MIT 14.32/14.320 **Spring 2023**
J. Angrist

Lecture Note 5

Intro to Multivariate Regression

1 Matchmaker, Matchmaker

We're often interested in the causal effect of one variable, X_{1i}, on an outcome variable, Y_i, in a scenario where the correlation between Y_i and X_{1i} at least partly reflects the fact that X_{1i} is correlated with another variable, X_{2i}, that also predicts Y_i.

- The strong correlation between health (Y_i) and health insurance (X_{1i}) in the NHIS might be explained by the higher schooling (X_{2i}) of the insured
- Econometricians use *multivariate regression* to control for such confounding factors
- Well-applied regression moves us along the path to *ceteris paribus* comparisons, mitigating and perhaps even eliminating selection bias.

Regression mitigates selection bias by conditioning on a set of possibly confounding variables so as to hold them constant. To keep things simple, suppose that X_{1i} is Bernoulli. "Holding things constant" in this case means replacing the unconditional comparison,

$$E[Y_i \mid X_{1i} = 1] - E[Y_i \mid X_{1i} = 0],$$

with conditional comparisons,

$$E[Y_i \mid X_{1i} = 1,\, X_{2i} = x] - E[Y_i \mid X_{1i} = 0,\, X_{2i} = x]. \tag{1}$$

In other words, we look at the CEF of Y given X_{1i}, *conditional on* $X_{2i} = x$. In Pset 1, for example, you're asked to compare the health of the insured and uninsured conditional on college graduation status.

- In terms of the potential outcomes notation introduced in LN4, we might be prepared to assume that X_{1i} is independent of potential $Y_i(0)$ conditional on X_{2i}:

  $$E[Y_i(0) \mid X_{1i} = 1,\, X_{2i} = x] = E[Y_i(0) \mid X_{1i} = 0,\, X_{2i} = x],$$

  where observed Y_i is determined by potentials according to $Y_i = (1 - X_{1i})Y_i(0) + X_{1i}Y_i(1)$. When this holds, the conditional comparison is causal

- Conditional comparisons of this sort are often said to be *effects* of X_{1i} computed while *matching* on values of X_{2i}.
  - Matching on X_{2i} ensures that all who contribute to the comparison of averages across values of X_{1i} have the same value of X_{2i} (at least)
  - Matching needn't yield 100% *ceteris paribus* comparisons to be useful and interesting
  - Even when the effect of X_{1i} is unlikely to be causal, as for the wage differences by ethnicity discussed below, the fact that conditioning on X_{2i} changes the size of the X_{1i} effect contributes to our understanding of it

- Define the conditional comparison:

$$\delta(X_{2i}) \equiv E[Y_i \mid X_{1i} = 1,\, X_{2i}] - E[Y_i \mid X_{1i} = 0,\, X_{2i}].$$

  Note that $\delta(X_{2i})$ is a function of X_{2i} and therefore has a distribution determined by the distribution of X_{2i}

1.1 Multivariate Regression Makes Me a Match

- Controls, X_{2i}, may take on many values (because there are many things to be controlled and/or because the individual controls take on many values, like SAT scores in the public-private college comparisons discussed below and in *MM* Chapter 2)
  - This threatens to overwhelm us with a multitude of conditional comparisons
- Regression solves this problem by fitting a linear model with a single conditional effect, while also generating the standard errors needed to do statistical inference for this effect

Regression is a many-splendored thing. We introduce it by assuming the CEF of Y_i given X_{1i} and X_{2i} is linear:

$$E[Y_i \mid X_{1i},\, X_{2i}] = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} \tag{2}$$

Defining $\varepsilon_i = Y_i - E[Y_i \mid X_{1i},\, X_{2i}]$, we have:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i,$$

where

$$E[\varepsilon_i \mid X_{1i,}\, X_{2i}] = 0.$$

By definition (as a CEF residual), ε_i is mean-zero and uncorrelated with regressors, X_{1i} and X_{2i}:

$$E[\varepsilon_i] = E[Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i}] = 0 \tag{3}$$
$$E[\varepsilon_i X_{1i}] = E[(Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i}) X_{1i}] = 0 \tag{4}$$
$$E[\varepsilon_i X_{2i}] = E[(Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i}) X_{2i}] = 0 \tag{5}$$

