Advanced Quantitative Techniques (Class 9)

Gregory M. Eirich
QMSS

Agenda

1. Propensity score matching
2. A first differences example: financial satisfaction
3. From our first differences example to our fixed effects model
4. Some thoughts on these models
5. Random effects example

1. Propensity score matching

Propensity Score Matching

- Want to observe the effect of the treatment on the treated, average treatment effect on the treated (ATT), counterfactual causal inference

- Want to control for "selection bias" – the fact that some individuals are more likely to be chosen to treatment than others (usually based on choice and resources)

Descriptively, we know that married people are happier than non-married people.

But is that causal?

For males…

```
> lm(Formula, data = sub, sex == 1) # for men

Call:
lm(formula = Formula, data = sub, subset = sex == 1)

Coefficients:
(Intercept)      dm1TRUE      wordsum      db1TRUE      df1TRUE
  1.7300423    0.2639883    0.0003186    0.0475963   -0.0077360    0.01
  f.region4    f.region5    f.region6    f.region7    f.region8    f.re
 -0.0269087    0.0040748    0.0122763    0.0488256    0.0491491   -0.01
```

For females…

```
> lm(Formula, data = sub, sex == 2) # for women

Call:
lm(formula = Formula, data = sub, subset = sex == 2)

Coefficients:
(Intercept)      dm1TRUE      wordsum      db1TRUE      df1TRUE         educ       paeduc
   1.600822     0.282436     0.001921     0.143887    -0.031071     0.014596     0.00463
  f.region4    f.region5    f.region6    f.region7    f.region8    f.region9
   0.023217     0.009337     0.071182    -0.004279     0.009210    -0.004647
```

OVerall ...

```
> lm(Formula, data = sub)          # overall

Call:
lm(formula = Formula, data = sub)

Coefficients:
(Intercept)      dm1TRUE      wordsum      db1TRUE      df1TRUE         educ       paeduc
   1.668949     0.271979     0.002351     0.098360    -0.020594     0.012125     0.001927     0.0
  f.region4    f.region5    f.region6    f.region7    f.region8    f.region9
  -0.001253     0.007039     0.048478     0.018964     0.025735    -0.010741
```

We know that married people are happier than non-married people. Is this relationship causal?
Why? Why not?

Propensity Score Matching (Old School Way)

1. Estimate a selection equation that predicts likelihood of receiving a treatment (using logit)

N.B., We predict treatment on *observable* characteristics, even if we suspect there are unobservable ones that drive likelihood of treatment as well

1. Estimate a selection equation that predicts likelihood of receiving a treatment

```
> # Estimate the propensity model
> xvars <- xvars[-1]
> Formula <- as.formula(paste("dm1 ~ ", paste(xvars, collapse = " + ")))
> propensity_model <- glm(Formula, data = sub, family = binomial)

  # Matching & ATT estimate
  # outcome
> Y <- sub$n.happy
  # treatment
> Tr <- sub$dm1
   # propensity scores
> pscore <- propensity_model$fitted
  # one-to-one matching
> matching  <- Match(Y = Y, Tr = Tr, X = pscore)
> summary(matching) # "Estimate" is the estimated ATT

Estimate...  0.27158
AI SE......  0.012508
T-stat.....  21.713
p.val......  < 2.22e-16

Original number of observations..............  11306
Original number of treated obs...............  6323
Matched number of observations...............  6323
Matched number of observations  (unweighted).  61140
```

Propensity Score Matching

2. Ensure balance between the treated and untreated "strata" or blocks on all covariates that predict treatment

- N.B., Identification of the optimal number of blocks. This number of blocks ensures that the mean propensity score is not different for treated and controls in each block
- Common support: need to have both treated and untreated with lots of values of X in common; if not, then treated individuals with highest probability of treatment are not matched with untreated individuals

2. Ensure balance between the treated and untreated “strata” or blocks on all covariates that predict treatment

2. Ensure balance between the treated and untreated “strata” or blocks on all covariates that predict treatment

```
> # Check/test for balance
> mb <- MatchBalance(Formula, data = sub, match.out = matching, nboots = 500)

***** (V1) wordsum *****
                       Before Matching          After Matching
mean treatment........     6.3135                  6.3135
mean control..........     6.1092                  6.3278
std mean diff.........     10.012                -0.70237

mean raw eQQ diff.....    0.20369                0.072015
med  raw eQQ diff.....          0                       0
max  raw eQQ diff.....          1                       1

mean eCDF diff........   0.018572               0.0065468
med  eCDF diff........   0.015394               0.0058227
max  eCDF diff........   0.041741                0.014181

var ratio (Tr/Co).....    0.91347                 0.95403
T-test p-value........ 2.6428e-07                 0.66742
KS Bootstrap p-value.. < 2.22e-16              < 2.22e-16
KS Naive p-value...... 0.00012124              9.1532e-06
KS Statistic..........   0.041741                0.014181
```

2. Ensure balance between the treated and untreated “strata” or blocks on all covariates that predict treatment

```
***** (V2) db1TRUE *****
                       Before Matching          After Matching
mean treatment........    0.91934                 0.91934
mean control..........    0.93498                  0.9233
std mean diff.........    -5.7419                 -1.4525

mean raw eQQ diff.....   0.015653               0.0040072
med  raw eQQ diff.....          0                       0
max  raw eQQ diff.....          1                       1

mean eCDF diff........  0.0078184               0.0020036
med  eCDF diff........  0.0078184               0.0020036
max  eCDF diff........   0.015637               0.0040072

var ratio (Tr/Co).....     1.2197                  1.0471
T-test p-value........  0.0013954                 0.37885
```

2. Ensure balance between the treated and untreated “strata” or blocks on all covariates that predict treatment

```
***** (V7) incom16 *****
                       Before Matching          After Matching
mean treatment........     2.8569                 2.8569
mean control..........     2.9466                 2.8614
std mean diff.........    -10.863               -0.54603

mean raw eQQ diff.....   0.089906                0.02545
med  raw eQQ diff.....          0                      0
max  raw eQQ diff.....          1                      1

mean eCDF diff........   0.017949                0.00509
med  eCDF diff........  0.0057548              0.0027969
max  eCDF diff........   0.044481               0.014835

var ratio (Tr/Co).....    0.94681                0.94993
T-test p-value........ 1.6812e-08                0.73641
KS Bootstrap p-value.. < 2.22e-16             < 2.22e-16
KS Naive p-value...... 3.2496e-05             2.8676e-06
KS Statistic..........   0.044481               0.014835
```

Caution!

- Cases may not balance, in which case you need to alter your selection model (e.g., I originally included CHILDS but it made everything unbalanced, so I removed it and got balance)

Propensity Score Matching

3. Estimate the size of the treatment on the treated.

3. Estimate the size of the treatment on the treated

```
> summary(matching) # "Estimate" is the estimated ATT

Estimate...  0.27158
AI SE......  0.012508
T-stat.....  21.713
p.val......  < 2.22e-16
```

Nearest Neighbor Matching

- A treated case is matched to an untreated case that has the closest probability (1 to 1 matching)

Other matching algorithms are possible

- attr = Caliper/radius matching
- atts = stratification/interval matching
- attk = Kernel matching

OLS vs. ATT

OLS = 0.2719

ATT = 0.2716

- Conclusion: It looks like even after we control for the fact that selection into marriage is not random (using a few predictors), the average treatment effect of marriage on those who are married (compared to those with an equal probability of being married – the ATT) is very similar to the OLS estimate

Heterogeneous treatment effects

Brand, Jennie E., and Yu Xie. "Who benefits most from college? Evidence for negative selection in heterogeneous economic returns to higher education." *American sociological review* 75.2 (2010): 273-302.

