STATS 209: What have we learned so far?

Small Recap

March 10, 2019

1 Week 1: Properties of regression models

Main takeouts from the week:

- Multilevel regression coefficients have to be interpreted as the effect of a given variable X on the outcome Y once the other variables have been accounted for. This makes these coefficients not very interpretable (ex: MOMED regressed out on all the other variables...)
- It is usually not recommended to standardize the variables in linear regression as the recovered coefficients depend on the population, and thus we can loose "general laws".
- Measurement errors make the estimated regression slopes less steep and can potentially introduce the appearance of a difference between two populations.

In more details:

- Notations: In everything that follows:
  - β_{YX} denote the coefficient of the one-predictor linear regression of Y against X: $Y = \beta_{YX}X + \alpha$
  - $\beta_{YX_1 \cdot X_2} = \beta_{Y(X_1 \cdot X_2)} = \beta_{(Y \cdot X_2)(X_1 \cdot X_2)}$ is the coefficient of the one-predictor linear regression of Y against the residuals $X_1 - \beta_{X_1X_2}X_2$. This quantifies in a way the "innovation" that X_1 brings for explaining Y, once that X_2 has been accounted for.
- **Regression Recursion**
  - **Fundamentals of linear regression**:
    - Simple regression:

$$\boxed{\beta_{YX} = \frac{Cov(X,Y)}{Var(X)} = r_{XY}\sqrt{\frac{Var(Y)}{Var(X)}}}$$

where $r_{XY} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$ is the correlation between X and Y.

* The formula for the estimated coefficients is:

$$\boxed{\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}}$$

- Reminder: properties of the covariance:

$$\begin{aligned}
Cov(A + B, C) &= Cov(A, C) + Cov(B, C) \\
Cov(A, B + C) &= Cov(A, B) + Cov(A, C) \\
Cov(\lambda A, B) &= Cov(A, \lambda B) = \lambda Cov(A, B) \\
Var(A) &= Cov(A, A) \\
Var(A + B) &= Var(A) + Var(B) + 2Cov(A, B) \\
Var(A - B) &= Var(A) + Var(B) - 2Cov(A, B)
\end{aligned}$$

- Simple linear regression model:

$$Y = \beta X + \alpha$$

where $\beta = \frac{Cov(X,Y)}{Var(X)}$ and the coefficients are estimated as:

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{and} \quad \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

- In multivariate regression: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$ In this case, the coefficients β_j capture the effect of variable X_j once the other variables have been accounted for. Using the notation of the handouts:

$$\beta_1 = \beta_{Y(X_1 \perp X_{\setminus X_1})} = \beta_{Y(X_1 \cdot X_2)}$$

with $X_1 \cdot X_2 = X_1 - \beta_{12} X_2$

- **Regression Recursion formula** Simple regression with two predictors:

$$\beta_{12} = \beta_{12 \cdot 3} + \beta_{32}\beta_{13 \cdot 2}$$

- **"Woes of regression"**: coefficients are thus difficult to interpret: the values of the coefficients depend crucially on what else is included in the linear fit. The dream is always to say that β_j represents the effect of a unit increase in variable X_j on the outcome Y. Note that this is a mathematical "fantasy", that is oftentimes difficult to translate in real life. Regression is good for fitting a model, but one has to be careful in the way to interpret the results.

- It is thus necessary to make a distinction between controlling and conditioning on other variables.

- **Error in variables**: Suppose we observe a corrupted version $X = \xi + \epsilon$ of a true measurement ξ. The linear regression model with one predictor becomes:

$$\begin{cases} \mathbb{E}[Y|\xi] = \beta_0 + \beta_1 \xi \\ \mathbb{E}[Y|X] = \gamma_0 + \gamma_1 X \end{cases}$$

where $\gamma_1 = \beta_1 R_x = \frac{Var(\xi)}{Var(X)}\beta_1$. $R_x = \frac{Var(\xi)}{Var(X)}$ is called the reliability.
Proof:

$$\gamma_1 = \frac{Cov(X,Y)}{Var(X)} = \frac{Cov(\xi+\epsilon, Y)}{Var(X)} = \frac{\overbrace{Cov(\xi,Y)/Var(\xi)}^{=\beta_1} \times Var(\xi) + \overbrace{Cov(\epsilon, y)}^{=0 \text{ by independence of } \epsilon \text{ and } Y}}{Var(X)}$$

$$\implies \gamma_1 = \beta_1 \frac{Var(\xi)}{Var(X)}$$

If we have two predictors (that is we observe $X_1 = \xi_1 + \epsilon_1$ and $X_2 = \xi_2 + \epsilon_2$), the formula becomes a little more involved:

$$\gamma_1 = \frac{\beta_1 R_1(1 - \rho^2 R_2) + \beta_2 \beta_{\xi_2 \xi_1} R_1(1 - R_2)}{1 - \rho^2 R_1 R_2}$$

(this can be derived in a similar fashion than before, using the recursion formula).

- **Effect of measurement errors on the regression coefficients:** measurement errors have thus the tendency to decrease the regression slopes. In certain studies (ex: years of education vs income), the study of two groups of population (ex: low income vs high income) can even produce the impression of a bias (i.e, higher intercept at 0 years of education for the high-income group)

- **Effect of standardizing the variables**. One could wish to standardize variables before running linear regression. This allows to put every variable on the same scale, and to make claims as to which variable is more important (i.e, causes a more drastic change in amplitude) for the response Y. The standardized coefficients can be obtained from the original ones via the formula:

$$\beta_{YX}^{(standardized)} = \frac{std(X)}{std(Y)}\beta_{YX}$$

This is in fact a relatively poor practice, since we are introducing population variance in the value of $\beta_{YX}^{(standardized)}$, and thus looses generalizability.

2 Week 2: Association vs Causation; Experiments vs observational studies; Neyman-Rubin-Holland formulation

Main takeouts from the week:

- **Spurious correlations:** they are more common than you think!! It is thus important to account for the existence of potential spurious correlations when looking for correlations between variables.
- **Design is better than analysis:** Randomized experiments are the gold standard in Causal Inference.
- Unfortunately, since we cannot only do clinical trials, we also have to consider observational studies. In these, we can make multiple corrections (ex: age, Socio-Economic Status, etc), to account for possible confounders. It seems that whatever we do though, the estimates of the coefficients (regression, odds ratio, etc) that we get are biased with respect to the ones from randomized experiments (cf Women Hormone paper), so this type of "correction" remains inherently limited.
- **Mediator Variables**: identify why and how the treatment works (e.g, treatment acts on a certain neurotransmitter, etc.).
- **Moderator Variables**: specify for whom or under what conditions the treatment works.

