MIT 14.32/14.320 **Spring 2023**

Review of second part of course

Raymond Han and Martina Uccioli [1]

1 Regression basics

1.1 Definition

Assuming a linear model for the data : $Y_i = \alpha + \beta X_i + \epsilon_i$.
Regression solves the population mean-squared error minimization problem. That is regression coefficients α and β solves

$$\min E[(Y_i - \alpha - \beta X_i)^2]$$

From the FOC of the problem, it follows that regression coefficients are chosen so that the regression error satisfies :

$$E[Y_i - \alpha - \beta X_i] = E[\epsilon_i] = 0$$
$$E[(Y_i - \alpha - \beta X_i)X_i] = E[\epsilon_i X_i] = 0$$

Three interpretation of regression coefficients:

1. When the CEF is linear : the regression formula is the CEF. That is $E[Y_i|X_i] = \alpha + \beta X_i$
2. Regression gives the best linear predictor of Y_i given X_i
3. Regression gives the best linear approximation of the CEF $E[Y_i|X_i]$

1.2 OVB

The regression coefficients will have a causal interpretation only if the CEF they approximate is causal. That is, a change in X_i affects $E[Y_i]$ only through X_i and not through another variable. e.g. people with more years of education have higher wages because of their education and not because they were more capable in the first place and so got more education.

The OVB formula :

Suppose the following long regression is causal

$$Y_i = \alpha + \beta X_i + \gamma A_i + \epsilon_i$$

Instead, you compute the following short regression

$$Y_i = \alpha^* + \beta^* X_i + \epsilon_i^*$$

[1] Based on notes by Viola Corradini, Clemence Idoux, Maddie McKelway, Benjamin Marx and Ryan Hill.

Let δ_{AX} be the slope coefficient of the regression of A_i on X_i. Then the coefficient β^* is biased (in a causal sense):

$$\beta^* = \beta + \gamma\delta_{AX}$$

The **sign of the bias** depends on

- whether the effect of the omitted variable A_i on Y_i is positive or negative
- whether the omitted variable A_i is positively or negatively correlated with the variable of interest X_i

1.3 Regressions with interactions

- Recall the simple model with two dummy variables and an interaction term ($x_i z_i$):

$$y_i = a + bx_i + cz_i + dx_i z_i + e_i.$$

Group (1) with x=0, z=0:	$y_i = a + e_i$	$\Rightarrow$	mean is a
Group (2) with x=1, z=0:	$y_i = a + b + e_i$	$\Rightarrow$	mean is $a + b$
Group (3) with x=0, z=1:	$y_i = a + c + e_i$	$\Rightarrow$	mean is $a + c$
Group (4) with x=1, z=1:	$y_i = a + b + c + d + e_i$	$\Rightarrow$	mean is $a + b + c + d$

- b tests for a difference in means between groups (1) and (2)
- c tests for a difference in means between groups (1) and (3)
- d tests whether there is a combined extra effect of having $x = 1$ and $z = 1$

Warning: if you want to do a triple interactions, you need to include all the double interactions and single coefficients (so you will have a constant, 3 coeffs on the single variables, 3 coeffs on the double interactions, 1 coeff on the triple interaction).

2 Estimating regressions: OLS

To estimate the regression parameters, we use their **sample analogs** $\hat{\alpha}$ and $\hat{\beta}$ defined such that :

$$\sum_{i=1}^{n} Y_i - \hat{\alpha} - \hat{\beta}X_i = \sum_{i=1}^{n} e_i = 0$$
$$\sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta}X_i)X_i = \sum_{i=1}^{n} e_i X_i = 0$$

where e_i are the sample regression residuals.

$\hat{\alpha}$ and $\hat{\beta}$ are called the **OLS estimators** as they can also be derived by minimizing the sample equivalent of the population MSE:

$$\min \frac{1}{n} \sum_{i=1}^{n} (Y_i - \alpha - \beta X_i)^2$$

2.1 Properties of OLS

The classical OLS assumptions are as follows:

1. **Linear CEF:** $E[Y_i|X_i] = \alpha + \beta X_i$, thus $E[\epsilon_i|X_i] = 0$
2. **Random sampling:** $E[\varepsilon_i \varepsilon_j] = 0$ for $i \neq j$
3. **Homoskedasticity:** $E[\varepsilon_i^2|X_i] = E[\varepsilon_i^2] = \sigma_\varepsilon^2$
4. **Normality:** the ε_i are normally distributed
5. X_i is fixed in repeated samples

Under these assumptions,

1. the OLS estimator $\hat{\alpha}$ and $\hat{\beta}$ are unbiased for α and β
2. The sampling variance of $\hat{\beta}$ is $\frac{\sigma_\varepsilon^2}{ns_X^2}$
3. $\hat{\beta} \sim N(\beta, \frac{\sigma_\varepsilon^2}{ns_X^2})$
4. OLS is the Best (lowest variance) Linear Unbiased Estimator of β (Gauss-Markov)

- Specific violations of the classical assumptions will break these results:
  - If the CEF is non linear, OLS is no longer unbiased, but it is still consistent.
  - Heteroskedasticity, non-random sampling: OLS is no longer BLUE
  - Non Normally distributed errors, OLS is no longer normally distributed in small sample, only asymptotically.