Negative selection into college-

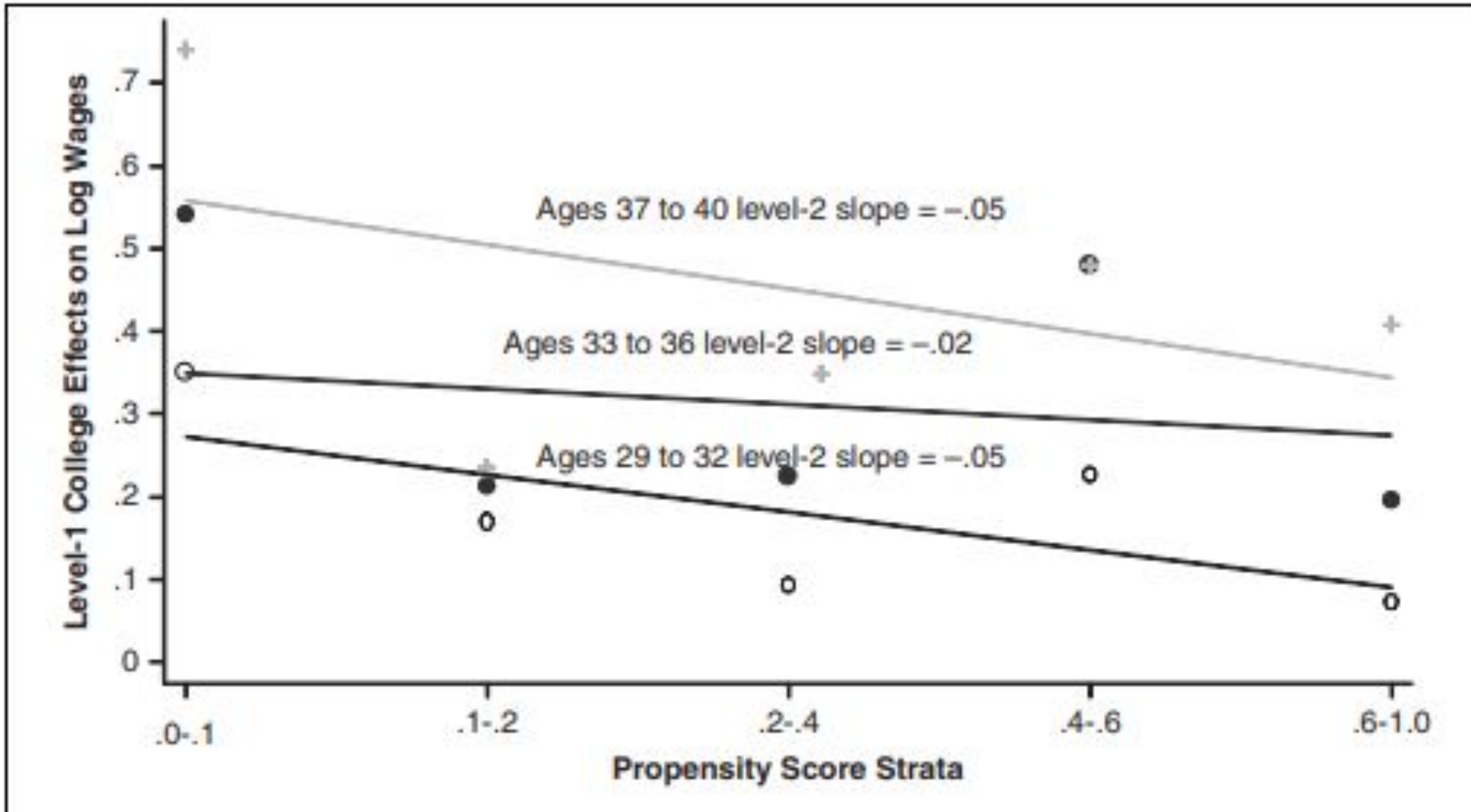


Figure 2. HLM of Economic Returns to College; NLSY Men

Do they know something we don't know?

Table 6. Proportion of College Majors for College-Educated Men by Propensity Score Strata: WLS Men

	Propensity Score Strata								
College Major	[.0–.05)	[.05–.1)	[.1–.15)	[.15–.2)	[.2–.4)	[.4–.6)	[.6–.7)	[.7–.8)	[.8–1.0)
Physical science	.00	.06	.04	.02	.03	.05	.05	.04	.05
Math	.00	.06	.04	.02	.06	.09	.08	.04	.05
Biological science	.11	.03	.04	.02	.09	.09	.11	.07	.12
Engineering	.04	.06	.13	.12	.06	.14	.13	.23	.22
Pre-professional	.00	.00	.00	.00	.00	.01	.01	.01	.02
Computer science	.04	.00	.04	.00	.01	.02	.01	.01	.01
Business	.19	.27	.17	.19	.16	.15	.10	.11	.10
Social science	.15	.15	.25	.17	.18	.19	.10	.22	.21
Humanities	.04	.03	.00	.10	.13	.08	.13	.11	.10
Art and music	.11	.09	.04	.07	.04	.05	.05	.01	.05
Education	.22	.18	.21	.14	.15	.08	.07	.06	.05
Communications	.04	.03	.00	.02	.06	.01	.01	.04	.01
Agriculture	.04	.00	.00	.02	.01	.01	.02	.04	.01
Other	.04	.03	.04	.10	.02	.03	.03	.04	.02
Number	27	33	24	42	145	196	120	171	375

Propensity score matching
(The modern way with MatchIt)

We know that married people are happier than non-married people.

OVerall …

```
> summary(lm(happy~ married + educ + age + childs + maeduc + attend, d2))

Call:
lm(formula = happy ~ married + educ + age + childs + maeduc +
    attend, data = d2)

Residuals:
    Min      1Q  Median      3Q     Max
-1.2727 -0.5612  0.0362  0.3665  1.5679

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.4757012  0.0684350  36.176  < 2e-16 ***
married     -0.3226741  0.0247600 -13.032  < 2e-16 ***
educ        -0.0229993  0.0044358  -5.185 2.33e-07 ***
age         -0.0009112  0.0008034  -1.134   0.2568
childs      -0.0045088  0.0085923  -0.525   0.5998
maeduc      -0.0062383  0.0035124  -1.776   0.0758 .
attend      -0.0200985  0.0043685  -4.601 4.41e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6066 on 2597 degrees of freedom
Multiple R-squared:  0.1012,  Adjusted R-squared:  0.09914
F-statistic: 48.74 on 6 and 2597 DF,  p-value: < 2.2e-16
```

We know that married people are happier than non-married people. Is this relationship causal?

Why? Why not?

1. Estimate a selection equation that predicts likelihood of receiving a treatment

```
install.packages("MatchIt")

d = read.csv(file.choose())


d$married = ifelse(d$marital==1, 1,0)

d2 = d %>% select("married","educ", "age", "childs", "maeduc", "attend",
"happy")

d2 = na.omit(d2)

m.out = matchit(married ~ educ + age + childs + maeduc + attend,
                data = d2, method = "nearest",
                ratio = 1)
```

2. Ensure balance between the treated and untreated on all covariates that predict treatment

```
> summary(m.out)

Call:
matchit(formula = married ~ educ + age + childs + maeduc + attend,
    data = d2, method = "nearest", ratio = 1)

Summary of balance for all data:
         Means Treated Means Control SD Control Mean Diff eQQ Med eQQ Mean eQQ Max
distance        0.5331        0.4535     0.1365    0.0796  0.0835   0.0799  0.1068
educ           13.7334       13.4557     3.1088    0.2777  0.0000   0.2938  2.0000
age            47.8730       44.9591    17.9141    2.9138  3.0000   3.7482  8.0000
childs          2.1902        1.4565     1.6628    0.7337  1.0000   0.7467  2.0000
maeduc         11.0070       11.5140     3.9082   -0.5070  0.0000   0.4996  2.0000
attend          4.0468        3.1605     2.7609    0.8863  1.0000   0.8885  2.0000


Summary of balance for matched data:
         Means Treated Means Control SD Control Mean Diff eQQ Med eQQ Mean eQQ Max
distance        0.5331        0.4598     0.1333    0.0733  0.0786   0.0733  0.1026
educ           13.7334       13.5705     3.0210    0.1629  0.0000   0.2330  2.0000
age            47.8730       45.0320    17.8423    2.8410  3.0000   3.6687  8.0000
childs          2.1902        1.4965     1.6689    0.6937  1.0000   0.7077  2.0000
maeduc         11.0070       11.5425     3.8877   -0.5355  0.0000   0.5355  2.0000
attend          4.0468        3.2416     2.7565    0.8051  1.0000   0.8051  2.0000
```