Distributing expectations and re-arranging terms gives 3 linear equations in 3 unknowns:

$$\beta_0 + \beta_1 E[X_{1i}] + \beta_2 E[X_{2i}] = E[Y_i] \tag{6}$$
$$\beta_0 E[X_{1i}] + \beta_1 E[X_{1i}^2] + \beta_2 E[X_{2i} X_{1i}] = E[Y_i X_{1i}] \tag{7}$$
$$\beta_0 E[X_{2i}] + \beta_1 E[X_{1i} X_{2i}] + \beta_2 E[X_{2i}^2] = E[Y_i X_{2i}] \tag{8}$$

The values of $\beta_0, \beta_1, \beta_2$ that solve this system define the *multivariate regression* of Y_i on X_{1i} and X_{2i}.

- Regression residuals are surely uncorrelated with the regressors that made them (why?)
- What if the CEF is nonlinear?
  - As explained in the appendix to MM Chpt 2 (and detailed in MHE Chpt 3), multivariate regression gives a best-in-class linear *approximation* to any CEF
  - An important consequence of this awesome approximation property is that regression is an *automatic matchmaker* (we "prove" this by computer below)

- With one independent variable, X_{1i}, we can write a linear CEF as $E[Y_i \mid X_{1i}] = \alpha + \beta X_{1i}$. In this case, the solution to equations (3)-(4) can be shown to be:

$$\beta = C(Y_i, X_{1i})/V(X_{1i})$$
$$\alpha = E[Y_i] - \beta E[X_{1i}]$$

  We use these *bivariate regression* formulas to understand *multivariate regression* (be sure you can derive this important special case).

2 Regression Talk

- Regression casts random variables in one of three roles
  - a variable to be explained, the *dependent variable*, denoted here by Y_i
    - In many of our examples, wages, grades, health, and test scores are the dependent variables
    - Dependent variables are sometimes called *outcome variables*, especially when the regressor of interest is a treatment dummy
  - *independent variables*, also known as *regressors,* of which, there are typically two types
    - *the regressor of interest*, like a treatment dummy in an experiment, a dummy for health insurance, years of education, or college characteristics
    - *control variables* that help us interpret the coefficient on the regressor of interest, making it more likely that this is a causal effect
    - sometimes regression talk references *covariates* - this can be a synonym for all regressors, but may also (depending on context) refer only to those included as *controls*
- Sometimes we lump all regressors together, labeling them with generic symbol X_i, which most often denotes a set or vector of regressors
- We often use notation that distinguishes the regressor of interest from control variables. We might, for example, use D_i for treatment status, with controls denoted by W_i. The vector X_i then contains both the focal regressor, D_i, and the controls, W_i. Formally, $X_i = [D_i \; W_i']$' (all vectors are column vectors).

3 Ordinary Least Squares

We *estimate* regression parameters with sample analogs. For example, bivariate regression estimators are given by:

$$\hat{\alpha} = \bar{Y} - \bar{X}_1 \hat{\beta}$$
$$\hat{\beta} = s_{X_1 Y}/s^2_{X_1}$$

These *Ordinary Least Squares* (OLS) estimators of α and β seem natural and have good statistical properties.

- Traditional 'metrics texts derive OLS as the solution to a sample least squares problem (hence the name). The traditional story in a nutshell:
  - Using observations on a pair of random variables: $\{(Y_i, X_{1i}); i = 1, \ldots, n\}$, you'd like to model Y_i as a linear function of X_{1i}

– How should you pick the slope and intercept? Minimizing the sample sum of squared errors,

$$SSE_{Y|X_1}(a, b) = \sum_i (Y_i - a - bX_{1i})^2,$$

generates $\hat{\alpha}$ and $\hat{\beta}$, above (be sure you can show this)

- OLS estimators, like sample means, are random and subject to sampling variance. This sampling variance is quantified by:
  1. standard errors, t-statistics, and confidence intervals
  2. F-statistics for joint tests
- Details TBD in LN7, but the ideas behind 1 are already familiar, so we'll start using these tools today.

