**Day 2 stuff (covariates).**

Male earnings grows by +2 pounds every year. Delta Y(0) = +2

Female earnings grows by +1 pounds every year. Delta Y(0) = +1

**Case 1. Equal male/female in treatment and control**

Assume that I have 50% of my sample in the treatment group is male and 50% of my control group is male.

Then calculate Delta E[Y(0)|D=1] and calculate Delta E[Y(0)|D=0]

Delta E[Y(0)|D=1] = 0.5 x 2 + 0.5x1 = 1+0.5 = 1.5

**Treatment group trend in Y(0) = 1.5**

Delta E[Y(0)|D=0] = 0.5 x 2 + 0.5 x 1 = 1.5

**Control group trend in Y(0) = 1.5**

Delta E[Y(0)|D=1] = Delta E[Y(0)|D=0] “parallel trends”

Case 1 is an example of “unconditional parallel trends” in this sense – the two groups, treatment and control, are evolving similarly on average on expected Y(0) trends and you do not need any covariates because the covariates have the same distribution in both groups.

**Case 2. Treatment is 75% male (25% female), but control is 25% male (75% female), but the trends we said (male +2, female + 1) is the same.**

Calculate Delta E[Y(0)|D=1] and calculate Delta E[Y(0)|D=0] in light of these *differing sex ratios in the treatment and control groups.*

Delta E[Y(0)|D=1] = 0.75 x 2 + 0.25 x 1 = 1.75

**Treatment group counterfactual trend in Y(0) = 1.75**. It’s higher than in the 50/50 case because it has more men.

Delta E[Y(0)|D=0] = 0.25 x 2 + 0.75 x 1 = 1.25

**Control group trend in Y(0) = 1.25, *which means unconditional parallel trends does not hold because of the “imbalance” in the sex ratio .***

**Case 3: male/female and young/old. Old men are seeing their earnings grow by +3, young men see their earnings grow by +1, old females are seeing their earnings grow by +2, and young females are seeing their earnings grow by +0.5**

**If it is the case:**

1. **Two observable populations have different trends in Delta E[Y(0)], *and***
2. **Treatment and control group are “imbalanced” on those very covariates describing those populations, *then***
3. **you will need to invoke a *different parallel trends assumption* called “conditional parallel trends”.**

**Conditional parallel trends means that the subpopulations (men +2, women +1) are on the same trend**

**To enter into the world of covariates with diff-in-diff where you invoke “conditional parallel trends” requires two tasks:**

1. **“What are the ordinary causes of your outcome trends?”** You need to figure out the Xs that you think cause Delta Y(0). But just because there are Xs that cause trends in Y(0) does not mean you need to include them in your model (unless you are focused on precision), *and*
2. **“Are those ordinary causes (observable covariates) differently distributed in the treatment group and the control group?”** You need a simple diagnostic to tell you whether or not the Xs are *on average* those Xs have different means in the treatment group and the control group at baseline.

This is me talking to you researcher to researcher, mortal human being to mortal flawed human being. You need to use your common sense, logic, deep institutional knowledge of the “left hand side variable”. And then once you have a candidate pool of covariates, you just need to figure out a way to reliably see if there is “problematic imbalance in those covariates between treatment and control”.

1. **Estimation**. You then need an econometric estimator that can accommodate those covariates into the estimate of the ATT *with minimal unnecessary assumptions.*  You goal is “what’s the least expensive method that I can use to get my target parmater while satisfying conditional parallel trends”. *Hint: it is not OLS.*

**Case 3: Two covariates male/female and young/old. Propensity scores *and* regression adjustment don’t just control for sex and age. They convert each of the covariates into a single dimension that looks like this:**

**Young men**

**Old men**

**Young females**

**Old females**

**So with two binary covariates, there are actualy *four dimensions*.**

**Three covariates: male/female, young/old, and black/white. How many dimensions are there now?**

**Young black male**

**Young black female**

**Young white male**

**Young white female**

**Old black male**

**Old black female**

**Old white male**

**Old white female**

**Adding in just one more dummy variable causes the dimensions to double again**

**10 covariates, all dummy variables. How many dimensions is that?**

**1,024 dimensions. For overlap, though, I would need at least two units per dimensions – one in the treatment, one in the control – just to satisfy common support.**

**Minimum dataset needed to satisfy common support is 1024 x 2 = 2048 units.**

**There’s this trade off when controlling for covariates to satisfy parallel trends:**

1. **YOU MUST SATISFY CONDITIONAL PARALLEL TRENDS, and that means to include it.**
2. **But the more covariates you throw into the model, the dimensions explode and that is called “curse of dimensionality”. What basically causes curse of dimensionality is that your dataset is simply too small**