2. Ensure balance between the treated and untreated on all covariates that predict treatment

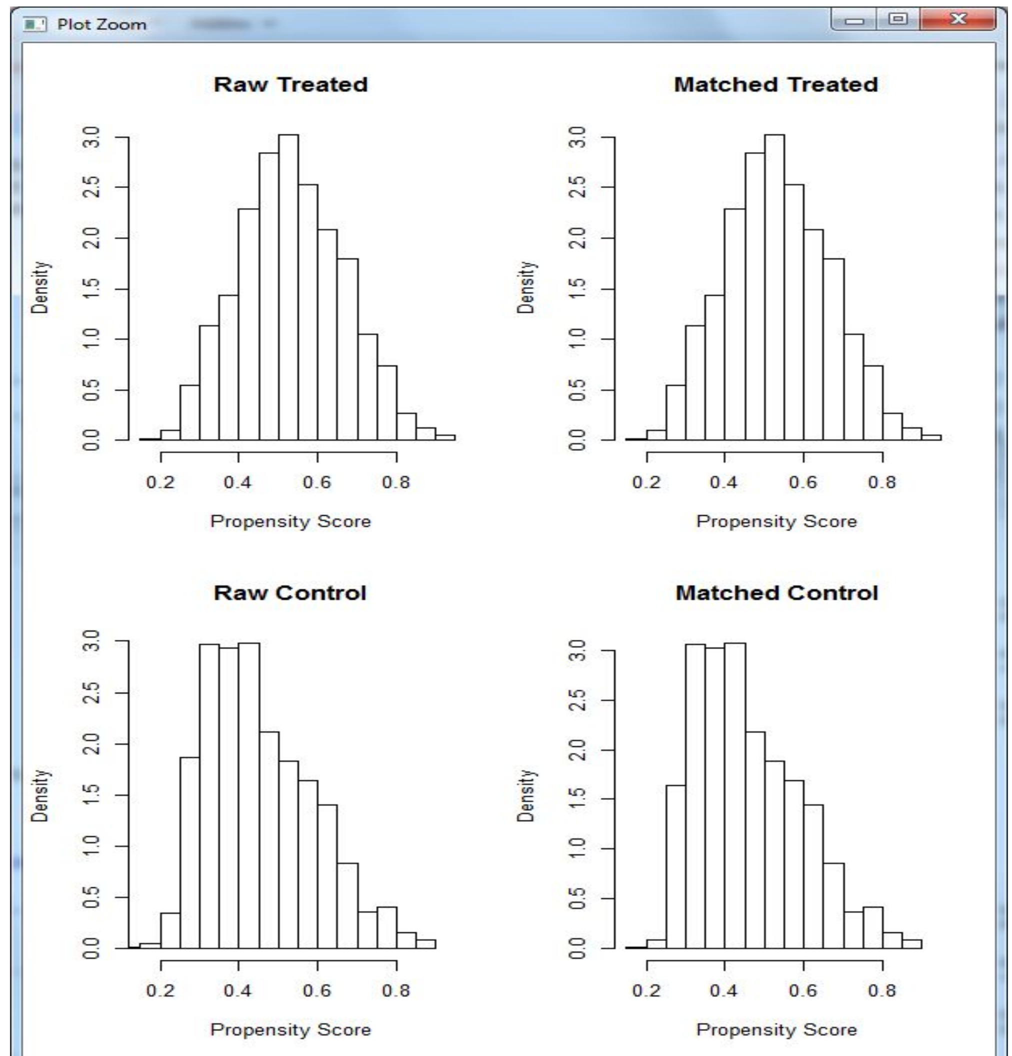
```
Summary of balance for matched data:
         Means Treated Means Control SD Control Mean Diff eQQ Med eQQ Mean eQQ Max
distance        0.5331        0.4598     0.1333    0.0733  0.0786   0.0733  0.1026
educ           13.7334       13.5705     3.0210    0.1629  0.0000   0.2330  2.0000
age            47.8730       45.0320    17.8423    2.8410  3.0000   3.6687  8.0000
childs          2.1902        1.4965     1.6689    0.6937  1.0000   0.7077  2.0000
maeduc         11.0070       11.5425     3.8877   -0.5355  0.0000   0.5355  2.0000
attend          4.0468        3.2416     2.7565    0.8051  1.0000   0.8051  2.0000

Percent Balance Improvement:
         Mean Diff. eQQ Med eQQ Mean eQQ Max
distance     7.9605  5.8492   8.2624  3.8428
educ        41.3444  0.0000  20.6897  0.0000
age          2.4996  0.0000   2.1210  0.0000
childs       5.4545  0.0000   5.2192  0.0000
maeduc      -5.6163  0.0000  -7.1763  0.0000
attend       9.1547  0.0000   9.3860  0.0000

Sample sizes:
          Control Treated
All          1321    1283
Matched      1283    1283
Unmatched      38       0
Discarded       0       0
```

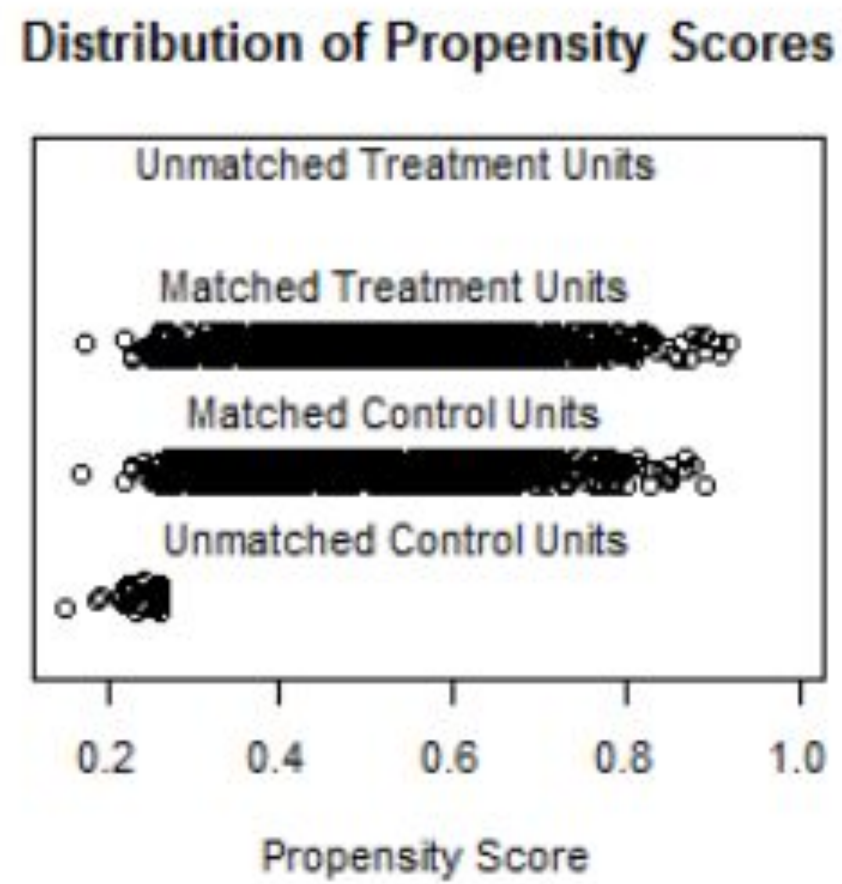
2. Does propensity to marry look the same for both groups?

```
plot(m.out, type = "hist")
```



2. Does propensity to married look the same for both?

```
plot(m.out, type = "jitter")
```



Propensity Score Matching

3. Estimate the size of average treatment

3. Estimate the size of the average treatment effect

```
> summary(lm(happy~ married, m.data1))

Call:
lm(formula = happy ~ married, data = m.data1)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.97973 -0.63367  0.02027  0.36633  1.36633 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.97973    0.01713  115.57   <2e-16 ***
married     -0.34606    0.02423  -14.29   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6136 on 2564 degrees of freedom
Multiple R-squared:  0.07372, Adjusted R-squared:  0.07336 
F-statistic: 204.1 on 1 and 2564 DF,  p-value: < 2.2e-16
```

Propensity Score Matching

3. How to estimate the size of treatment effect on the treated?

3. Estimate the size of the average treatment effect on the treated

The model is used to impute the value that the outcome variable would take among the treated units if those treated units were actually controls.

Fit a model to the matched data and create simulated predicted values of the dependent variable for the treated units with T_i switched counterfactually from 1 to 0.

Then, given this fitted model, the missing outcomes Yi(0) are imputed for the matched treated units by using the values of the explanatory variables for the treated units.

In this way, we get an estimate of what values the treated units would have taken if those treated units were actually controls.

Some bad news ...

PA

Why Propensity Scores Should Not Be Used for Matching

Gary King[1] **and Richard Nielsen**[2]

[1] *Institute for Quantitative Social Science, Harvard University, 1737 Cambridge Street, Cambridge, MA 02138, USA. Email: king@harvard.edu, URL: http://GaryKing.org*
[2] *Department of Political Science, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA. Email: rnielsen@mit.edu, URL: http://www.mit.edu/~rnielsen*

Abstract

We show that propensity score matching (PSM), an enormously popular method of preprocessing data for causal inference, often accomplishes the opposite of its intended goal—thus increasing imbalance, inefficiency, model dependence, and bias. The weakness of PSM comes from its attempts to approximate a completely randomized experiment, rather than, as with other matching methods, a more efficient fully blocked randomized experiment. PSM is thus uniquely blind to the often large portion of imbalance that can be eliminated by approximating full blocking with other matching methods. Moreover, in data balanced enough to approximate complete randomization, either to begin with or after pruning some observations, PSM approximates random matching which, we show, increases imbalance even relative to the original data. Although these results suggest researchers replace PSM with one of the other available matching methods, propensity scores have other productive uses.

What to do instead?