4 Regression for Dummies

When the independent variable is a dummy, say D_i, then $E[Y_i \mid D_i]$ is linear:

$$E[Y_i \mid D_i] = \underbrace{E[Y_i \mid D_i = 0]}_{\alpha} + (\underbrace{E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0]}_{\beta})D_i$$

The bivariate regression slope and intercept must therefore be:

$$\begin{aligned}\alpha =& E[Y_i|D_i = 0] = E[Y_i] - E[D_i]\beta \\ \beta =& E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = C(D_i, Y_i)/V(D_i)\end{aligned}$$

(Show this in Pset 2)

- *Regression estimates differences in means.* No surprise: we've repeatedly seen regression used to estimate differences in means in comparisons of treatment and control groups using data from experiments like ALO (2009) and CGW (2017).
- When regressors are discrete and our regression model includes dummies for all of the possible values they might assume, the regression model is said to be *saturated.* The regression function and CEF coincide for saturated models (why?)

Matchmaker Reprise

- We introduced regression by assuming the CEF is linear, in which case the regression function is it. Otherwise, regression approximates a nonlinear CEF. That leaves an open ticket on my claim that regression is an automatic matchmaker, that is, a simple strategy to make *ceteris paribus* comparisons across values of D_i, while holding control variables, W_i, fixed.
- Here's the formal result behind this claim: Consider coefficient δ in the regression of Y_i on dummy D_i and a vector of saturated dummy controls, W_i, including constant:

$$Y_i = W_i'\gamma + \delta D_i + \varepsilon_i, \tag{9}$$

- Note that (9) has a single treatment effect. As noted above, however, the conditional difference in means with D_i switched on and off is a function of W_i. That is,

$$\delta(W_i) = E[Y_i|D_i = 1, W_i] - E[Y_i|D_i = 0, W_i]$$

- The regression coefficient δ in (9) can be shown to be a weighted average of $\delta(W_i)$. In particular:

$$\delta = \frac{E[\delta(W_i)\sigma_D^2(W_i)]}{E[\sigma_D^2(W_i)]}, \tag{10}$$

where $\sigma_D^2(W_i)$ is the conditional variance function for D_i given W_i. MHE Section 3.3 proves this. We show this below by computer.

5 Asians and Whites Under Control

- In a sample of prime age male high school grads in the 2016 American Community Survey, Asians (75% of whom are foreign-born) earn more than (non-Asian) Whites

```
------------------------------------------------------------------------------
   . summarize

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
        agep |     57,822    44.63239    2.835719         40         49
        wagp |     57,822    85133.18    88524.26          0     714000
        wkhp |     57,822    45.16629    10.01506          1         99
      racasn |     57,822    .0985092    .2980045          0          1
       racpi |     57,822    .0031476    .0560155          0          1
-------------+---------------------------------------------------------
      racwht |     57,822    .9063505     .291343          0          1
         uhe |     57,050    34.79542    29.36205          0   201.5789
      loguhe |     53,874    3.360682    .7238038  -6.437752   5.306181
       immig |     57,822    .1754868    .3803863          0          1
     yearsEd |     57,822    14.53499    2.428069         12         21
-------------+---------------------------------------------------------
      hsgrad |     57,822           1           0          1          1
     somecol |     57,822    .5345024    .4988125          0          1
     colgrad |     57,822     .442029    .4966323          0          1
    asianpac |     57,822    .1009304    .3012392          0          1
       white |     57,822    .8990696    .3012392          0          1

. bys asianpac: summarize loguhe yearsEd colgrad immig

------------------------------------------------------------------------------
-> asianpac = 0

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
      loguhe |     48,397    3.345498    .7155838  -6.437752   5.306181
     yearsEd |     51,986    14.40621    2.377636         12         21
     colgrad |     51,986    .4188435    .4933744          0          1
       immig |     51,986    .1102605    .3132171          0          1

------------------------------------------------------------------------------
-> asianpac = 1

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
      loguhe |      5,477    3.494851    .7800729  -3.912023    5.30231
     yearsEd |      5,836    15.68215    2.567458         12         21
     colgrad |      5,836    .6485607    .4774608          0          1
       immig |      5,836    .7565113    .4292243          0          1
```

