Covariates and Regression

1. Covariates: other variables that we can measure in addition the treatment variable and the outcome variable. Help check for problems in the experiment and to improve the precision of our estimate of the experimental treatment effect.
   1. Covariate Imbalance in an observational study; covariates can show us when we fail to have an apples-to-apples comparison.
      1. Example of a successful randomization check: all groups are identical when looking at demographics.
      2. Results from RAND Health Insurance experiments:
         1. There were many small treatment groups, so many treatments. There is not enough power. Can fix it in post, but try not to because it will spread observations too thin (pool treatments together).
         2. Increase coverage causes increased spending on health care. Not large improvement in physical health, but increase mental health and financial status. No evidence of increased preventative-care or decreased emergency department utilization.
         3. Noncompliance: not everyone in the treatment group actually received the treatment.
            1. The effects of the Oregon insurance are 4x larger for each person who actually received insurance, than the “intent to treat” estimates.
      3. Joint test for significance of covariates using an F-test. T-tests are often sufficient, but if covariates are correlated, t-tests could fail to reject when they should.
         1. Regress the binary treatment variable on the whole list of covariates. Interpret the regression F statistic to test the hypothesis that all covariate differences are jointly equal to 0.
2. History of Experiments and Randomization
   1. Milestones
      1. Religious prophet Daniel: proposed an experiment on a vegetarian diet
      2. James Lind (1742): experimented with citrus in the diet and demonstrated that it cured scurvy (didn’t randomize, but chose his 12 subjects with covariate balance in mind)
      3. Charles Peirce (1885): first recorded use of random assignment
      4. Sir Ronald Fisher (1935): detailed theory of randomized experiments
      5. Galileo (1500s): proposing the experimental method
   2. Use an F-test to ensure proper randomization.
   3. Use regression as an estimator. Pure matching requires a match from the untreated group for each member of the treated group. Regression summarizes this effort by specifying a single equation simultaneously containing both the treatment variable and the covariates we are accounting for.
      1. Regression with experimental data: Tennessee STAR Experiment
      2. Summary of using regression with experiments
         1. With randomized treatment assignment, we know that the treatment is uncorrelated with everything else: both observable covariates and unobservables we can’t measure.
            1. In an experiment, we don’t have to worry about omitted-variable bias, because we should get approximately the same answer no matter how many covariates we include.
         2. What including covariates does for us in an experiment is explain some of the residual variance in the regression, allowing us to shrink the standard error on the treatment effect.
      3. Differences in Differences redefines the outcome variable to be a before-vs.-after change. Use the change in Y. Similar to inserting the past (lagged) value of the outcome into the regression as a covariate. Controlling for their predisposition of their treatment.
      4. Omitted variable bias example: Tennessee STAR experiment and sending only one college application. Note the direction of bias. Omitted variable bias can be mitigated by controlling for college selectivity.
         1. Omitted variable bias theory: remember intuition about the direction of bias. There can be overestimation bias and underestimation bias.
            1. The direction and magnitude of the bias depend on how correlated is with . If omitted variable has no effect of its own, there’s no bias.
            2. The direction and magnitude of the bias depend on how correlated is with . If omitted variable is not correlated with the included variable, then there’s no bias.
      5. Intellectual history of regression: developed n the late 1800s by Galton and Yule. Started with Galton’s first linear model (1886) with “regression to the mean”.
         1. Conditional expectation function: models don’t have to be linear.
         2. Regression specification: how to choose the covariates and functional form of a regression.
            1. Saturated model: each value of the covariate gets a different dummy variable to represent it. With multiple covariates, we also have to include all possible interactions between these dummy variables.
            2. Why choose a model that is not fully saturated? Worry about too little data per cell to get identification and precision.

Linear in a continuous variable. Exclude interaction effects.

* + - * 1. Using lots of covariates increases statistical precision (standard errors decrease), but “fishing expeditions” happens often with observational data. Not much downside in an experiment, if we have made sure that X is independent of all covariates.