Use CEM, coarsened exact matching, because it better approximates a fully blocked experimental design

Statistical Modeling, Causal Inference, and Social Science

Home Authors Blogs We Read Sponsors

« Kristof/Brooks update: NYT columnists correct their mistakes!

Ma conférence 11 h, lundi 23 juin à l'Université Paris Dauphine »

It's not matching *or* regression, it's matching *and* regression.

Posted on June 22, 2014 1:36 PM by Andrew

A colleague writes:

> *Why do people keep praising matching over regression for being non parametric? Isn't it f'ing parametric in the matching stage, in effect, given how many types of matching there are... you're making structural assumptions about how to deal with similarities and differences.... the*

- Art
- Bayesian Statistics
- Causal Inference
- Decision Theory
- Economics
- Jobs
- Literature
- Miscellaneous Scie
- Miscellaneous Stat
- Multilevel Modeling
- Political Science
- Public Health
- Sociology
- Sports
- Stan
- Statistical computi
- Statistical graphics
- Teaching

2. A first differences example: financial satisfaction

Some organizing ...

```
panel=read.csv(file.choose())

library(QMSS)
library(plyr)
library(psych)
library(VGAM)
library(plm)

pd <- arrange(panel,idnum,panelwave)
```

An example

If someone increases their family income, do they also increase their satisfaction with their present financial situation?

Financial satisfaction

"We are interested in how people are getting along financially these days. So far as you and your family are concerned, would you say that you are (1) pretty well satisfied with your present financial situation, (2) more or less satisfied, or (3) not satisfied at all?"

```
# make reverse-coded version of "satfin" variable called "n.satfin"

> pd$n.satfin <- ReverseThis(pd$satfin)

> Tab(pd$n.satfin)
  Count   Pct Cum.Pct
1  1320 27.52   27.52
2  2102 43.82   71.34
3  1375 28.66  100.00

> with(pd, table(satfin, n.satfin)) ## compare the recode ##
      n.satfin
satfin    1    2    3
     1    0    0 1375
     2    0 2102    0
     3 1320    0    0
```

The simplest OLS results, with clustered S.E.s

```
> pd$realinc10k <- pd$realinc/10000

> # make subset of data with needed variables for faster processing
> pd.sub <- pd[,c("idnum","panelwave","n.satfin","realinc10k")]


> ols.satfin <- plm(n.satfin ~ realinc10k, data = pd.sub, index =
c("idnum", "panelwave"), model = "pooling")

> clusterSE(ols.satfin, cluster.var = "idnum")

t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept) 1.7596638  0.0200967  87.560 < 2.2e-16 ***
realinc10k  0.0693951  0.0038919  17.831 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The simplest OLS results, with clustered S.E.s

For every $10k increase in someone's family income, there is a 0.069*** point increase in their satisfaction with their financial situation, indicating greater financial satisfaction (the t-stat goes from ≈21 to ≈18)

```
> ols.satfin <- plm(n.satfin ~ realinc10k, data = pd.sub,
+                   index = c("idnum", "panelwave"),
+                   model = "pooling")

> clusterSE(ols.satfin, cluster.var = "idnum")

t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept) 1.7596638  0.0200967  87.560 < 2.2e-16 ***
realinc10k  0.0693951  0.0038919  17.831 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Run OLS regression with clustered standard errors I

- Remember, with a panel we have the same person multiple times (2000 individuals x 3 waves = 6000 “person-years”)
- That means that we really don’t have 6000 independent observations; we have less than that

Run OLS regression with clustered standard errors II

- If we act like we have 6000 independent observations, then we will underestimate our standard errors, because observations are serially correlated across waves
- We should apply clustered standard errors, which relax the independence assumption of i.i.d. errors

In matrix form, homoskedasticity

$$Var(\beta)=\sigma^2(X'X)^{-1}\,X'\sigma^2 I X\,(X'X)^{-1}$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

In matrix form, robust standard errors

$$Var(\beta)=\sigma^2(X'X)^{-1}\, X'\sigma^2\boldsymbol{\Omega}X\,(X'X)^{-1}$$

$$\begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ & & \\ 0 & 0 & \sigma_n^2 \end{bmatrix}$$

In matrix form, clustered standard errors

$$\mathrm{Var}(\beta)=\sigma^2(X'X)^{-1}\, X'\sigma^2\mathbf{C}X\,(X'X)^{-1}$$

$$\begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \cdots & \sigma_{2p} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \cdots & \sigma_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \sigma_{p3} & \cdots & \sigma_p^2 \end{pmatrix}$$

Why isn’t this good enough?

- We might imagine that even after controlling for many, many things, people who make more money are just fundamentally and essentially different from people who make less money

- Maybe wealthier people (even if they weren’t wealthier), would be more satisfied with their financial situation; maybe they are just positive and happy with whatever they have

Why isn’t this good enough?

Why isn’t this good enough?

- Maybe wealthy people held the same opinion on their financial situation even before they got more money

- Maybe earning more didn’t change their opinion at all.

- Maybe high income people are truly incomparable with low income people

- This is known as “individual heterogeneity”

How can we overcome this very fundamental concern?

- We could run an experiment where we randomly gave more money to some people and not to others; and
- Then we could see if their opinions on their financial situations change
- Quasi-experiments like this have been run (lotteries, tax credits, welfare, etc.) …

The unobserved error

We can now imagine that our equation has two errors:

$$Y_{it} = \alpha_0 + \beta_1 x_{it} + a_i + u_{it}$$

$a_{i} =$ the unobserved, time-invariant factors that affect y_{it}

u_{it} = the idiosyncratic, time-varying factors that affect y_{it}

The problem with running the simple naïve OLS regression

We want to allow the unmeasured factors in a_i (whether personality, genetics, or other factors) that affect earning money to also be correlated with feelings of satisfaction

Remember how first differencing works

- For each variable, we subtract the old value (at time t) from the new value (at time t+1)
- For constant variables, their difference goes to zero
- E.g., female at t (1) – female at t+1 (1)

  = 1-1=0
- So all constant variables drop out of the equation (since they are all zeros)

What happens to the error when we difference the equation?

If we take the difference between the 2 time periods:

$$\Delta Y_{it} = \beta_0 + \beta_1 \Delta x_{it} + \Delta a_i + \Delta u_{it}, \ t=1,2$$

The $\Delta a_i = 0$ because a_i are the unobserved, time-invariant factors that affect y_{it} … YAY!

All that is left of the error is the Δu_{it} which are the idiosyncratic, time-varying factors that affect y_{it} (which we hope is really random)

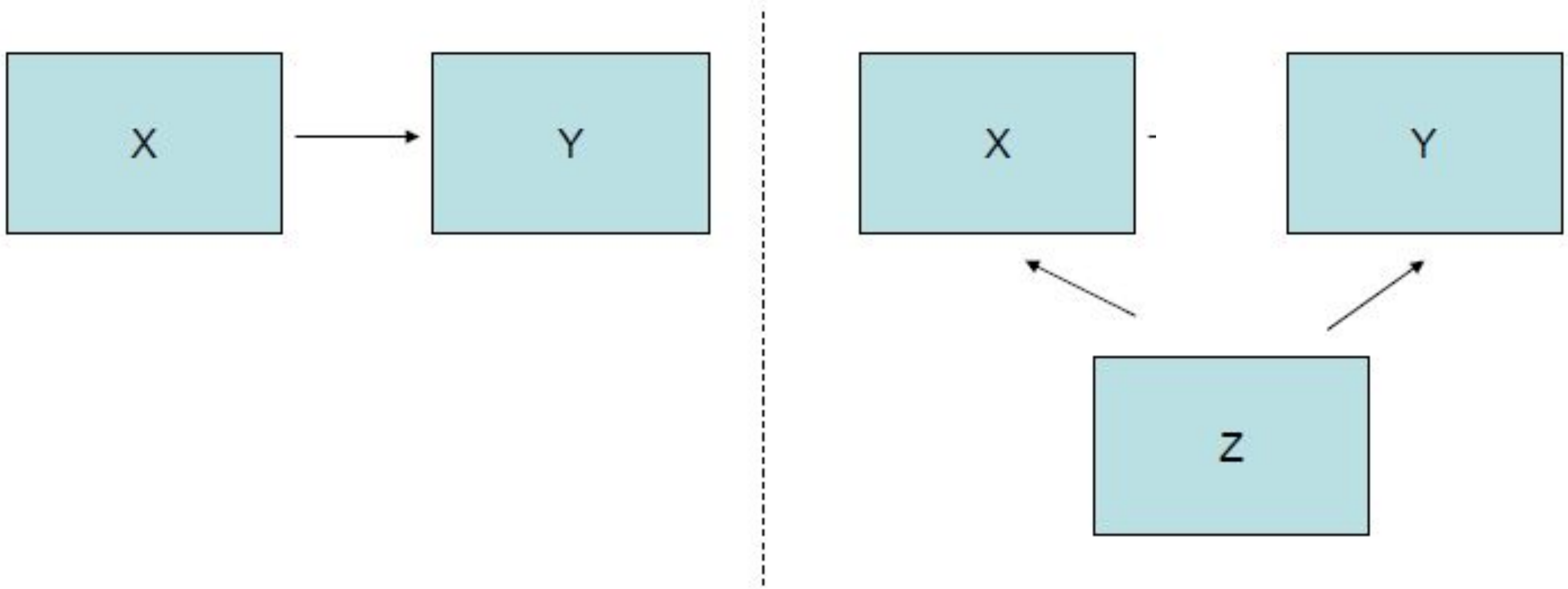
First differencing, cont'd

- For all time-varying variables, we just get the result of the old value (at time t) being subtracted from the new value (at time t+1)
- Then, we just regress these differenced X variables on the differenced Y (independent variable)
- We are estimating the effect of *changes* in the explanatory variables on *changes* in the dependent variable