- Is the Asian wage effect causal? (Ponder potential outcomes)
- Either way, an ethnicity gap in college graduation rates might explain it

```
. reg loguhe asianpac

      Source |       SS           df       MS      Number of obs   =    53,874
-------------+----------------------------------   F(1, 53872)     =    210.31
       Model |  109.751403         1  109.751403   Prob > F        =    0.0000
    Residual |  28113.8826    53,872  .521864468   R-squared       =    0.0039
-------------+----------------------------------   Adj R-squared   =    0.0039
       Total |   28223.634    53,873  .523892005   Root MSE        =     .7224

------------------------------------------------------------------------------
      loguhe | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
    asianpac |    .149353   .0102988    14.50   0.000     .1291672    .1695388
       _cons |   3.345498   .0032837  1018.81   0.000     3.339062    3.351934
------------------------------------------------------------------------------

. reg loguhe asianpac colgrad

      Source |       SS           df       MS      Number of obs   =    53,874
-------------+----------------------------------   F(2, 53871)     =   5612.77
       Model |  4867.00805         2  2433.50402   Prob > F        =    0.0000
    Residual |   23356.626    53,871  .433565851   R-squared       =    0.1724
-------------+----------------------------------   Adj R-squared   =    0.1724
       Total |   28223.634    53,873  .523892005   Root MSE        =    .65846

------------------------------------------------------------------------------
      loguhe | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
    asianpac |   .0094287   .0094818     0.99   0.320    -.0091556    .0280131
     colgrad |   .6041877   .0057679   104.75   0.000     .5928825     .615493
       _cons |   3.091823   .0038501   803.05   0.000     3.084277     3.09937
------------------------------------------------------------------------------

. reg loguhe asianpac yearsEd

      Source |       SS           df       MS      Number of obs   =    53,874
-------------+----------------------------------   F(2, 53871)     =   6267.00
       Model |   5327.2321         2  2663.61605   Prob > F        =    0.0000
    Residual |  22896.4019    53,871  .425022775   R-squared       =    0.1888
-------------+----------------------------------   Adj R-squared   =    0.1887
       Total |   28223.634    53,873  .523892005   Root MSE        =    .65194

------------------------------------------------------------------------------
      loguhe | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
    asianpac |  -.0170668   .0094149    -1.81   0.070      -.03552    .0013864
     yearsEd |   .1300443   .0011737   110.80   0.000     .1277437    .1323448
       _cons |   1.471337    .017173    85.68   0.000     1.437678    1.504996
------------------------------------------------------------------------------

.
. bys colgrad: reg loguhe asianpac

------------------------------------------------------------------------------
-> colgrad = 0

      Source |       SS           df       MS      Number of obs   =    29,986
-------------+----------------------------------   F(1, 29984)     =     28.38
       Model |  11.0183369         1  11.0183369   Prob > F        =    0.0000
    Residual |  11639.6575    29,984  .388195619   R-squared       =    0.0009
-------------+----------------------------------   Adj R-squared   =    0.0009
       Total |  11650.6758    29,985  .388550135   Root MSE        =    .62305

------------------------------------------------------------------------------
      loguhe | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
    asianpac |  -.0785125   .0147369    -5.33   0.000    -.1073975   -.0496276
       _cons |   3.097422   .0037183   833.01   0.000     3.090134     3.10471
------------------------------------------------------------------------------

------------------------------------------------------------------------------
-> colgrad = 1

      Source |       SS           df       MS      Number of obs   =    23,888
-------------+----------------------------------   F(1, 23886)     =     23.23
       Model |  11.3753421         1  11.3753421   Prob > F        =    0.0000
    Residual |  11695.0036    23,886  .489617498   R-squared       =    0.0010
-------------+----------------------------------   Adj R-squared   =    0.0009
       Total |  11706.3789    23,887  .490073216   Root MSE        =    .69973

------------------------------------------------------------------------------
      loguhe | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
    asianpac |   .0612206   .0127012     4.82   0.000     .0363255    .0861158
       _cons |   3.688275   .0049087   751.37   0.000     3.678654    3.697897
------------------------------------------------------------------------------
```