In more details:

- **Third Variables and spurious correlations** Consider X_1, X_2, X_3 3 variables.

$$r_{13\cdot 2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{(1-r_{12}^2)(1-r_{23}^2)}}$$

$$R^2_{Y\cdot X_1X_2} = r^2_{YX_1} + r^2_{Y(X_2\cdot X_1)}$$

- **Simpson's paradox:** Simpson's paradox (or Simpson's reversal, YuleSimpson effect, amalgamation paradox, or reversal paradox), is a phenomenon in probability and statistics, in which a trend appears in several different groups of data but disappears or reverses when these groups are combined.
- **Neyman-Rubin-Holland Formulation**. Ref: Chap 19 in Rubin's Causal Inference in Retrospective studies.

$$\begin{cases} U &= \text{Units in population} \\ K &= \text{Causes (treatment)} \\ S &= \text{Assignement Rule} \end{cases}$$

A little bit of notations:

- $S(u)$ denotes the actual exposure of unit u to the treatment, and we observe $Y(u, S(u))$
- Unit-level Causal effect $Y(u,t) - Y(u,c)$
- Unit homogeneity: $Y(u,s) = Y(v,s) \forall s \in K, u, v \in U$
- Causal effect: $T_{tc}(u) = Y_t(u) - Y_c(u)$
- Average Causal Effect $ACE_{tc} = \mathbb{E}[T_{tc}] = \mathbb{E}[Y_t] - \mathbb{E}[Y_c]$
- We in fact only observe the FACE: $FACE_{tc}(Y) = \mathbb{E}[Y_t|S = t] - \mathbb{E}[Y_c|S = c]$. Randomization makes it possible to have the assignment independent of Y, so that $FACE = ACE$
- The two main assumptions to get an unbiased estimator of the causal effect are:
  * SUTVA: stable unit treatment value assumption : assumes the treatment status of any unit does not affect the potential outcomes of the other units (non-interference) and that the treatments for all units are comparable (no variation in treatment).
  * Ignorability/ Unconfoundedness: $(Y_t, Y_c) \perp T$, that is, the treatment assignment is independent of the outcome.

  Potential problems arise when considering the actual compliance of subjects to their treatment assignments (Never-taker: Unit never takes treatment, always-takers, defier: Unit takes treatment when not assigned and control when assigned ...). This can introduce bias in the estimates of the Average Causal Effect.

- Mediation/ Moderation: moderators identify on whom and under what circumstances treatments have different effects. Mediators identify why and how treatments have effects.

  - **Moderator**: Treatments moderators specify for whom or under what conditions the treatment works. They may identify subpopulations with possibly different causal mechanisms or course of illness.

    "In general terms, a moderator is a qualitative (e.g., sex, race, class) or quantitative (e.g., level of reward) variable that affects the direction and/or strength of the relation between an independent or predictor variable and a dependent or criterion variable. Specifically within a correlational analysis framework, a moderator is a third variable that affects the zero-order correlation between two other variables. ... In the more familiar analysis of variance (ANOVA) terms, a basic moderator effect can be represented as an interaction between a focal independent variable and a factor that specifies the appropriate conditions for its operation." p. 1174 (Kenny).

  - **Mediators:** "In general, a given variable may be said to function as a mediator to the extent that it accounts for the relation between the predictor and the criterion. Mediators explain how external physical events take on internal psychological significance. Whereas moderator variables specify when certain effects

will hold, mediators speak to how or why such effects occur." p. 1176.

To show that M is a mediator of treatment, M would have to measure and event or change occurring during the treatment ad then it must correlate with treatment choice, hence possibly be a result of treatment and have either a main or interactive effect on the outcome.

- **Example: difference between moderator and mediator variables**: a moderator variable is one that influences the strength of a relationship between two other variables, and a mediator variable is one that explains the relationship between the two other variables. As an example, let's consider the relation between social class (SES) and frequency of breast self-exams (BSE). Age might be a moderator variable, in that the relation between SES and BSE could be stronger for older women and less strong or nonexistent for younger women. Education might be a mediator variable in that it explains why there is a relation between SES and BSE. When you remove the effect of education, the relation between SES and BSE disappears. [1]

[1] `http://psych.wisc.edu/henriques/mediator.html`

3 Week 3: Path analysis and causal modeling

Main takeouts from the week:

- Path analysis is inherently limited. Sounds like a risky road to take from a statistical perspective, although people publishing studies in papers seem to thoroughly enjoy them...
- Basically, when studying the relationship between two variables in path analysis, there are two types of effects: the direct effect and the indirect effect.
- Two main estimators to estimate the indirect effect :
  - Wald estimator: assumes direct effect τ is 0 (see figure 1). Hence:

$$\hat{\beta} = \frac{\bar{Y}_t - \bar{Y}_C}{\bar{R}_t - \bar{R}_C}$$

  - ALICE estimator: takes into account the τ which can be potentially non zero, but has to assume a simple formula for the aggregation biased. Seems to be a less robust and justifiable assumption than the Wald estimator. Hence:

$$\hat{\beta} = \hat{\beta}_{Y(R \cdot G)} + \frac{\hat{\beta}_{Y(G \cdot R)}}{\hat{\beta}_{RG}}$$

In more details:

- **Formulation for Encouragement Designs: "Counterfactual Data"** (Holland 1988, p471)
  1. $R_t(u) - R_c(u) = \rho(u)$
  2. $Y_{Gr}(u) - Y_{Gr'}(u) = (r - r')\beta(u)$
  3. $Y_{tr}(u) - Y_{cr}(u) = \tau(u)$
  4. $Y_{tR_t}(u) - Y_{cR_c}(u) = \tau(u) + \rho(u)\beta(u)$ (sum of Direct + Indirect effect)

- ALICE framework: Additive Linear Constant Effect model: this is a causal theory/model with constant effects. At the individual level, this model is represented by the following diagram.
  1. $R_t(u) - R_c(u) = \rho$
  2. $Y_{Gr}(u) - Y_{Gr'}(u) = (r - r')\beta$
  3. $Y_{tr}(u) - Y_{cr}(u) = \tau$