Here is how we can make more causal statements

- This is essentially a before-and-after portrait
- Since we can control for “you” (what stably makes you *you*), then any semi-exogenous shocks to you will produce changes in you that are a result of the shocks and not who you stably are
- We can rule out a certain form of spurious correlation (linked to your stable error term)

Remember spurious correlation



First, pooled OLS; now, using plm

```
> pooled.satfin <- plm(n.satfin ~ realinc10k + panelwave, index=c("idnum",
"panelwave"), model="pooling", data=d)

> summary(pooled.satfin)
Oneway (individual) effect Pooling Model

Unbalanced Panel: n=1879, T=1-3, N=4269

Coefficients :
              Estimate Std. Error t-value Pr(>|t|)
(Intercept)  1.7943626  0.0205819 87.1815  < 2e-16 ***
realinc10k   0.0694776  0.0032844 21.1537  < 2e-16 ***
panelwave2  -0.0546719  0.0256957 -2.1277  0.03342 *
panelwave3  -0.0630928  0.0272278 -2.3172  0.02054 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    2405
Residual Sum of Squares: 2173.5
R-Squared      :  0.096262
      Adj. R-Squared :  0.096172
F-statistic: 151.43 on 3 and 4265 DF, p-value: < 2.22e-16
```

These results are the same as earlier

The first differenced results

```
> fd.satfin <- plm(n.satfin ~ realinc10k + panelwave, index=c("idnum",
"panelwave"), model="fd", data=pd.sub)

> summary(fd.satfin)

Oneway (individual) effect First-Difference Model

Call:
plm(formula = n.satfin ~ realinc10k + panelwave, data = pd.sub,
    model = "fd", index = c("idnum", "panelwave"))

Unbalanced Panel: n=1879, T=1-3, N=4269

Coefficients :
              Estimate Std. Error t-value  Pr(>|t|)
(intercept)  0.1322291  0.1388286  0.9525    0.3410
realinc10k   0.0434935  0.0059709  7.2843 4.373e-13 ***
panelwave2  -0.1710695  0.1371856 -1.2470    0.2125
panelwave3  -0.3096928  0.2723985 -1.1369    0.2557
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    1385
Residual Sum of Squares: 1354
R-Squared      :  0.02238
      Adj. R-Squared :  0.022343
F-statistic: 18.2072 on 3 and 2386 DF, p-value: 1.1127e-11
```

The first differenced results

For every $10k positive *change* in someone's family income, it produces a 0.043*** point positive *change* in their financial satisfaction, on average, for the same person across 3 waves of data, net of wave

```
> summary(fd.satfin)
Oneway (individual) effect First-Difference Model

Call:
plm(formula = n.satfin ~ realinc10k + panelwave, data = pd.sub,
    model = "fd", index = c("idnum", "panelwave"))

Unbalanced Panel: n=1879, T=1-3, N=4269

Coefficients :
              Estimate Std. Error t-value  Pr(>|t|)
(intercept)  0.1322291  0.1388286  0.9525    0.3410
realinc10k   0.0434935  0.0059709  7.2843 4.373e-13 ***
panelwave2  -0.1710695  0.1371856 -1.2470    0.2125
panelwave3  -0.3096928  0.2723985 -1.1369    0.2557
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



```
Total Sum of Squares:    1385
```

The first differenced results

- Apparently, when someone's family earns more money, they actually become more satisfied with their financial situation

- Apparently, the process of earning more money is at least partially driving the results, not only that "the type of people who earn more money" are more inclined to just be happier with their financial situation, whatever it is.

What to make of the coefficients

- But the coefficient from the first difference model is 0.043***, vs. 0.069*** in the naïve OLS regression

- So there is a meaningful drop (38% reduction) in the size of the coefficient between the types of models, so this might -- at first sight -- seem to imply that there may be something to our original critique … that some unmeasured traits make people both earn more money and be more financially happy, but this is more likely measurement error

What to make of the adjusted R-sqs?

- Look at the adjusted R-sqs, however. The naïve OLS had an adjusted R2 = 0.095, while the first differences had an adjusted R2 = 0.021.

- The first differences adj. R2 is almost 5 times smaller than the naive OLS.

3. From our first differences model to the fixed effects model

I am going to rework my original example

- I am going to make it a balanced 2-wave panel

This is for teaching purposes only – you will never need to do this … do not follow this code for anything !!!!!

I start with this recoding …
for teaching purposes only !!!!

```
> # take only obs for individuals without missingness on "n.satfin" and
"realinc10k" for both waves 1 and 2 and drop all obs from panelwave 3
(for demonstration purposes only)

> good_ids1 <- with(pd.sub, idnum[which(!is.na(n.satfin) &
!is.na(realinc10k) & panelwave==1)])

> good_ids2 <- with(pd.sub, idnum[which(!is.na(n.satfin) &
!is.na(realinc10k) & panelwave==2)])

> temp <- subset(pd.sub, idnum %in% good_ids1 & idnum %in% good_ids2 &
panelwave < 3)
```

- This “goodids” gives me only people who answered all my questions for the first 2 waves of the data only

The first differenced results

For every $10k positive *change* in someone's family income, it produces a 0.028*** point positive *change* in their financial satisfaction, for the same person across the first 2 waves of this panel

```
> fd.satfin2 <- plm(n.satfin ~ realinc10k, index = c("idnum", "panelwave"), model
= "fd", data = temp)

> summary(fd.satfin2)
Oneway (individual) effect First-Difference Model

Call:
plm(formula = n.satfin ~ realinc10k, data = temp, model = "fd",
    index = c("idnum", "panelwave"))

Balanced Panel: n=1255, T=2, N=2510

Coefficients :
             Estimate Std. Error t-value  Pr(>|t|)
(intercept) -0.035765   0.021510 -1.6627 0.0966261 .
realinc10k   0.027870   0.007950  3.5057 0.0004715 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The first differenced results

Continued ...

```
> fd.satfin2 <- plm(n.satfin ~ realinc10k, index = c("idnum", "panelwave"), model
= "fd", data = temp)

> summary(fd.satfin2)
Oneway (individual) effect First-Difference Model

Call:
plm(formula = n.satfin ~ realinc10k, data = temp, model = "fd",
    index = c("idnum", "panelwave"))

Balanced Panel: n=1255, T=2, N=2510

Coefficients :
              Estimate Std. Error t-value  Pr(>|t|)
(intercept) -0.035765   0.021510 -1.6627 0.0966261 .
realinc10k   0.027870   0.007950  3.5057 0.0004715 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    730.85
Residual Sum of Squares: 723.75
R-Squared      :  0.0097131
      Adj. R-Squared :  0.0096976
F-statistic: 12.2899 on 1 and 1253 DF, p-value: 0.0004715
```

Dummy variable model

```
> dummy.satfin <- lm(n.satfin ~ realinc10k + panelwave + as.factor(idnum),
data = temp)

> summary(dummy.satfin)$coef[1:3,] # don't print the nearly 2000 coefficients
for the dummies
               Estimate  Std. Error   t value     Pr(>|t|)
(Intercept)  1.47695596 0.381864491  3.867749 0.0001154885
realinc10k   0.02787012 0.007949958  3.505695 0.0004714987
panelwave   -0.03576505 0.021510448 -1.662683 0.0966260753
```

Dummy variable model

For every $10k positive *change* in someone's family income, it produces a 0.028*** point positive *change* in their financial satisfaction, net of any particular person, across the first 2 waves of this panel

```
> dummy.satfin <- lm(n.satfin ~ realinc10k + panelwave + as.factor(idnum),
data = temp)

> summary(dummy.satfin)$coef[1:3,] # don't print the nearly 2000 coefficients
for the dummies
               Estimate  Std. Error   t value     Pr(>|t|)
(Intercept)  1.47695596 0.381864491  3.867749 0.0001154885
realinc10k   0.02787012 0.007949958  3.505695 0.0004714987
panelwave   -0.03576505 0.021510448 -1.662683 0.0966260753
```

Dummy variable model

This model assumes the same slope for every individual, on average

But it allows the intercepts to be different for each person

That is: Some people just have greater financial satisfaction, net of their actual family income level, because of perhaps (relatively-stable) omitted variables not included in the model

(Notice also that I put in a panelwave variable to capture trend over time for everyone.)