- Compare the college-controlled regression estimate of 0.009 (above) to the average conditional-on-college Asian effect:
$$-.0785(.558) + .0612(.442) \simeq -.017$$
Both are close to to zero
- Regression makes me a match! (on college graduation status)

6 Regression Fission

Every regression splits the dependent variable into two parts: residuals and fitted values. Consider a regression equation with two regressors:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i. \tag{11}$$

This can be written:

$$Y_i = \hat{Y}_i^* + \varepsilon_i,$$

where:

$$\hat{Y}_i^* \equiv \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$$

are the regression *fitted values* and the residual, ε_i, is the part left over. Note that fitted values are a linear combination of regressors.

- We'll have more to say about residuals and fitted values in LN8. For now, it's enough to note that residuals are mean-zero and uncorrelated with fitted values. That is,
$$E[\varepsilon_i] = 0$$
$$E[\hat{Y}_i^* \varepsilon_i] = 0$$
Look back at the equations that define regression parameters (3-5) to see why this must be so.
- Stata's `predict` command computes residuals and fitted values *in your data*. Sample fitted values and residuals are defined as:
$$Y_i = \hat{Y}_i + e_i,$$
where:
$$\hat{Y}_i \equiv \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i},$$
and hats on the right-hand side denote OLS estimates. Sample resids and fits satisfy:
$$\sum_{i=1}^{n} e_i = 0$$
$$\sum_{i=1}^{n} \hat{Y}_i e_i = 0$$
in the sample that made them.

7 Regression Anatomy

System (3) doesn't immediately reveal how multivariate regression works its matching magic. Let's look inside the system of equations that defines regression.

- Start with a regression equation with two regressors:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \tag{12}$$

- Add the following two *auxiliary regressions*:

$$X_{1i} = \delta_{10} + \delta_{12} X_{2i} + \tilde{x}_{1i}$$
$$X_{2i} = \delta_{20} + \delta_{21} X_{1i} + \tilde{x}_{2i}$$

  where the δ's are bivariate regression coefficients [e.g., $\delta_{12} = COV(X_{1i}, X_{2i})/V(X_{2i})$], while $\tilde{x}_{1i}$ is the residual from a regression of X_{1i} on X_{2i} and $\tilde{x}_{2i}$ is the residual from a regression of X_{2i} on X_{1i}

The following theoretical result is key to our understanding of multivariate regression models:

The Regression-Anatomy Theorem.

$$\beta_1 = COV(Y_i, \tilde{x}_{1i})/V(\tilde{x}_{1i})$$
$$\beta_2 = COV(Y_i, \tilde{x}_{2i})/V(\tilde{x}_{2i})$$

Proof. Substitute for Y_i using $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$, and use the properties of regression residuals and fitted values highlighted above.