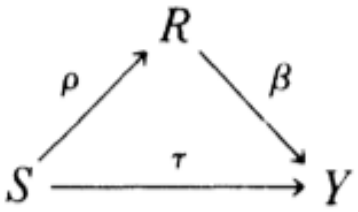


Figure 1: ALICE model at the individual level

4. In this setting, the parameters of the ALICE model (ρ, β and γ) may be used to express the four ACEs of the model
   - $ACE_{tc}(R) = \rho$
   - $ACE_{tc}(Y) = \tau + \beta\rho$
   - $ACE_{tc}(Y(\cdot, \cdot, r) = r$
   - $ACE_{tc}(Y(\cdot, s, \cdot) = r\beta(r - r')$
   - $FACE_{tc}(Y(\cdot, \cdot, r) = \tau + \mu_c(r - \rho) - \mu_c(r)$
5. In our model, every individual comes with his own baseline, $\mu_s(r) = \mathbb{E}[Y_{c0}|R_s = r]$ for$s = t, c$ is the average value of test scores for students when they are not encouraged to study and they do not study, for all students who would study an amount r when they are not encouraged to study. Thus $\mu_(r)$ is a counterfactual regression because Y_{c0} and R_s can never be simultaneously observed except when $R_s = 0$... (by design, we observe the student studying r units of time, and thus have no data for his baseline performance). Thus $\mu_c(r)$ us inherently inobservable, and has to be estimated...
6. This becomes particularly important at the group level (aggregated data), because we are aggregating all of these baseline performances, which introduce a bias in our estimation of β. Indeed, suppose that this (in fact complicated) quantity is in fact linear:
   $$\mu_c(r) = \gamma + \delta r$$
   (interpretation: people self-select into levels of studying).
   Positive δ: the more a student would study when no encouragement the higher he would score on the test without study or encouragement. So in fact, what we have is:
   $$\mathbb{E}[Y_S R_s|S, R_S] = \mu_c(R_S - \rho S) + \tau S + \beta R_S = \gamma + (\tau - \delta\rho)S + (\beta + \delta)R_S$$
7. So our estimate of both β and τ will be biased. The overall effect however will remain unbiased.

- Traditional Path Analysis

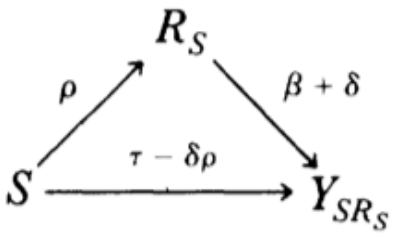


Figure 2: ALICE model at the group level

1. Solve normal equations to get association decompositions into Direct and Indirect effects:
2. Notre Dame example

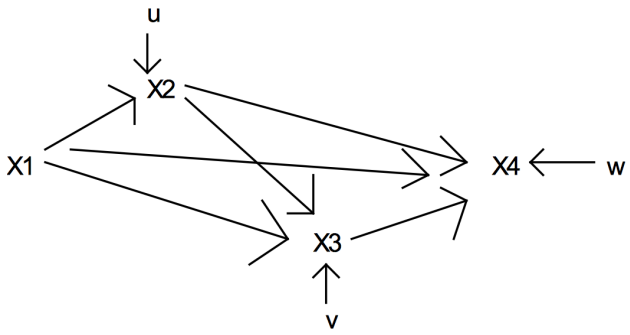


Figure 3: Example Path Diagram

Structural equations in the above diagram are:

$$X_2 = \beta_{21} X_1 + u$$

$$X_3 = \beta_{31} X_1 + \beta_{32} X_2 + v$$

$$X_4 = \beta_{41} X_1 + \beta_{42} X_2 + \beta_{43} X_3 + w$$

3. **Sewell Wrights multiplication rule:** To find the correlation between X_h and X_j, where X_j appears later in the model:
   (a) begin at X_j and read back to X_h along each distinct direct and indirect (compound) path, forming the product of the coefficients along that path. (This will give you the correlation between X_j and X_h that is due to the direct and indirect effects of X_h on X_j)
   (b) After reading back, read forward (if necessary), but only one reversal from back to forward is permitted. (This will give you correlation that is due to common causes.)
   (c) A double-headed arrow may be read either forward or backward, but you can only pass through 1 double-headed arrow on each transit. (This will give you correlation due to correlated causes)

(d) If you pass through a variable, you may not return to it on that transit.

(e) Sum the products obtained for all the linkages between X_j and X_h. (The main trick to using Wrights rule is to make sure you dont miss any linkages, count linkages twice, or make illegal double reversals.) This will give you the total correlation between the 2 variables.

4 Week 4: Group Comparisons and Causal Inference with Multilevel data: Contextual effects, aggregation bias, Mixed-effects (lmer) models

Main takeouts from the week:

- Aggregating data can create weird phenomena (ecological fallacy, aggregation bias, etc.) that hinders the accuracy of the study.
- Oftentimes, the data is organized in nested groups or clusters, the "contextual" effect of which has to be taken into account when performing the analysis.
- We now study data that can potentially belong to different groups/ clusters (i.e, in a way, data organized in groups or hierarchy).
- Important notion: in the context of this multilevel data analysis, we can model group effects via either fixed effect vs random effects
- The R-package lmer is the go-to package for fitting these types of models.

In more details...

- **Fixed/ Mixed/Random effects models:** Suppose we have N subjects, belonging to two groups (Seattle vs San Diego). We measure their level of happiness Y_{ijt} through time vs their chocolate consumption X_{ijt}.
  - Fixed effect: model parameters are fixed, non random quantities, e.g:
  $$Y_{ijt} = \alpha_j + \beta_j X_{ijt}$$
  - Mixed effect: combination of random and fixed models parameters:
  $$Y_{ijt} = \alpha_{ij} + \beta_j X_{ijt}$$
  with the parameter β_j being fixed and non-random for the population, but the intercepts α_{ij} are random variables (subject effects)
  - Random effects: all the parameters are themselves random:
  $$Y_{ijt} = \alpha_{ij} + \beta_{ij} X_{ijt}$$
  , with random slopes and random intercepts for the different subjects.
- **Multilevel Data**: the idea here is that the data is "multiscale", and is typically stratified in different groups or something of the like. Notation: there are thus some

"between effects" (indexed by the upper symbol $(\cdot)^b$ and "within-group" (referred as "within-pooled") effects, the latter being indexed by the upper mark $(\cdot)^{w-p}$
Starting point:

$$Y_{ij} = \bar{Y}_{i\cdot} + (Y_{ij} - \bar{Y}_{i\cdot}) \quad X_{ij} = \bar{X}_{i\cdot} + (X_{ij} - \bar{X}_{i\cdot})$$

Basic levels of regression:

$$\beta^t_{YX} = n^2_X \beta^b_{\bar{Y}\bar{X}} + (1 - n^2_X)\beta^{w-p}_{YX}$$

where n_X is a grouping measure: $n_X = \frac{Var(\bar{X})}{Var(X)}$, β^t_{YX} is the coeff ignoring group assignments, $\beta^b_{\bar{Y}\bar{X}}$ is the group mean "between", and β^{w-p}_{YX} is the within pooling relative standing.
Note that the coefficient $\beta^b_{\bar{Y}\bar{X}}$ is not equal to $\beta_{\bar{Y}\bar{X}}$ obtained by OLS regression of $\bar{Y}$ onto $\bar{X}$: it is in fact a weighted regression (with the weights as the proportion of people in each cluster). Another simple way of obtaining this coefficient is by introducing additional columns $\bar{X}$ and $\bar{Y}$ in the individual level data, and directly running linear regression: the replicated entries in the table will take care of this weighting issue.