2.2 Standard error issues

3 main types of issues can affect your standard errors

- These are all violations of the Gauss-Markov assumptions: OLS is no longer BLUE
- (But $\hat{\beta}^{OLS}$ is still consistent)

1. **Heteroskedasticity**: $E[\varepsilon_i^2|X_i] \neq \sigma_\varepsilon^2$
   - Likely to occur if CEF is non-linear, e.g. linear probability models
   - Conventional SEs are likely to be too small in this case (we overreject), but not a big difference in practice
   - Fix: robust standard errors.
2. **Serial correlation**: $E[\varepsilon_t \varepsilon_s] \neq 0$ for $s \neq t$
   - Occurs in time-series applications: $Y_t = \alpha + \beta X_t + \varepsilon_t$
   - Durbin-Watson test diagnoses serial correlation - look if DW statistic is close to 2 when testing if $\rho = 0$

- Fix 1: assume an AR(1) process and do quasi-differencing
  - Assume $\varepsilon_t = \rho\varepsilon_{t-1} + \nu_t$
  - quasi-differenced the model $Y_t - \rho Y_{t-1} = \alpha(1-\rho) + \beta(X_t - \rho X_{t-1}) + \nu_t$. ν_t is not serially correlated by assumption.
  - In practice, you use an estimate of ρ, $\hat{\rho}$
- Fix 2: Newey-West (HAC) SEs allow for autocorrelation + heteroskedasticity

3. **Clustering**:

- Occurs when data has a group structure: $Y_{ig} = \beta_0 + \beta_1 x_g + \varepsilon_{ig}$
- Residuals are correlated withing clusters: $E[\varepsilon_{ig}\varepsilon_{jg}] = \rho\sigma_\varepsilon^2 > 0$
- Each observation does not carry as much information as in the iid case since other observations from the same group already provides similar information. To reduce sampling variance, it is more efficient to sample more groups than more observations within each group.
- Fix 1: Moulton correction (multiply conventional SE by Moulton factor. If regressor constant withing group and equally sized groups, is $\sqrt{1+(n-1)\rho}$)
- Fix 2: group means using collapsed data, weighting by group size
- Fix 3: clustered SEs at the group level (may be unreliable with few clusters, need at least 20-30 clusters)

2.3 Measurement error

Measurement error in the regressor happens when you observe X_i^* instead of X_i :

$$X_i = X_i^* + u_i$$

The measurement error is classic when

- $E[u_i] = 0$
- $Cov(X_i^*, u_i) = Cov(u_i, \epsilon_i) = 0$

Main facts about classical measurement error:

- measurement error biases the coefficient toward zero (it cannot change its sign!). This is called attenuation bias.
  - $\beta^* = \beta r$ where $r = \frac{Var(X_i^*)}{Var(X_i^*)+Var(u_i)} < 1$
- Measurement error gets worse when
  - controls are added to the regression
  - fixed effects are added to the the regression
- **Intuition** : controls and fixed effects absorb parts of the "true" variation in X , leaving us with more noise (reduce variance of the signal in X^*, while leaving variance of noise unchanged).
- What's the fix? IV

3 Interpreting regressions

3.1 R-squared

R^2 is a measure of “goodness of fit”: how much of the variance in our dependent variable (Y) is explained by the right-hand side variables (the X’s)’:

$$R^2 = \frac{s^2_{\hat{Y}}}{s^2_Y} = \frac{\text{“explained variance”}}{\text{total variance}}$$

A high R^2 is good for prediction. Nothing to do with causality. Remember that we denote with $\hat{Y}$ the fitted values:

- Population fitted values: $\hat{Y}_i = \alpha + \beta X_i$, they satisfy $E[\hat{Y}_i \epsilon_i] = 0$
- Estimated fitted values: $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$, they satisfy $\sum_i^n \hat{Y}_i e_i = 0$

3.2 Testing hypothesis (under classical assumptions)

For **single hypothesis** $H_0 : \beta = \beta_0$, use a t-test:

$$T_n = \frac{\hat{\beta} - \beta_0}{\sigma_\varepsilon / \sqrt{n} s_X} \sim t(n-2)$$

