Statistics Cheat Sheet

Data Sources

**Observational Studies:**

* Panel data (multiple N, multiple t)
* Cross sectional data (large N, one t)

**Quasi Experiments / Natural Experiment**

* Diff-in-Diff
* Regression Discontinuity
* Instrumental Variables

**Experiments**

* Randomized Controlled Trials

**Basics**

**Mean:**

**Variance:**

**Standard deviation:**

**Corrleation coefficient:**

**T-Test:**

D= difference between values

Ordinary Least Squares

**Measures of Fit**

* R2: fraction of variance of explained by
* Standard Error of Regression (SER): average distance btw values and reg line:
* Root Mean Square Error (RMSE): average error

**Homoskedasticity**

* var(u | X) is constant
* Assumption: E[u | X] = 0

**Heteroskedasticity**

* var(u | X) not constant. u = u(X)
* Assumption: E[u | X] = 0
* Standard Error too small if not robust

**Omitted Variable Bias**

* Z is determinant of Y (Z part of u)
* Z correlated with regressor X

**Assumptions**

* unbiased estimator,
* are i.i.d
* X and Y have finite fourth moments

Multiple Linear Regression Model

**Measures of fit**

* R2, adjusted R2: penalizes R2 when too many X, show when data overfitted

**Multicollinearity**

* problem

**Assumptions**

* unbiased estimator, E[u | X1,…,Xn] = 0
* (X1i,…,Xni, Yi) are i.i.d
* Large outliers are rare
* No perfect multicollinearity
* Non-Lin. Regression Functions

**Polynomials**

* give table Delta Y for Delta X

**Interpretation of Coefficients:**

* **Lin-Log**: a 1% change in X associated with 0.01 change in Y
* **Log-Lin**: a unit change in X associated with 100 % change in Y
* **Log-Log**: a 1% change in X associated with % change in Y

**Interaction between Independent Variables**

* Binary – continuous: create one regression line per group
* binary – binary: different slope for each dummy
* continuous – continuous:

Linear Probability Model

* Very simple to interpret

**Disadvantage**

* predicted probabilities >1 or <0
* assumption that

Probit Regression

**Advantage**

* bounded probability and

**Interpretation of Coefficients**

* is the change in the z-value of unit change in X
* = z-value
* To get probabilities evaluate z in cumulative standard normal distribution

**Measures of Fit**

* pseudo-R2: improvement in value of log likelihood relative to having no X

Logit Regression

* Same advantage as Probit
* Same interpretation of coefficient but evaluate z in logistic distribution
* Coefficients are odd rations

Validity

**Internal Validity,**

* OVB
* Simultaneous Causality Bias
* Wrong functional form
* Errors in variable bias
* Sample selection bias

**External Validity**

* Generalization of data to other time
* to other country, urban area?

Panel Regression

* contains observation on multiple entities at two or more points in time
* balanced panel: have data for each entity for each time

**Fixed Differences**

* two time periods, unobserved variable Z can be controlled for

**Fixed Effects**

* Add constant shift in intercept for each entity/time
* One model 🡪 one slope

**Entity Fixed Effects**

* Same slope for all entities, different intercepts
* Control for OV which varies across entities but not over time
* **Assumption**:

**Time Fixed Effects**

* Control for OV which varies over time but not across entities

**Assumptions**

* are i.i.d
* have finite fourth moments (same as OLS)
* No perfect multicollinearity (same as OLS)

**Autocorrelation**

* data is i.i.d across clusters but not within
* for
* Use clustered standard errors (assume variables are not i.i.d within entities)

**Advantages**

* Control for factors that vary across entities but not over time
* Control for unobserved and unmeasured variables
* Omitted variable does not change over time
* More observation gives more information

**Limitations and Challenges**

* Time lag effects can be important
* Need to use clustered standard errors
* unobserved variable determinant of Y but uncorrelated with X
* unobserved variable varies across entities and over time
* Data collection issues, non-response

**Random Effect Regression**

* if OV random and uncorrelated with regressors
* if OV time invariant and random
* **Assumption**:
* Hausman Test to decide if random or fixed effects

Instrumental Variable Regression

* breaks X into two parts, one correlated with u, one not. Uncorrelated part is IV called Zi.
* Endogeneity: variable correlated with u
* Exogeneity: variable uncorrelated with u

**Condition for valid Instruments**

* **Relevance**:
  + - at least one must be relevant
* **Exogeneity**: (Exclusion Restriction Principle)
  + - all must be exogenous

**Two Stage Least Squares**

* First stage: regress X on the IV Z
* Second stage: regress Y on the estimated X
* Include control variables W in both steps
* Endogenous coefficient X is:
  + - m IV, k endogenous variables
    - over-identified: exactly identified: under-identified:

**Checking Instrument Validity**

* Relevance: at least one pi is nonzero
* Weak instruments:
  + - all pi zero or close to zero
    - with weak instruments, 2SLS can be biased in direction of OLS estimator
    - check: compute F statistic (>10) drop weakest
* Exogeneity: only poss. if m>k, do J-test

**Assumptions**

* E[u | W1i,…, Wri] = 0 (exogenous regressors are exogenous)
* (Yi, X1i,…, Xki, W1i,…, Wri, Z1i,…, Zmi) are i.i.d
* (X, W, Z, Y) have finite fourth moments
* The instruments (Z1i,…, Zmi) are valid

Difference in Differences

**Comparison Group**

* Quality of comparison group determines quality of policy evolution
* Counterfactual: what would have happened to same people if policy not implemented

**Diff-in-Diff Estimator**

* difference between two before after differences
* Treatment effect isolable 🡪 Causality

**Weakness**

* Non-random treatment
* biased estimation if other determinant of jump than policy
* Can never really know counterfactual