Define:

$$\begin{cases} \beta^b_{YX} = \frac{Cov(\mathbb{E}[X|U], \mathbb{E}[Y|U])}{Var(\mathbb{E}[X|U])} \\ \beta^{w-p}_{YX} = \frac{\mathbb{E}[Cov(X,Y|U]}{\mathbb{E}[Var(X|U)]} \\ n^2_x = \frac{\mathbb{E}[Var(X|U)]}{Var(x)} \end{cases}$$

Proof.

$$\begin{aligned} Var(X) &= \mathbb{E}[Var(X|U)] + Var(\mathbb{E}[X|U]) \quad \text{(fact from probability)} \\ Cov(X,Y) &= \mathbb{E}[Cov(X,Y|U)] + Cov(\mathbb{E}[X|U], \mathbb{E}[Y|U]) \quad \text{(total covariance formula)} \\ \beta = \frac{Cov(X,Y)}{Var(X)} &= \frac{\mathbb{E}[Cov(X,Y|U)] + Cov(\mathbb{E}[X|U], \mathbb{E}[Y|U]))}{Var(X)} \\ &= \mathbb{E}\Big[\frac{Cov(X,Y|U)}{\mathbb{E}[Var(X|U)]} \times \frac{\mathbb{E}[Var(X|U)]}{Var(X)}\Big] + \frac{Cov(\mathbb{E}[X|U], \mathbb{E}[Y|U]))}{Var(\mathbb{E}[X|U]))} \frac{Var(\mathbb{E}[X|U]))}{Var(X)} \\ &= \beta^{w-p}_{YX}(1 - n^2_x) + \beta^b_{YX} n^2_x \end{aligned}$$

- These allow to define some nice quantities:
  - **Aggregation bias:** $\beta^t_{YX} - \beta^b_{\bar{Y}\bar{X}}$
  - **Contextual effect:** $\beta^t_{Y\bar{X}\cdot X} = \beta^b_{\bar{Y}\bar{X}} - \beta^{w-p}_{\bar{Y}\bar{X}}$.
    Note that this can be proven using the regression recursion formula:

$$\beta^t_{Y\bar{X}\cdot X} = \beta_{Y\bar{X}} - \beta_{X\bar{X}}\beta_{YX\cdot\bar{X}}$$

    Now we know that $\beta_{Y\bar{X}} = \beta^b_{\bar{Y}\bar{X}}$ since: $\beta_{Y\bar{X}} = Cov(\bar{Y} + (Y - \bar{Y}), \bar{X}) = Cov(\bar{Y}, \bar{X})$ (because $(Y - \bar{Y}) \perp \bar{X}$). We also know that $\beta X\bar{X} = 1 (= Cov(\bar{X} + (X - \bar{X}), \bar{X}) = Cov(\bar{X}, \bar{X}))$

- **"Ecological inference"**– a conclusion about individual behavior drawn from data about aggregate behavior.

- **Ecological fallacy**: The "ecological fallacy" consists in thinking that relationships observed for groups necessarily hold for individuals: if countries with more Protestants tend to have higher suicide rates, then Protestants must be more likely to commit suicide.Statistical procedures have been proposed for disentangling individual-level from group-level behavior, including "ecological regression" and "cross-level" or "hierarchical" regression models. However, each method makes its own rather strong behavioral assumptions, which seem implausible when stated explicitly. For instance, ecological regression makes the "constancy assumption." According to this assumption, with an application like Durkheim's, individual behavior cannot depend on geographical location. Protestants all over Europe must have similar propensities to commit suicide; and Catholics are homogeneous too.

- Hence: danger of looking at aggregated data: it might not be reflective of individual trends if the assignment of the people in the groups is not random. The problem of confounding must be dealt with in any observational study. But the second problem is specific to ecological studies: putative causes and effects are measured for groups rather than individuals. If there is no confounding, the expected difference between effects for groups and for individuals is "aggregation bias"; in general, the difference is partly attributable to confounding and partly to aggregation bias.

- **"Smart first-year student"**: instead of running lmer, which computes the hierarchical model, a simple approach would be to run the regression groupwise, and compare the results.

5 Week 5.–The many uses and forms of analysis of covariance

Main takeouts from the week:

- The goal here is to draw conclusions as to the effect of a given treatment while controlling for additional covariates.
- Two main estimators: ANCOVA and CNRL (Comparing Nonparallel Regression Lines), which is basically a more flexible ANCOVA where the regression slopes are allowed to vary from group to group.
- Heterogeneous Treatment Effect
- Uses of ANCOVA with haphazard and with systematic assignment.
  - ANCOVA regression adjustments fail in observational studies
  - Non-random assignment on the basis of the covariate

Main takeouts

- ANCOVA: Analysis of Covariance: blends ANOVA and regression. ANCOVA evaluates whether the means of a dependent variable Y are equal across levels of a categorical independent variable G often called a treatment, while statistically controlling for the effects of other continuous variables that are not of primary interest, known as covariates (X) or nuisance variables. Mathematically, ANCOVA decomposes the variance of Y into variance explained by X, variance explained by the categorical treatment/group assignment (G), and residual variance. Intuitively, ANCOVA can be thought of as 'adjusting' the DV by the group means of the CV(s).
  This amounts to evaluating a model of the form:

  $$y_{ij} = \mu + \tau_i + \beta(x_{ij} - \bar{x}) + \epsilon_{ij}$$

  with y_{ij} the jth observation under the ith category, μ the grand man, $\bar{x}$ the global mean for covariate x. What we are really interested in is the effect of the treatment: τ_i
- Setting: suppose the data can be split into multiple clusters (for simplicity, let us first assume that we have 2 groups). Define $G_i = 1$ if unit i is in group 1 and 0 otherwise. Let us define:

  $$\mathbb{E}[Y|G=0] = \mu_0 \quad \mathbb{E}[Y|G=1] = \mu_1 \mathbb{E}[Y|G] = \beta_0 + \beta_1 G$$

  with $\beta_0 = \mu_0$ and $\beta_1 = \mu_1 - \mu_0$.
  The goal in ANCOVA is to fit two straight-line regressions within each group.

  $$\mathbb{E}[Y|X] = \alpha_1 + \gamma_1 X \quad \text{Group 1}$$

$$\mathbb{E}[Y|X] = \alpha_0 + \gamma_0 X \quad \text{Group 0}$$

such that: $\gamma_1 = \gamma_0$. If we fit a regression line to each group ($\hat{Y} = \bar{Y}_i + \hat{\gamma}_i(X - \bar{X}_i)$), then we can recover the common γ by:

$$\hat{\gamma}_p = \frac{\hat{\gamma}_1 SSX_1 + \hat{\gamma}_0 SSX_0}{SSX_1 + SSX_0}$$

This assumes that the regression lines are parallel within each group, and the only thing that is different is the intercept.