What happens when we run a fixed effects equation? - I

If we take the difference between the 2 time periods:

$$Y_{it} = \beta_0 + \beta_1 x_{it} + a_i + u_{it}, \ t=1,2$$

The a_i = is a series of dummy variables representing the average unobserved, time-invariant factors that affect y_{it}

Beyond the typical way

- What I will do here is use a person earlier as a natural control for themselves later: this is known as fixed effects.

- The demeaned fixed effects model looks like this:

$$(Y_{ij} - \overline{Y_i}) = \beta_1 (X_{ij} - \overline{X_i}) + (u_{ij} - \overline{u_i})$$

where *j* is for the individual year and *i* is the person

Fixed effects

Use “within” for the plm function

```
> fe.satfin <- plm(n.satfin ~ realinc10k + panelwave,index=c("idnum",
"panelwave"), model="within", # set model = "within" for fixed effects
data= temp)

> summary(fe.satfin)
Oneway (individual) effect Within Model

Call:
plm(formula = n.satfin ~ realinc10k + panelwave, data = temp,
    model = "within", index = c("idnum", "panelwave"))

Balanced Panel: n=1255, T=2, N=2510

Coefficients :
            Estimate Std. Error t-value  Pr(>|t|)
realinc10k  0.027870   0.007950  3.5057 0.0004715 ***
panelwave2 -0.035765   0.021510 -1.6627 0.0966261 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    366
Residual Sum of Squares: 361.88
R-Squared       :  0.01127
      Adj. R-Squared :  0.0056259
```

Fixed effects

For every $10k positive *change* in someone's family income, it produces a 0.028*** point positive *change* in their financial satisfaction, net of any particular person, across the first 2 waves of this panel

```
> fe.satfin <- plm(n.satfin ~ realinc10k + panelwave,index=c("idnum",
"panelwave"), model="within", # set model = "within" for fixed effects
data= temp)

> summary(fe.satfin)
Oneway (individual) effect Within Model

Call:
plm(formula = n.satfin ~ realinc10k + panelwave, data = temp,
    model = "within", index = c("idnum", "panelwave"))

Balanced Panel: n=1255, T=2, N=2510

Coefficients :
            Estimate Std. Error t-value  Pr(>|t|)
realinc10k  0.027870   0.007950  3.5057 0.0004715 ***
panelwave2 -0.035765   0.021510 -1.6627 0.0966261 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fixed effects

What about all this stuff down the bottom?

```
> summary(fe.satfin)
[omitted]

Total Sum of Squares:    366
Residual Sum of Squares: 361.88
R-Squared      :  0.01127
      Adj. R-Squared :  0.0056259
F-statistic: 7.14094 on 2 and 1253 DF, p-value: 0.00082465



> #get sigma_u, sigma_e, rho (using sigmaRho function in QMSS package)
> sigmaRho(fe.satfin)
sigma_u =  0.61164
sigma_e =  0.53741
    rho =  0.56434 (fraction of variance due to u_i)
```

Rho

- What does rho = 0.56 mean?
- Rho = Proportion of error variance due to unit effects (the fact that these 2 observations came from the same person)
- 56% of the error variance is due to the fact that the same person is being analyzed each time, as opposed to any time-varying factors coming into play

More on Rho

- If Rho=0, that would mean that knowing the same person was being analyzed each time would not at all help predict their financial satisfaction (the changes would totally trump)
- If Rho=1, that would mean that knowing the same person was being analyzed would totally predict their average financial satisfaction (the attitude is totally set, immune to outside changes)

Always include dummies for the Waves in fixed effects

This will account for any time-specific events affecting all observations at that time

This is like the “difference in differences” model where we need to account for common trends over time across treatment and control groups

All of these R-squares?

R-squares for difference models or fixed effects are not worth much

When we put in dummies for each person, naturally our R-sq goes up (because most of the variation in ave. happiness can be accounted for due to person effects)

The within regression R-sq is the most reliable for our purposes, but it is still not particularly informative

Multivariate fixed effects model

Same as before with first differences

What if I didn't make the panel be balanced?

```
> summary(fe.satfin2)
Oneway (individual) effect Within Model

Call:
plm(formula = n.satfin ~ realinc10k + panelwave, data = pd.sub,
    model = "within", index = c("idnum", "panelwave"))

Unbalanced Panel: n=1879, T=1-3, N=4269

Coefficients :
             Estimate Std. Error t-value  Pr(>|t|)
realinc10k  0.0421566  0.0060423  6.9770 3.892e-12 ***
panelwave2 -0.0373494  0.0213513 -1.7493   0.08037 .
panelwave3 -0.0415640  0.0228971 -1.8152   0.06961 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    713.83
Residual Sum of Squares: 698.21
R-Squared      :  0.021892
      Adj. R-Squared :  0.012241
F-statistic: 17.8084 on 3 and 2387 DF, p-value: 1.9762e-11

> sigmaRho(fe.satfin2)
sigma_u =  0.62257
sigma_e =  0.54084
    rho =  0.56991 (fraction of variance due to u_i)
```

What if I didn't make the panel be balanced?

For every $10k positive *change* in someone's family income, it produces a 0.028*** point positive *change* in their financial satisfaction, for the same person across the 3 waves of this panel. People are only in the panel 2.3 times.

```
> summary(fe.satfin2)
Oneway (individual) effect Within Model

Call:
plm(formula = n.satfin ~ realinc10k + panelwave, data = pd.sub,
    model = "within", index = c("idnum", "panelwave"))

Unbalanced Panel: n=1879, T=1-3, N=4269

Coefficients :
             Estimate Std. Error t-value  Pr(>|t|)
realinc10k  0.0421566  0.0060423  6.9770 3.892e-12 ***
panelwave2 -0.0373494  0.0213513 -1.7493   0.08037 .
panelwave3 -0.0415640  0.0228971 -1.8152   0.06961 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

What if I had more waves than 2?

Things start to diverge a tiny bit …

Remember: The first differenced results

For every $10k positive *change* in someone's family income, it produces a 0.043*** point positive *change* in their financial satisfaction, for the same person across 3 waves of data, net of wave

```
. reg d.nsatfin d.realinc10k b2.panelwave, cluster(idnum)

Linear regression                                      Number of obs =    2359
                                                       F(  2,  1356) =   22.95
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0214
                                                       Root MSE      =   .7514

                               (Std. Err. adjusted for 1357 clusters in idnum)
------------------------------------------------------------------------------
             |               Robust
   D.nsatfin |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  realinc10k |
         D1. |   .0430892   .0063612     6.77   0.000     .0306105     .055568
             |
 3.panelwave |   .0322156   .0364819     0.88   0.377    -.0393515    .1037828
       _cons |  -.0387609    .021544    -1.80   0.072    -.0810242    .0035024
```


*

Now: Fixed effects for a 3 wave panel

```
. xtreg nsatfin realinc10k i.panelwave, fe robust

Fixed-effects (within) regression               Number of obs      =      4269
Group variable: idnum                           Number of groups   =      1879

R-sq:  within  = 0.0219                         Obs per group: min =         1
       between = 0.1236                                        avg =       2.3
       overall = 0.0962                                        max =         3

                                                F(3,1878)          =     13.65
corr(u_i, Xb)  = 0.1536                         Prob > F           =    0.0000

                               (Std. Err. adjusted for 1879 clusters in idnum)
------------------------------------------------------------------------------
             |               Robust
     nsatfin |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  realinc10k |   .0421566   .0068934     6.12   0.000     .0286371    .0556762
             |
   panelwave |
          2  |  -.0373495   .0212712    -1.76   0.079    -.0790672    .0043683
          3  |   -.041564   .0235617    -1.76   0.078    -.0877739     .004646
             |
       _cons |   1.877946   .0262665    71.50   0.000     1.826431     1.92946
-------------+----------------------------------------------------------------
     sigma_u |  .62257247
     sigma_e |  .54083608
         rho |  .56991089   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Now: Fixed effects for a 3 wave panel

For every $10k positive *change* in someone's family income, it produces a 0.042*** point positive *change* in their financial satisfaction, for the same person, across 3 waves of data