For **multiple hypothesis or linear restrictions**, use an F-test:

$$F = \frac{n-k-1}{q} \frac{s^2_R - s^2_U}{s^2_U} \sim F_{(q,n-k-1)}$$

where

- n is the number of observations in the sample
- k is the number of regressors
- q is the number of linear restrictions
- $s^2_R = \sum \hat{\epsilon}_i^{R2}$, the residuals sum of squares of the restricted model
- $s^2_U = \sum \hat{\epsilon}_i^2$, the residuals sum of squares of the unrestricted model

The F-statistic can also be written in terms of R-squared:

$$F = \frac{n-k-1}{q} \frac{R^2_U - R^2_R}{1 - R^2_U} \sim F_{(q,n-k-1)}$$

4 Causal Effects

We are typically interested in measuring different kinds of treatment effects:

- $ATE = E[Y_i(1) - Y_i(0)]$
- $ATT = E[Y_i(1) - Y_i(0) | D_i = 1]$

- $LATE = E[Y_i(1) - Y_i(0)|D_i(Z_i = 1) - D_i(Z = 0) = 1]$

To avoid OVB problems, the easiest way is to randomize the variable of interest D_i (also referred as treatment)

- **Randomization**:
  - avoids **selection bias** (and OVB more generally) by yielding comparable groups along both observable *and unobservable* dimensions.
  - ensures the regressor of interest is truly exogenous to the model
- **What if we can't randomize?**
  - We discussed 3 other methods to get at causality:
    1. Instrumental Variables (IV)
    2. Differences-in-Differences (DID of just DD)
    3. Regression Discontinuity (RD)

5 Instrumental Variables (IV)

Intuition: IV only uses the part of the variation in the regressor of interest that is caused by some variable exogenous to the model (the "instrument")

Main identification assumptions:

1. **Relevance :** $Cov(Z_i, D_i) \neq 0$. Z predicts some part of D_i
2. **Exclusion Restriction:** $cov(Z_i, \epsilon_i) = 0$. Z is only correlated with Y through D

IV lingo:

$$D_i = \beta_0 + \beta_1 Z_i + \beta_2 X_{2i} + \epsilon_{1i} \quad \text{"First stage"} \tag{1}$$

$$Y_i = \beta_0 + \beta_1 Z_i + \beta_2 X_{2i} + \epsilon_{1i} \quad \text{"Reduced form"} \tag{2}$$

- $\beta_{IV} = \frac{Cov(Y_i,Z_i)}{Cov(D_i,Z_i)} = \frac{ReducedForm}{Firststage}$
- When the instrument is a dummy, the IV estimator is sometimes called the Wald estimator, $\hat{\beta}_{IV} = \frac{\bar{Y}_1 - \bar{Y}_0}{\bar{D}_1 - \bar{D}_0}$

Regression estimation via 2SLS:

$$Y_i = \alpha_0 + \alpha_1 \hat{D}_i + \alpha_2 X_{2i} + \epsilon_{2i} \quad \text{"Second stage"} \tag{3}$$

- Z_i is the instrument – it is **excluded** from the second stage (the controls, X_{2i}, are not)
- Remember $\beta_{2SLS} = \frac{Cov(Y_i,\hat{D}_i)}{Var(\hat{D}_i)} = ... = \frac{Cov(Y_i,Z_i)}{Cov(D_i,Z_i)} = \beta_{IV}$
- the asymptotic variance of the 2sls estimator is $\frac{\sigma_\epsilon^2}{n\sigma_{\hat{D}}^2} = \frac{\sigma_\epsilon^2/\sigma_D^2}{nR_f^2}$ where R_f^2 is the first stage R^2 (if you want the asymptotic standard error, just take the square roots of that). Thus, a more predictive first stage decreases the std error of the 2sls estimate.

- Doing 2sls "manually" (ols on second stage, after substituting fitted values) will give you the wrong s.e. (because it estimates the variance of $\epsilon_2 + \beta_{2SLS}(D_i - \hat{D}_i)$ rather than just $\sigma^2_{\epsilon_2}$)
- In a world of heterogeneous potential outcomes/heterogeneous treatment effects, 2sls estimates the LATE (local average treatment effect) for compliers of the instrument. Compliers are individuals who take the treatment only when the instrument assigns them to it (different from always takers or never takers). Different instruments can have different complier groups and thus lead to different 2sls estimates. (If few always takers, LATE close to ATT).

6 Difference-in-Differences

To implement DD, we need to observe a treated and an untreated group (not necessarily randomized) over several periods of time.

In DD, we measure the "treatment effect" by comparing the change in the treatment group with the change in the untreated ("control") group over time.

- Thus DD "nets out" permanent differences between the treatment and the comparison groups, as well as differences that can be attributed to time.