- In this context, since we are clustering, weird hybrids between moderator and mediators start appearing: Moderated mediation: individual differences in individuals (ex: mediation variable works differently for males and females for instance). Mediated moderation: individual responses to the moderation: why are there individual differences in response to a treatment.

- The purpose of including covariates in ANOVA is two-fold:

  - To reduce within-group error variance: In ANOVA we assess the effect of an experiment by comparing the amount of variability in the data that the experiment can explain, against the variability that it cannot explain. If we can explain some of this unexplained variance (SSR) in terms of covariates, then we reduce the error variance, allowing us to more accurately assess the effect of the experimental manipulation (SSM).
  - Elimination of Confounds: In any experiment, there may be unmeasured variables that confound the results (i.e. a variable that varies systematically with the experimental manipulation). If any variables are known to influence the dependent variable being measured, then ANCOVA is ideally suited to remove the bias of these variables. Once a possible confounding variable has been identified, it can be measured and entered into the analysis as a covariate

- Confidence bounds for regression:

  - Simultaneous Confidence bounds (Working Hotelling bounds): contains the regression line with probability $1 - \alpha$, $\forall x$: Ex: is given by, for m new points:

$$\hat{\beta}_1 x_{n+1} + \hat{\beta}_0 \pm \sqrt{F_{2,n-2}^{1-\alpha}} s \sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_X X}}$$

  - Non simultaneous ones/ Pointwise confidence bands: Ex: is given by, for average of m new points at x_{n+1}:

$$\hat{\beta}_1 x_{n+1} + \hat{\beta}_0 \pm t_{n-2}^{1-\alpha/2} s \sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_X X}}$$

- **CNRL**: More general model than ANCOVA:

$$Y = \beta_1 + \beta_2 G + \beta_3 X + \beta_4 XG + \epsilon$$

where XG is an interaction term between control and treatment.
Stated more simply:

$$\begin{cases} \mathbb{E}[Y|X, G = 1] = \beta_1 + \beta_2 + (\beta_3 + \beta_4)X & \text{for Group 1} \\ \mathbb{E}[Y|X, G = 0] = \beta_1 + \beta_3 X & \text{for Group 0} \end{cases}$$

In this setting:

  * The treatment effect is simply t the difference of regressions:

$$\Delta(X) = \beta_2 + \beta_4 X$$

  * Another potentially interesting quantity to consider: the abscissa of the point of intersection $X^0 = -\frac{\beta_2}{\beta_4}$ (This allows to compare the difference in groups at variables $X = 0$).
  * Inference for $\Delta(X)$: we can compute the OLS estimates for each of the coefficients in the afore-described problem. Pick a point (your favorite: $\bar{X}_G$ to get the average treatment effect ofa given group G for instance, or $C_a = -\frac{S_{24}}{S_{44}} = -\widehat{(\frac{\beta_2}{\beta_4})}$ (the point at the abscissa x=0) to get the ANCOVA treatment effect): estimate the ratio $\frac{D(X)}{S_{D(X)}}$ via t-distribution, with N-4 degrees of freedom/
  * R Non simultaneous $D(X) \pm \sqrt{F^{\alpha}_{1,N-4} S^2_{D(X)}}$
  * R Simultaneous (Working Hotelling bands) $D(X) \pm \sqrt{2F^{\alpha}_{2,N-4} {S^2_{D(X)}}^K}$

- **Pick-a-point Procedure:** the overall treatment effect, in which dependence on X of the treatment effect is ignored, can be estimated by evaluating the difference between the sample regression lines at a prespecified value of X. Bottom-line: pick a certain value for X and evaluate treatment effect for that X.

- This differs from the simultaneous coverage procedures, where we are interested in comparing the regression over whole ranges of X.

- **Study designs**:

  - Random assignment: randomized control trials (RCT). Here the Average treatment effect (ATE) can be obtained via a simple t-test: $\mu_0 - \mu_1$ or using ANCOVA. But we might also be interested in more subtle effects (Mediation :path analysis..., Moderation: Conditional Average Treatment Effect...)
  - Observational studies: haphazard assignments:
    * in haphazard assignments, we try to recreate random assignments. It doesnt usually work, because of selection bias: where you knowingly or unknowingly create unrepresentative samples.

* **Regression Adjustments for Quasi-Experiments**: suppose we have a mis-specified model:

$$Y = \beta_0 + \beta_1 G + U$$

where the group assignment G and U (effect) are not independent.

* First step: We try to adjust/pre-measure X:

$$\hat{\alpha} = \bar{Y}_1 - \bar{Y}_0 - \hat{\beta}(\bar{X}_1 - \bar{X}_0)$$

* Second step: we evaluate via a t-test of $\hat{\beta} = 0$.If it's not significantly different from 0, then no adjustment is required. Otherwise, will have to adjust

– RD systematic assignment

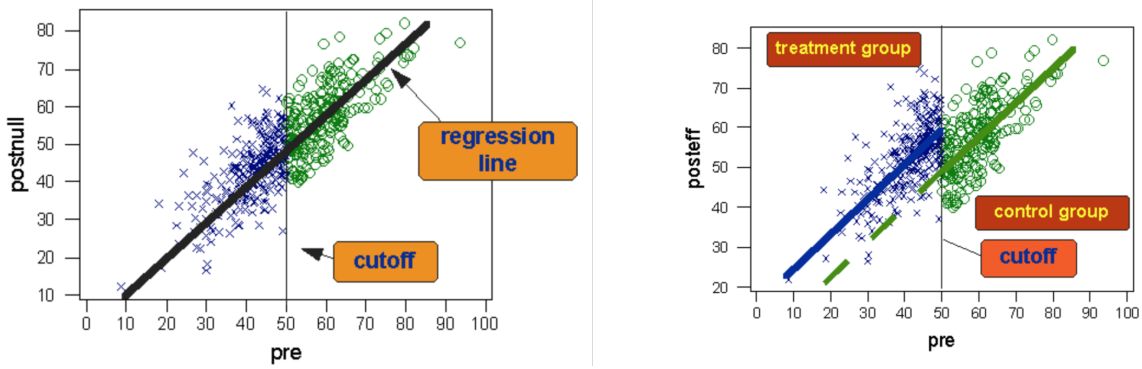
$$\hat{\beta} = 0$$

– **Regression discontinuity**(solves the ethical questions of "we don't want to deprive some people of the treatment if they need it). Sharp threshold to determine if a subject should be control or trial.
Helpful illustration from Wikipedia:
"The intuition behind the RDD is well illustrated using the evaluation of merit-based scholarships. The main problem with estimating the causal effect of such an intervention is the endogeneity of performance to the assignment of treatment (e.g. scholarship award): Since high-performing students are more likely to be awarded the merit scholarship and continue performing well at the same time, comparing the outcomes of awardees and non-recipients would lead to an upward bias of the estimates. Even if the scholarship did not improve grades at all, awardees would have performed better than non-recipients, simply because scholarships were given to students who were performing well ex ante.