```
> summary(fe.satfin2)
Oneway (individual) effect Within Model

Call:
plm(formula = n.satfin ~ realinc10k + panelwave, data = pd.sub,
    model = "within", index = c("idnum", "panelwave"))

Unbalanced Panel: n=1879, T=1-3, N=4269

Coefficients :
             Estimate Std. Error t-value  Pr(>|t|)
realinc10k  0.0421566  0.0060423  6.9770 3.892e-12 ***
panelwave2 -0.0373494  0.0213513 -1.7493   0.08037 .
panelwave3 -0.0415640  0.0228971 -1.8152   0.06961 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    713.83
Residual Sum of Squares: 698.21
R-Squared      :  0.021892
      Adj. R-Squared :  0.012241
```

Okay. To be clear: this last model on the previous slide is the only kind of fixed effects model you actually ever really need to run in practice … all the other models were just there to highlight how fixed effects and first differences can be similar, under certain conditions. Okay?

First differences *vs*. fixed effects

With 3 waves of data at work, the coefficients on family income are:

First differences = **0.043** (p=.000)

Fixed effects = **0.042** (p=.000)

Why are first differences and fixed effects different now at all?

With 3 waves of data at work, first differences does the difference between Wave 1 and Wave2 and then does the difference between Wave 2 and Wave 3 (but Wave 1 and Wave 3 are completely unrelated)

Fixed effects takes all 3 waves of data and deviates each year from the average for that person over 3 waves

Why are first differences and fixed effects different now at all?

Also, fixed effects allows people to enter into the “between” estimates even if they contribute nothing to the “within” estimates

That is, fixed effects allows people to be counted even if they are only in the data once, see the “min” on the fixed effects output, (and hence have no first difference to offer) – this can be consequential for some estimates in the model

For first differences

Risk of serial correlation grows, as number of waves increases (especially when the n is small)

That is why fixed effects is usually preferred

Could I run ordinal logit fixed effects?

- No. Such a model does not have very good properties, except when your units (i.e., persons) have 20 to 30 observations each. Even then, it is questionable.

- But you can easily do a fixed effects conditional binary logit, among other kinds of models …

Unbalanced panel

Why unbalanced?

How unbalanced?

4. Some considerations

This issue of variance …

- Differences in Xs and Ys often have much less variance than the distribution of the original Xs and Ys; this makes it harder to get efficient estimates of coefficients (i.e., large standard errors)

Our Xs

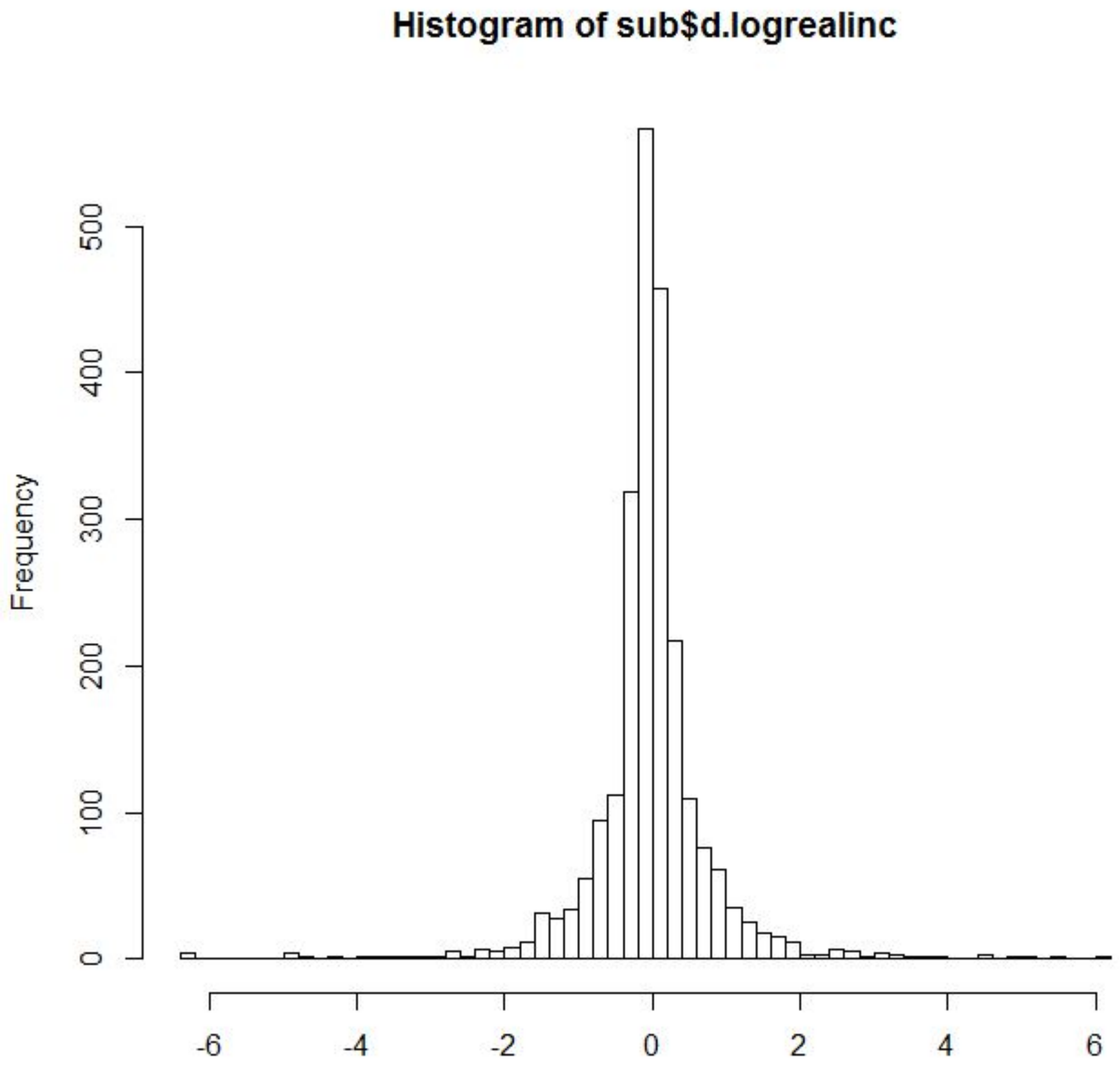
We are regressing the CHANGES in Y on the CHANGES in X, so they are invariably on a different scale from their original variables

```
> describe(sub$logrealinc)
  vars    n  mean   sd median trimmed mad  min   max range  skew kurtosis   se
1    1 4272 10.01 1.07  10.15   10.08 0.9 5.56 11.89  6.34 -1.02     2.37 0.02
```