Set up

- two groups (group 1 or 0) → $g_i \in \{0, 1\}$,
- two time periods (period 1 or 0) → $period_t \in \{0, 1\}$,
- in period 1, only group 1 is subject to a policy.

Estimation:

$$y_{st} = \beta post_t + \gamma treat_s + \delta post_t \times treat_s + \varepsilon_{st} \tag{4}$$

- β captures any preexisting differences between the two groups,
- γ tells us whether anything else changed between pre- and post- period
- δ is the treatment effect we want to measure–the effect of the policy.

Generalizing to many states and periods is straightforward using a TWFE (Two-Way Fixed Effects) model:

$$y_{st} = \beta_s + \gamma_t + \delta TREAT_{st} + \varepsilon_{st} \tag{5}$$

where β_s are state effects and γ_t are time effects.

Main DiD assumption: Parallel trends - outcomes in both groups would have followed the same trend in the absence of the treatment.

Careful with SE's:

- typically DiD models are group regressions, where all the individuals in a region/place are treated simultaneously. So don't forget to cluster the standard errors!

7 Event Studies

Event studies generalize the DD logic to more than two time periods. This involves a bit of notation but has three main advantages:

- Provides a way to look for pre-trends
- Allows us to measure dynamic treatment effects

To accomodate staggered treatment (e.g. different states lower MLDA in different years), we work in event time (time relative to when the group is treated). The event study model looks like:

$$Y_{st} = \sum_{j=-m}^{-2} \tau_j \Delta D_{st-j} + \sum_{j=0}^{q} \tau_j D_{st-j} + \gamma_s + \lambda_t + \eta_{st} \tag{6}$$

where ΔD_{st-j} is an indicator that equals one for state s in the year j years after state s is treated. The index j here is "event-time", the number of years since treatment. This setup allows us to estimate a different TE in each period (τ_j).

- We omit the effect in period -1, the period right before treatment. The other TE's are measured relative to this period.
- The first summation corresponds to *leading effects*, differences between treatment and control units in the periods leading up to treatment. Leads close to zero support the parallel trends assumption.
- The second summation corresponds to *lagged effects*, allowing us to see how treatment effects evolve in the periods after treatment.

Intuitively, think of event study models as implementing differences-in-differences for each event-time period j, where the post period is the period for each state corresponding to event-horizon j and the pre period is the period right before treatment ($j = 1$). (See recitation 13 notes)

8 Regression Discontinuity

see RD handout

- **Intuition:** around arbitrary policy cutoffs or thresholds, individuals are extremely similar to each other, but some individuals are exposed to a treatment while others are not
- **Main identification assumption: Continuity** of potential outcomes (observables and unobservables) at the cutoff. The only thing that changes as the cutoff affecting outcome is the treatment status.

$$\lim_{x \downarrow c} E[Y_i(0)|x_i] = \lim_{x \uparrow c} E[Y_i(0)|x_i] \tag{7}$$

- **Sharp RD vs Fuzzy RD**
  - Sharp RD: deterministic treatment, perfect function of your running variable (everyone above a certain cutoff is treated, everyone below is not)
  - Fuzzy RD: treatment is not deterministically determined by the threshold-crossing rule, but treatment intensity or treatment probability changes discontinuously at the cutoff

- **Sharp RD - Regression estimation:**

  $$Y_i = \alpha + \beta Treat_i + f(x_i) + \varepsilon_i \tag{8}$$

  where $Treat_i = 1(x_i >$ some cutoff $c)$ and $f(.)$ is typically a polynomial in x_i

- **Fuzzy RD - Regression estimation:** Implemented as 2SLS!

  $$Y_i = \alpha + \beta Treat_i + f(x_i) + \varepsilon_i \quad \text{“Second stage”} \tag{9}$$
  $$Treat_i = \gamma + \delta Z_i + g(x_i) + u_i \quad \text{“First stage”} \tag{10}$$

  where the instrument is $Z_i = 1(x_i > c)$

- Additional comments:
  - Often slopes of the running variable polynomial allowed to be different on either side of the cutoff
  - Typically only use observations close enough to cutoff (in the “bandwidth”) - there are methods to choose bandwidth “optimally” (typically to minimize MSE)
  - Can also weight observations closer to c differently than those further away (kernel-weighting)

- **What can mess up RD?**
  - Purposeful sorting around the cutoff will do: when agents strive to avoid or cross the threshold, $E[Y_i(0)|x_i]$ is unlikely to be continuous around c.
  - for instance, families might choose in which side of a school district border to live depending on their interest in education and the difference in quality of the two school districts.
  - Typically test for this by looking at: covariate balance tests on both sides of cutoff, density smoothness (McCrary test)