Despite the absence of an experimental design, a RDD can exploit exogenous characteristics of the intervention to elicit causal effects. If all students above a given gradefor example 80%are given the scholarship, it is possible to elicit the local treatment effect by comparing students around the 80% cut-off: The intuition here is that a student scoring 79% is likely to be very similar to a student scoring 81%given the pre-defined threshold of 80%, however, one student will receive the scholarship while the other will not. Comparing the outcome of the awardee (treatment group) to the counterfactual outcome of the non-recipient (control group) will hence deliver the local treatment effect."[2]

* **benefits** In experimental or other quasi- experimental designs we either assume or try to provide evidence that the program and comparison groups are equivalent prior to the program so that post-program differences can be attributed to the manipulation. The RD design involves no such assumption. Instead, with RD designs we assume that in the absence of the program the pre-post relationship would be equivalent for the two groups. Thus, the strength of the RD design is dependent on two major factors. The first is

[2] `https://en.wikipedia.org/wiki/Regression_discontinuity_design`



(a) Pre-Post distribution with no treatment effect.

(b) Regression-Discontinuity Design with Ten-point Treatment Effect

Figure 4: Figure (b) is identical to Figure (a) except that all points to the left of the cutoff (i.e., the treatment group) have been raised by 10 points on the post-test. The dashed line in Figure (b) shows what we would expect the treated group's regression line to look like if the program had no effect (as was the case in Figure (a)).

the assumption that there is no spurious discontinuity in the pre-post relationship which happens to coincide with the cutoff point. The second factor concerns the degree to which we can know and correctly model the pre-post relationship and constitutes the major problem in the statistical analysis of the RD design which will be discussed below.

- **Regression Discontinuity analysis:**
  - 5 central assumptions:
    1. The Cutoff Criterion. The cutoff criterion must be followed without exception. When there is misassignment relative to the cutoff value (unless it is known to be random), a selection threat arises and estimates of the effect of the program are likely to be biased. Misassignment relative to the cutoff, often termed a "fuzzy" RD design, introduces analytic complexities that are outside the scope of this discussion.
    2. The Pre-Post Distribution. It is assumed that the pre-post distribution is describable as a polynomial function. If the true pre-post relationship is logarithmic, exponential or some other function, the model given below is misspecified and estimates of the effect of the program are likely to be biased. Of course, if the data can be transformed to create a polynomial distribution prior to analysis the model below may be appropriate although it is likely to be more problematic to interpret.
    3. Continuous Pretest Distribution. Both groups must come from a single continuous pretest distribution with the division between groups determined by the cutoff. In some cases one might be able to find intact groups (e.g., two groups of patients from two different geographic locations) which divide on some measure so as to imply some cutoff. Such naturally discontinuous groups

must be used with caution because of the greater likelihood that if they differed naturally at the cutoff prior to the program such a difference could reflect a selection bias which could introduce natural pre-post discontinuities at that point.

4. Program Implementation. It is assumed that the program is uniformly delivered to all recipients, that is, that they all receive the same dosage, length of stay, amount of training, or whatever.

- **Grand tour of the regression adjustements**: Suppose we have a model of the form:

$$Y = \alpha + \gamma G + U$$

where U and G are in fact not independent of each other. This is a little problematic, because it biases the t-test for the difference in means models. What we want to do is take out the effect of other covariates X in the estimation of the treatment effect, so we basically want to estimate:

$$\bar{Y} - \hat{\beta}\bar{X}$$

But what $\hat{\beta}$ should we plug in?

Here is a little grand tour of the several adjustments that people have made to try to correct this t-test:

  - Vanilla game: no adjustment, we assume $\hat{\beta} = 0$.
  - Estimate the effect of gain: $\hat{\beta} = 1$. This is typically for longitudinal studies in which the end point $\bar{Y}$ is not of interest in itself, but we rather want to assess the existence of a significant difference in the way that people have progressed from time 1 to time 2.
  - Use the adjusted regression coefficient $\hat{\beta} = \beta_{YG \cdot X}$ (standard ANCOVA).
  - Use an adjusted ANCOVA $\hat{\beta} = \frac{\beta_{YG \cdot X}}{\text{Reliability}(X)}$ (because people used to be convinced that ANCOVA under-adjusts. It has been proven not to be the case, it can under-adjust, over adjusts, etc.)
  - And a second type of adjusted ANCOVA coefficient: $\hat{\beta} = \frac{\beta_{YG \cdot X}}{r_{(X,Y)}}$
  - Belson-(Peters) estimator: estimate the slope from the control group $\hat{\beta} = \hat{\beta}_0$. While this method is pretty old, it has regained increased interest over the past few years in the field of prognostics.

6 Week 6. Instrumental variable methods, simultaneous equations, reciprocal effects

Main takeouts from the week: 3 main topics

- **Instrumental Variable methods**: the idea is that in many situations, it is hard to estimate the true relationship between a given output Y and an input X, because there might be some confounding factors that we are not putting in the regression. To alleviate this problem, IV methods have been developed: the idea is to replace X by Z where Z is correlated with Y and X but not with any of the residual noise, and to run the regression. These yield consistent, yet biased estimates of the effects. It also seems hard to find good instrumental variables.
- Random assignment as an Instrumental Variable (AIR paper): the idea here is that we can use random assignments (e.g. lottery draws) as instrumental variables: ex: for the Vietnam war, it's difficult to assess the impact of having served in the military on future outcomes, because a large fraction of the veterans volunteered (ie not random assignment). To assess this effect, we can use the lottery draw as IV.
- **"As-if-by-experiments"**: broken regression observational studies
- **Random Assignments in RCT** (randomized controlled trials). Ex: draft, lottery, encouragement designs. We can use G (the group assignment) as an instrumental variable
- **Simultaneity**: ex: supply and demand. (small bias correction). Reciprocal ideas (try to use longitudinal data, and entangle associations by looking at data over time).