**Assumption**:

* Common trend (also parallel trends)
* Special cast of panel data, use clustered SE because of autocorrelation

**Test Common Trend Assumption**

* Placebo DD with fake treatment group (0 effect)
* Placebo DD with different outcome var (0 effect)
* Different comparison group (find same results)

Randomized Controlled Trial

**Measurement error: precision**

* Increase sample size to get rid of it

**Systematic error: accuracy (bias)**

* get better comparison grp (close to treatment grp)

**Main Idea**

* Treatment has causal effect on person
* Treatment X randomly assigned, so independent of u 🡪 is unbiased
* No OVB as X randomly assigned, independent of any W
* Having baseline (W) still increases precision

**Mechanisms of Randomization**

* **Pure**:(list of participants, computer)
* **Systematic**:(dice)
* **Oversubscription**:(take first who show/sign up)
* **Pipeline**:(all get treatment, randomize when)
* **Encouragement**:(Discount, when ethical hazards)
  + - Run IV Reg. with getting encouragement as IV
* Think of which unit of randomization! 🡪 cluster SE

**Challenges with RCT**

* Ethical concerns (vaccines)
* focus on programs easier to measure?

**Remaining Threats Internal Validity**

* Does the study provides unbiased estimate?
* Partial Compliance (fail to follow treatment protocol)
* Attrition (subject dropping out of study)
* Experimental effects (Experimenter bias)
* Spillover effects (Positive or Negative)
* Small Samples

**Remaining Threats External Validity**

* Can the study be generalized?
* Non representative sample (diff. btw. population)
* Non representative treatment (small-scale well monitored to large scale)
* General Equilibrium Effects (small experiment to large permanent changes economic environment)

Regression Discontinuity

* Impact evaluation method

**Conditions/Assumptions**

* Need continuous eligibility index W and clearly defined threshold w0.
* Eligibility index must be continuous
* Cutoff must be unique to the program
* Only driver of having the treatment is W score.

**Main Idea**

* Compare people just above and under threshold
* Treatment effect is difference around threshold
* Effect of treatment shown as jump in Y
* No need for control group
* W called running variable

**Sharp RD Design**

* Everyone above threshold gets treatment
* Interaction term allows for having two different curves left and right of threshold
* No OVB by definition, running variable determinant of getting treatment or not.

**Fuzzy RD Design**

* Crossing threshold changes probability to get treatment
* IV Regression with probability as IV

**Challenges and Limitations**

* Local average Treatment Effect
  + - estimation around threshold point not always generalizable (not externally valid)
* Statistical Power
  + - effect estimated at discontinuity, fewer observations than in experiment with same sample size
* Sensitivity to functional form
  + - jump might be simply due to nonlinear functional form

**Robustness Checks**

* Functional Form (include polynomials)
* Statistical Power (change bandwidth)
* Placebo RD with other threshold (no jump)
* Placebo RD with other outcome var (no jump)
* Placebo RD with fake treatment group (no jump)
* Check for manipulation of data (plot)

**Multiple Choice:**

**Linear regression**

* Binary variables can take on only two values
* In the simple linear regression model, the regression slope indicates by how many units Y increases, given a one unit increase in X.
* The OLS estimator is derived by minimizing the sum of squared residuals.
* The slope estimator, β1, has a smaller standard error, other things equal, if there is more variation in the explanatory variable, X.
* When the estimated slope coefficient in the simple regression model, b1, is zero, then R2 = 0

**Non-linear regression**

* The interpretation of the slope coefficient in the model Yi = β0 + β1 ln(Xi) + ui is as follows a 1% change in X is associated with a change in Y of 0.01 β1.
* The interpretation of the slope coefficient in the model ln(Yi) = β0 + β1Xi + ui is as follows a change in X by one unit is associated with a 100 β1 % change in Y.
* The interpretation of the slope coefficient in the model ln(Yi) = β0 + β1 ln(Xi)+ ui is as follows a 1% change in X is associated with a β1 % change in Y
* The best way to interpret polynomial regressions is to plot the estimated regression function and to calculate the estimated effect on Y associated with a change in X for one or more values of X.
* The binary variable interaction regression allows the effect of changing one of the binary independent variables to depend on the value of the other binary variable

**Regression with binary dependent variable**

* The linear probability model is the application of the linear multiple regression model to a binary dependent variable.
* The probit model forces the predicted values to lie between 0 and 1.
* In the probit regression, the coefficient β1 indicates the change in the z- value associated with a unit change in X.
* To measure the fit of the probit model, you should use the fraction correctly predicted or the pseudo R2
* When testing joint hypothesis, you can use either the F-statistic or the chi-square statistic.

**Panel Regression**

* The Fixed Effects regression model has n different intercepts.
* In the Fixed Time Effects regression model, you should exclude one of the binary variables for the time periods when an intercept is present in the equation to avoid perfect multicollinearity
* In panel data, the regression error is likely to be correlated over time within an entity.
* It is advisable to use clustered standard errors in panel regressions because the fixed effects estimator is asymptotically normally distributed when n is large.
* When you add state fixed effects to a simple regression model for U.S. states over a certain time period, then state fixed effects account for a large amount of the variation in the data

**IV Regression**

* The rule-of-thumb for checking for weak instruments is as follows: for the case of a single endogenous regressor a first stage F < 10 indicates that the instruments are weak.
* The distinction between endogenous and exogenous variables is whether or not the variables are correlated with the error term.
* Let W be the included exogenous variables in a regression function that also has endogenous regressors (X). The W variables can all of the above
* The logic of control variables in IV regressions parallels the logic of control variables in OLS
* The IV estimator can be used to potentially eliminate bias resulting from errors in variables