- Multivariate β_1 captures the effect of $\tilde{x}_{1i}$, that is, the part of X_{1i} that is not explained (in a regression sense) by X_{2i}
- Multivariate β_2 captures the effect of $\tilde{x}_{2i}$, that is, the part of X_{2i} that is not explained (in a regression sense) by X_{1i}

```
. ***regression anatomy***
.
. reg loguhe asianpac yearsEd age

      Source |       SS           df       MS      Number of obs   =    53,874
-------------+----------------------------------   F(3, 53870)     =   4210.94
       Model |  5361.33436         3  1787.11145   Prob > F        =    0.0000
    Residual |  22862.2996    53,870  .424397617   R-squared       =    0.1900
-------------+----------------------------------   Adj R-squared   =    0.1899
       Total |   28223.634    53,873  .523892005   Root MSE        =    .65146

------------------------------------------------------------------------------
      loguhe | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
    asianpac |  -.0142974    .009413    -1.52   0.129     -.032747    .0041521
     yearsEd |   .1301485   .0011729   110.96   0.000     .1278496    .1324475
        agep |   .0088739   .0009899     8.96   0.000     .0069336    .0108142
       _cons |   1.073621   .0475708    22.57   0.000     .9803819     1.16686
------------------------------------------------------------------------------

.
. **step 1
.
. reg asianpac age yearsEd if e(sample)==1

      Source |       SS           df       MS      Number of obs   =    53,874
-------------+----------------------------------   F(2, 53871)     =   733.29
       Model |  130.397429         2  65.1987147   Prob > F        =    0.0000
    Residual |  4789.79355    53,871  .088912282   R-squared       =    0.0265
-------------+----------------------------------   Adj R-squared   =    0.0265
       Total |  4920.19098    53,873  .091329441   Root MSE        =    .29818

------------------------------------------------------------------------------
    asianpac | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        agep |  -.0034517   .0004529    -7.62   0.000    -.0043393    -.002564
     yearsEd |   .0198273     .00053    37.41   0.000     .0187885    .0208662
       _cons |  -.0326626   .0217734    -1.50   0.134    -.0753386    .0100134
------------------------------------------------------------------------------

. predict ap_resid, residuals

.
. **step 2
.
. reg loguhe ap_resid

      Source |       SS           df       MS      Number of obs   =    53,874
-------------+----------------------------------   F(1, 53872)     =      1.87
       Model |   .97911525         1   .97911525   Prob > F        =    0.1716
    Residual |  28222.6549    53,872  .523883555   R-squared       =    0.0000
-------------+----------------------------------   Adj R-squared   =    0.0000
       Total |   28223.634    53,873  .523892005   Root MSE        =     .7238

------------------------------------------------------------------------------
      loguhe | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
    ap_resid |  -.0142974   .0104582    -1.37   0.172    -.0347957    .0062008
       _cons |   3.360682   .0031184  1077.70   0.000      3.35457    3.366794
------------------------------------------------------------------------------
```

- It works! Phew!

8 Regression Checks for Balance

RCT research studies typically begin with evidence of *covariate balance* - the covariates here are random variables that describe the sample. We first encountered covariate balance in the intro slides, showing that treatment and control groups in the CGW2017 laptop experiment look similar. We also checked for balance in LN2, in the context of the pay-for-grades RCT discussed in ALO2009.

- An easy way to check for balance is to regress covariates on a treatment dummy, including any needed *stratification controls*
- Regression on dummy variables estimates differences in means
- The Angrist, Oreopoulos, and Williams (2014; AOW2014) RCT is stratified and therefore reports regression estimates of treatment-control differences *controlling for strata.*
  - An RCT is stratified when treatment is randomly assigned within covariate-defined subgroups called strata (a simple RCT randomizes unconditionally).
  - The AOW2014 section describing experimental design notes: *Treatment assignment was stratified by year (first and second) and sex, with 100 in each group. Within sex-year cells, assignment was stratified by high school GPA quartile, with 25 in each group. (The analysis below controls for strata.)*
- Pset 4 asks you to replicate the AOW2014 balance estimates; balance looks good within strata, as we'd expect given the RCT research design
- Pset 4 also asks you to check for balance without strata controls. This needn't work out well (though it might).