In more details:

- **Definition: Instrumental Variable:** used to estimate causal relationships when controlled experiments are not feasible or when a treatment is not successfully delivered to every unit in a randomized experiment. IVs are used when an explanatory variable of interest is correlated with the error term, in which case ordinary least squares and ANOVA give biased results. Such correlation may occur 1) when changes in the dependent variable change the value of at least one of the covariates ("reverse" causation), 2) when there are omitted variables that affect both the dependent and independent variables, or 3) when the covariates are subject to non-random measurement error. Explanatory variables which suffer from one or more of these issues in the context of a regression are sometimes referred to as endogenous.
  IV can be used either:
  a. For controlling for measurement error
  b. in an omitted variable setting (fixing broken regression): $Y = \beta_0 + \beta_1 X + U$ such

that $Cov(X, U) \neq 0$

- More formallly:
  - Let Z be an exogenous variable ith no partial effect on Y.
  - X on Z: $X = \pi_0 + \pi_1 Z + V$
  - $Cov(Z, Y) = \beta_1 Cov(X, Z) + \underbrace{Cov(Z, U)}_{=0 \text{ by assumption}}$

$$\implies \beta_1 = \frac{Cov(Z, Y)}{Cov(Z, X)}$$

- Weak instrumental variables inflates variance . The MSE is always worse, even if the bias is 0
- $Y = \beta_0 + \beta_1 Ed + U$

$$Cov(X, Y) = \beta_1 Var(X) + Cov(CX, U)$$

- Fix broken regression: we have omitted variables: Dose-response model. $Cov(G, U) \neq 0$. D and G are in fact correlated with omitted variables in U. This means that OLS will fail for $Y = \beta_0 + \beta_1 X + U$ when $Cov(X, U) \neq 0$

$$Y = \beta_0 + \beta_1 D + U$$

$$Y = \beta_0 + \beta_1 G + U$$

- Instrumental variables are a good way to proceed. Let's pick Z such that $Cov(Z, U) = 0$ and $Cov(X, Z) \neq 0$ Omitted variables induce bias. People self-select in levels of education (rather than random assignment)/
- Randomized Control Trials IV:
  - Encouragement design: for estimating dose response, (with the dose either binary of measured) $G = 1, 0$ encourage or not. D: self-selected dose. Y: outcome. IV assume that the assignment has no effect on Y: The Wald estimator for the IV effect is:

$$\hat{\beta}_{IV} = \frac{S_{YG}}{S_{DG}} = \frac{\hat{Y}_1 - \hat{Y}_0}{\hat{D}_1 - \hat{D}_0}$$

  - Compliance Design: salvage broken protocols in RT
- Reciprocal Effects, Simultaneous Equations
  - **CLC : cross-lagged correlation**. We want to compare $r_{X_1 Y_2}$ to $r_{Y_1 X_2}$. Causal predominance to the larger. This seems to be particularly useful for determining causal relationship in longitudinal studies. In the following drawing, we have two "waves" of variables X and Y, observed sequentially at times and 2. Yet it was shown (Rogosa 1980) that this is not a useful procedure for the analysis of panel data.It is in fact difficult to make any causal claims or determine if the correlation that we observe is spurious or not.

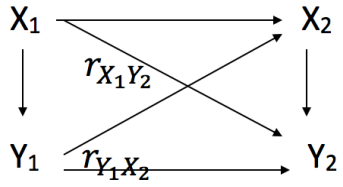


Figure 5: Reciprocal Effects

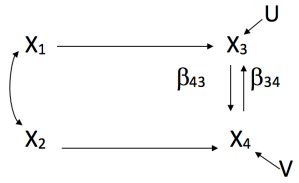


Figure 6: Reciprocal Effects

– Non-recursive (feedbacks). This screws up the OLS estimates, which are no longer consistent.

$$X_3 = \beta_{31}X_1 + \beta_{34}X_4 + U$$
$$X_4 = \beta_{42}X_2 + \beta_{43}X_3 + V$$

WLOG:

$$\mathbb{E}[X_j] = 0, \mathbb{E}[U] = \mathbb{E}[V]$$

OLS estimates:simultaneity bias:

$$\mathbb{E}[\hat{\beta}_{31}^{OLS}] = \beta_{31} - \frac{\sigma_{14}\sigma_{4u}}{\sigma_{11}\sigma_{44} - \sigma_{14}^2}$$

$$\mathbb{E}[\hat{\beta}_{34}^{OLS}] = \beta_{34} + \frac{\sigma_{11}\sigma_{4u}}{\sigma_{11}\sigma_{44} - \sigma_{14}^2}$$

$$\mathbb{E}[\hat{\beta}_{42}^{OLS}] = \beta_{42} - \frac{\sigma_{23}\sigma_{3v}}{\sigma_{22}\sigma_{33} - \sigma_{23}^2}$$

$$\mathbb{E}[\hat{\beta}_{43}^{OLS}] = \beta_{43} + \frac{\sigma_{22}\sigma_{3v}}{\sigma_{22}\sigma_{33} - \sigma_{23}^2}$$

We have to replace this with the IV estimates.

7 Week 7.Compliance and experimental protocols; intent to treat and compliance adjustments (under construction)

Main takeouts from the week:

- We have to take into account the fact that people tend to comply or not to an experimental setup, which can potentially bias our estimates of the treatment effect and hinder our analysis.
- We can define several types of populations of compliers/ non-compliers, as well as adjusted treatment effects. (Intent to treatm as-treated, oer-protocol, Compliance Average Effect (CACE))

Treatment	Control
(A) Compliers	(a) Potential compliers
(B) Non Compliers	(b) Potential Non Compliers

Table 1: Compliers: Notations

- Post-treatment variables measured after the treatment assignment and before the assessment of final outcomes of interest.These variables can potentially mediate the effect of treatment assignment on the outcome.
- Sometimes, we want to make inferences taking into account these intermittent post-treatment variables (vs. ITT analysis ignoring these variables).Pos0ttreatment variables are usually not randomized and already affected by treatment assignment
- There are several quantities that we might be interested in evaluating:
  - Intent-to Treat (ITT) analyses (A+B) vs (a +b): assumes that everyone complied and assesses the evaluated effect ("intent" to treat, not actual treatment, so non compliers are not a problem)
  - As-treated analysis (A) vs (B +a+b): want to just have the effect of the treatment on the compliers in the treatment.
  - Per protocol Analysis (A) vs (a+b)
  - CACE (Compliance Average Causal effect) (A) vs (a)
- Setting: randomized trials, where successful placebo control is unlikely, 2 conditions: intervention and control ($Z = 0$ and $Z = 1$). Two compliance types (c: receives treatment if assigned, and does not if not assigned. Compliance rate: π_c) and non compliers (does not receive treatment even if assigned to receive it: Non compliance rate: $1 - \pi_c$)

