\mathrm{P(Labor)_i} = \delta_0 + \textstyle\sum_k \delta_k z_{k,i} + \epsilon_i \tag{25}$$

Where cor(ε, ϵ) = ρ > 0

That is, an individual's labor supply generates a selection bias on mincerian equation estimates. That is particularly true for women.

Heckman adjustment controls this correlation, adding to the mincerian equation the cdf of predicted P(Labor).

Quantile Regression

Instead of using a mean function of linear regression, one may use the conditional median function $\mathrm{Q_q(y|x)}$, where q is the Xth percentile. While OLS minimizes $\sum \varepsilon_i^2$, a quantile regression, also known as least-absolute-deviations (LAD) regression, minimizes $\sum |\varepsilon_i|$.

Quantile regressions provide snapshots of different points of a conditional distribution. They constitute a parsimonious way of describing the whole distribution and should bring much value-added if the relationship between the regressors and the independent variable evolves across its conditional distribution.

For example:

$$Y_i = \beta_{0,0.5th} + \beta_{1,0.5th} S_i + \beta_{k,0.5th,} \gamma_{k,i} + \varepsilon_i \tag{26}$$

So, $\beta_{1,0.5\mathrm{th}}$ is the estimate of having one more year of study over the median (it may be any quantile of the distribution).

Median regression is more robust to outliers than least squares regression, and is semiparametric as it avoids assumptions about the parametric distribution of the error process. While OLS can be inefficient if the errors are highly non-normal, QR is more robust to non-normal errors and outliers. QR also provides a richer characterization of the data, allowing us to consider the impact of a covariate on the entire distribution of y, not merely its conditional mean.

For visualization and application of quantile regressions, check the link below:

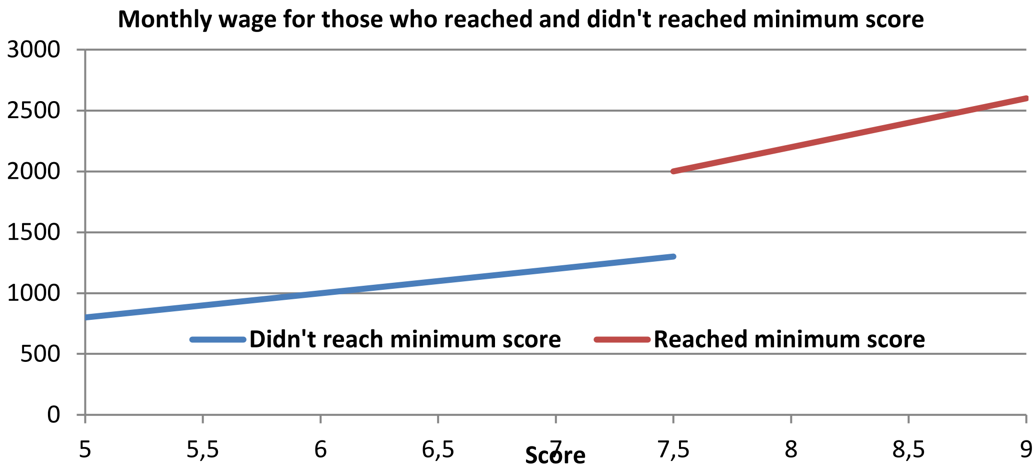
https://data.library.virginia.edu/getting-started-with-quantile-regression/

Regression Discontinuity Design

Currently, this is an useful causal evaluation method. This is because RDD estimates the average treatment effect by exploring a selection rule based on a discontinuity in a particular observable variable.

For example, let's think about the college entrance examination, in which the students enter for college if they take a certain score. In principle, students with higher grades in the college entrance examination tend to perform better in the job market, which may bias the estimate.

But in this case, we can compare students who took notes very close to the minimum score, up and down.



Say that f [.] is the function, x is the continuous variable (such as the score). xm is the treatment threshold separating the units (individuals) into two groups: those receiving treatment (D = 1) and those who did not receive (D = 0). So, our regression is like:

$$D_i \;=\; D(x_i) = f[x_i \;>\; x_m] \tag{27}$$

$$Y_i \;=\; \beta_0 + \beta_1 Score_i + \beta_2 D_i + \beta_3 Score_i \times D_i + \varepsilon_i \tag{28}$$

Usually, the sample is subset into a window of x around xm. If not small enough, our regression may add higher polynomial degrees for the running variable X, in order to have a better fit. In this first example presented, we have the case of a "sharp RDD", in which the cutoff 100% determinant for the treatment. However, in some cases it may not be so.

Sometimes passing from the left to the right side of the cutoff does not guarantee treatment, but significantly increases its likelihood. In this case, we have what is called "fuzzy RDD", in which the methods of "sharp RDD" and instrumental variables are combined. More specifically, the discontinuity will instrument the treatment, as showed below.

$$X_i \;=\; \theta_0 + \theta_1 Score_i + \theta_2 D_i + \theta_3 Score_i \times D_i + \varepsilon_i \tag{29}$$

$$Y_i \;=\; \beta_0 + \beta_1 \hat{X}_i + \varepsilon_i \tag{30}$$