- Estimator of interest:

$$ITT = \mu_1 - \mu_0 = \pi_c(\mu_{c1} - \mu_{c0}) + (1 - \pi_c)(\mu_{n1} - \mu_{n0})$$

- Dean Eckles at MIT has nice compliers estimates with inverse propensity scoring.
- **Intent to Treat and Non-Compliance**. We have to say in whether people comply or not, we can just observe the outcome.
- Compliance: The IV-approach: some exogenous factor (independent of treatment/control assignment) induce some people not to comply to the treatment.
- Compliance Efron Feldman:
  - $Z(u)$: compliance of patient u.
  - $Y_0(u)$ response if Placebo
  - $Y_X(u)$ response if assigned treatment at dose X
  - $Y_X(u) = G_X + (1 + H_X)Y_0(u) + e_X(u)$

$$G_0 = H_0 = 0$$

$$\delta(X) = \mathbb{E}[Y_X(u) - Y_0(u)] = G_X + H_X \times \mathbb{E}[Y_0(u)]$$

  - CACE estimate: $\frac{\hat{\mu}_1 - \hat{\mu}_0}{\pi_C}$ Pb this depends on the proportion of compliers, and hence on the definition of the compliers. This is pretty problematic (eg: for drugs: what threshold makes a person who skips few pills a complier or a non-complier).

8 Week 8. Matching and propensity score methods

Main takeouts from the week:

- Suppose you have an outcome of interest Y, group assignments G (defined by haphazard, etc.)and additional covariates X. So far we have been trying to compute effects on Y while adjusting for the groups $Y \sim G$.
- Instead of adjusting $Y \sim G$, we adjust $G \sim X$, and we try to re-create the experiments that we would have liked to do $\implies$ This is called matching.
- Modern matching techniques
- Propensity scores: people with equal propensity: treatment is independent of all the covariates.

In more details:

- **Goal:** find subgroups (subclasses split on a given covariate or matched pairs) in which the treatment and control have balance– that is, the same distribution of observed covariates.
- **1: Most classical method: subclassification:** Instead of putting a linear model that accounts for all of the covariates, we bin subjects according to the value of a given covariate (ex: age). This allows comparisons that do not rely on any particular functional form (e.g, linearity) for the relationship between Y and the covariate within each treatment group. Note that this becomes tough when we have several covariates... In that case we might use propensity scoring (see 3.)
- **2: Matching:** can d one-to one matching, optimal matching, k to 0ne, etc. There are a number of R packages that can take care of this easily.
- **3: Propensity scoring:** The propensity score is the probability of treatment vs control as a function of observed covariates $e(X) = \mathbb{P}Z = 1|X]$. Models the reasons for treatment vs control at the level of decision makers. We can then use to subclassify or match on the propensity score as if it were the only one covariate.
  - **Theorem: balancing score(at large)**: Balancing score $b(X)$ is a quantity such that the conditional distribution X given $b(x)$ is the same between treated and control units: $X \perp Z|b(x)$. The coarsest balancing score is this propensity score.
  - **theorem** There is approximatively 90% reduction in bias for sublassifying at quintiles of population propensity score.

9 Week 9. Longitudinal (mainly time-1, time-2) data analysis for Experimental designs and Observational studies. (TO BE COMPLETED)

Main takeouts from the week:

- Cross-over designs: patients are subject to two treatments, A and B, order is assigned at random: $A \to B$ or $B \to A$.
- Comparing groups on time-1 time-2 : several options: repeated measure ANOVA or t-test.
- Diff in diffs with matching: technique to get the ATE under constant selection bias assumptions.
- Lord's paradox

In more details:

- **Cross-over designs:** :
  - Carryover is the persistence of a treatment effect applied in one period in a subsequent period of treatment.
  - With no carryover effects, the treatment effect is estimated by $(\bar{d}_1 + \bar{d}_2)/2$ which has variance σ_w^2/n, compared to (in the case of two independent groups, and 2n subjects in each group): a variance of $\sigma_b^2 + \sigma_w^2)/n$
  - Thus, the crossover design has the potential to substantially increase the precision of the estimate of the treatment effect.
- **Diffs in diffs**: Time 1 vs Time 2:

$$\text{Time 1:} \quad \bar{Y}_{1T} - \bar{Y}_{1C} : \text{ selection bias}$$

$$\text{Time 2:} \quad \bar{Y}_{2T} - \bar{Y}_{2C} : \text{ ATE effect + selection bias (assumed to remain identical)}$$

Hence the ATE effect is:

$$ATE = (\bar{Y}_{2T} - \bar{Y}_{2C}) - (\bar{Y}_{1T} - \bar{Y}_{1C})$$

It's also a good idea to use matching in the first step to reduce the selection bias.

10 Causal Inference-y jargon

- **SUTVA:** Single Unit Treatment Variable assignment. Main assumption in many causal inference problem, amounts to saying that all units are independent (i.e, for instance, impact of treatment on one unit does not impact the other units).
- **Confounders:** In statistics, a confounder (also confounding variable, confounding factor, or lurking variable) is a variable that influences both the dependent variable and independent variable causing a spurious association
- **IV: Instrumental Variables:** used to counter measurement errors or correlated assignments.
- **Endogenous variables:** Explanatory variable is correlated with the error term. The distinction between endogenous and exogenous variables originated in simultaneous equations models, where one separates variables whose values are determined by the model from variables which are predetermined; ignoring simultaneity in the estimation leads to biased estimates as it violates the exogeneity assumption of the GaussMarkov theorem.
  Examples:
  - Omitted variable $y = \alpha + \beta x + \gamma z + u$, but we omit $z \implies y = \alpha + \beta x + \epsilon$ with $\epsilon = \gamma z + u$
  - Measurement error: we have $y = \alpha + \beta x^* + \epsilon$ but we observe $x = x^* + v$. Since both x_i and u_i depend on ν_i, they are correlated, so the OLS estimation of β will be biased downwards.
  - Simultaneity:
    $$y_i = \beta_1 x_i + \gamma_1 z_i + u_i$$
    $$z_i = \beta_2 x_i + \gamma_2 y_i + v_i$$
- **Reliability:** The reliability coefficient $\rho_{xx'}$ provides an index of the relative influence of true and error scores on attained test scores. For a given variable X, the variance is the sum of the variance of true scores plus the variance of errors of measurement: $\sigma_X^2 = \sigma_T^2 + \sigma_E^2$ In its general form, the reliability coefficient is defined as the ratio of true score variance to the total variance of test scores:$\rho_{xx'} = \frac{\sigma_T^2}{\sigma_X^2} = 1 - \frac{\sigma_E^2}{\sigma_X^2}$