```
> describe(sub$d.logrealinc)
  vars    n mean   sd median trimmed mad   min max range  skew kurtosis   se
1    1 2362    0 0.85  -0.03       0 0.3 -6.34 6.1 12.44 -0.37    14.52 0.02
```

Looking at “locally appropriate”

- These models are essentially identified by the “changers”
- Everybody changed some, but most not by much



This issue of variance …

- There is much less structured variance in the first difference model because it is correlating changes in Ys with changes in Xs … but much of the change in Y or X in a given year is random noise

- Also, with first differences, outliers can have a greater effect on coefficients but also, then, R-sq

- That is why the first differences adj. R2 is 7 times smaller than the naive OLS, in our case

Looking at “locally appropriate”

- These models are essentially identified by the “changers” (a small part of the total sample)

Hey, wait a minute! Δ in variables?

- Why would there even be coefficients at all on half of the variables in the data set
- There should not be changes in variables that don't change.
- Sex doesn't change, usually. Race doesn't change, usually. Age should be going up constantly for everyone (i.e., everyone should be 2 years older one wave to the next wave, so Δage should = 2 for everyone). What is going on here?

One other issue

- Classical errors-in-variables would cause a greater problem for fixed effects models and bias our fixed effects estimates toward zero

- How much bias is introduced by measurement error? It is a function of observed reliability (R_o) and the correlation between siblings on education (ρ)

$$(R_\Delta) = R_o(1 - \rho) \,/\, 1 - \rho R_o \;=\; 0.85*(1 - 0.5) \,/\, 1 - (0.7*0.85) = 0.63$$

- Our FE estimate is attenuated probably by -1+0.63 = -0.37 (or 37% too low)
- From Card, David. "The causal effect of education on earnings." *Handbook of labor economics* 3 (1999): 1801-1863.

Measurement error ….

Union Wage Effect Estimates with Union Status Measurement Error Reduced Through Averaging

	Actual variables		*2 year averages (1991–1992), (1993–1994) and (1995–1996)*		*3 year averages (1991–1993) and (1994–1996)*	
	OLS	*Fixed-effects*	*OLS*	*Fixed-effects*	*OLS*	*Fixed-effects*
Waves 1-6						
Female employees						
Member V	0.088	0.024	0.101	0.049	0.106	0.065
	(5.63)	(1.82)	(4.44)	(2.43)	(3.71)	(2.23)
Adj. R^2	0.487	0.229	0.535	0.349	0.555	0.444
No. of observations	3,294		1,647		1,098	
No. of individuals	549		549		549	
Male manual full-time employees						
Member V	0.104	0.050	0.102	0.076	0.121	0.251
	(4.24)	(1.49)	(3.23)	(1.69)	(3.10)	(3.44)
Adj. R^2	0.265	0.220	0.321	0.360	0.347	0.482
No. of observations	960		480		320	
No. of individuals	160		160		160	

Swaffield, Joanna K. "Does Measurement Error Bias Fixed-effects Estimates of the Union Wage Effect?." *Oxford Bulletin of Economics and Statistics* 63.4 (2001): 437-457.

Measurement error ….

Union Wage Effect Estimates with Union Status Measurement Error Reduced Through Averaging

	Actual variables		2 year averages (1991–1992), (1993–1994) and (1995–1996)		3 year averages (1991–1993) and (1994–1996)	
	OLS	*Fixed-effects*	*OLS*	*Fixed-effects*	*OLS*	*Fixed-effects*
Waves 1-6						
Female employees						
Member V	0.088	0.024	0.101	0.049	0.106	0.065
	(5.63)	(1.82)	(4.44)	(2.43)	(3.71)	(2.23)
Adj. R^2	0.487	0.229	0.535	0.349	0.555	0.444
No. of observations	3,294		1,647		1,098	
No. of individuals	549		549		549	
Male manual full-time employees						
Member V	0.104	0.050	0.102	0.076	0.121	0.251
	(4.24)	(1.49)	(3.23)	(1.69)	(3.10)	(3.44)
Adj. R^2	0.265	0.220	0.321	0.360	0.347	0.482
No. of observations	960		480		320	
No. of individuals	160		160		160	

See how when we average away some of the measurement error (by aggregating across years), the coefficient on "Member V" goes up

Item #5: Random effects vs. fixed effects

The question

Does being married make someone more disapproving of homosexual marriage or not?

Married people …

Half of people are married at a given time, while a small percentage change the marital status across waves

```
> # create indicator variable for "married"

> pd.sub$married <- ifelse(pd.sub$marital == 1, 1, 0)

> Tab(pd.sub$married)
  Count   Pct Cum.Pct
0  2455 51.06   51.06
1  2353 48.94  100.00

> # create first-differenced variables d.married and d.marhomo

> pd.sub <- ddply(pd.sub, "idnum", mutate,
+                 d.married = firstD(married),
+                 d.marhomo = firstD(marhomo))

> Tab(pd.sub$d.married)
   Count   Pct Cum.Pct
-1   117  4.17    4.17
0   2573 91.63   95.80
1    118  4.20  100.00
```

Is it okay for homosexuals to marry?

MARHOMO **HOMOSEXUALS SHOULD HAVE RIGHT TO MARRY**

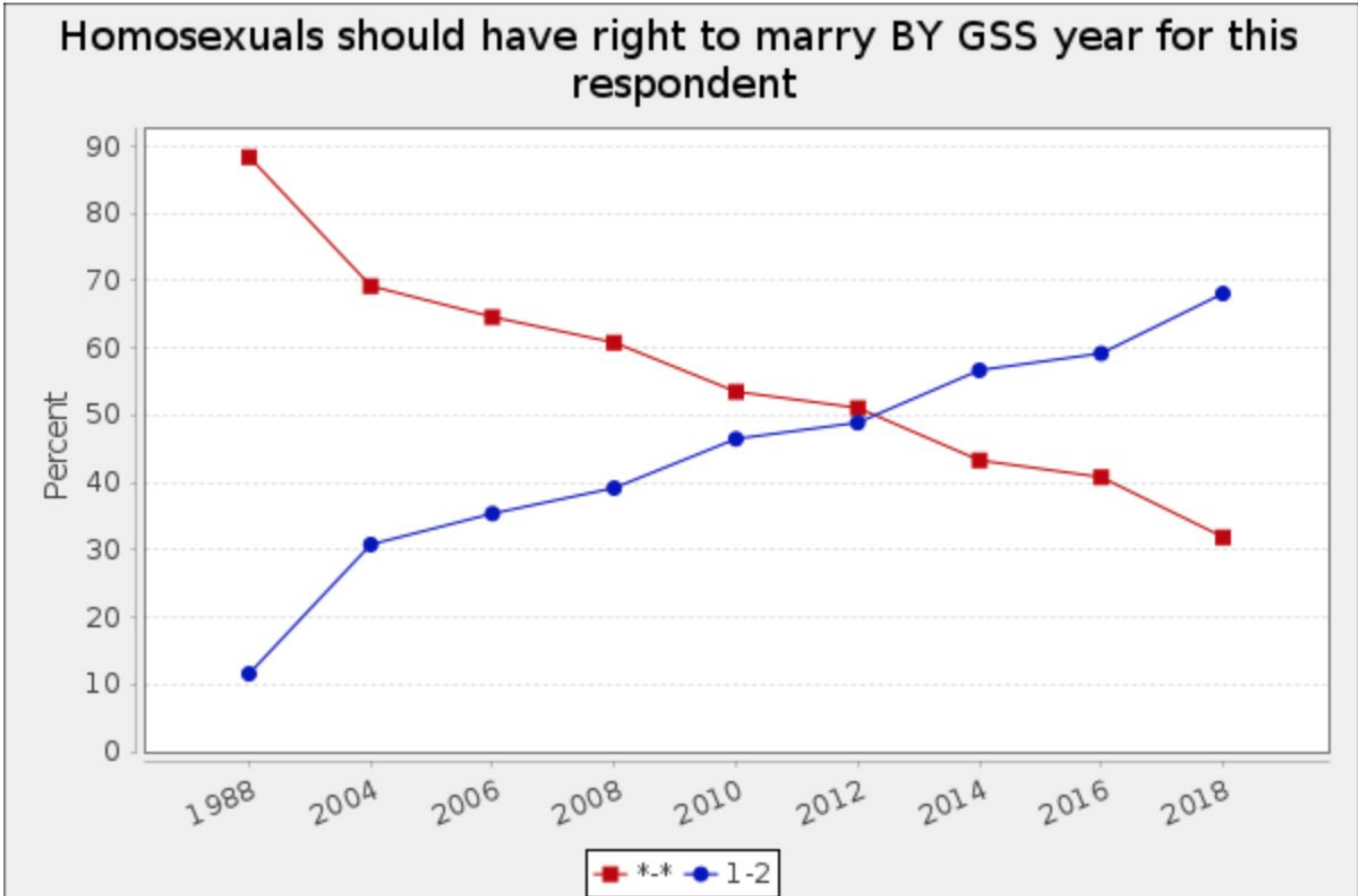
Description of the Variable

1280. Do you agree or disagree? j. Homosexual couples should have the right to marry one another.

Percent	N	Value	Label
15.6	1,302	1	STRONGLY AGREE
19.9	1,663	2	AGREE
13.4	1,118	3	NEITHER AGREE NOR DISAGREE
17.9	1,492	4	DISAGREE
33.3	2,783	5	STRONGLY DISAGREE
	48,487	0	IAP
	162	8	CANT CHOOSE
	54	9	NA
100.0	**57,061**		**Total**

```
> Tab(pd.sub$d.marhomo)
   Count   Pct Cum.Pct
-4    12  0.64    0.64
-3    45  2.40    3.04
-2   109  5.81    8.84
-1   332 17.69   26.53
0    987 52.58   79.12
1    285 15.18   94.30
2     61  3.25   97.55
3     30  1.60   99.15
4     16  0.85  100.00
```

Overall context



Or this way ...

```
> # Note: the firstD function in QMSS package can be used in different ways.
We could also have created the d.married and d.marhomo like this

> pd.sub$d.married <- firstD(married, idnum, pd.sub) # or with(pd.sub,
firstD(married, idnum))

> pd.sub$d.marhomo <- firstD(marhomo, idnum, pd.sub) # or with(pd.sub,
firstD(marhomo, idnum))

> Tab(pd.sub$d.married)
   Count   Pct Cum.Pct
-1   117  4.17    4.17
0   2573 91.63   95.80
1    118  4.20  100.00

> Tab(pd.sub$d.marhomo)
   Count   Pct Cum.Pct
-4    12  0.64    0.64
-3    45  2.40    3.04
-2   109  5.81    8.84
-1   332 17.69   26.53
0    987 52.58   79.12
1    285 15.18   94.30
2     61  3.25   97.55
3     30  1.60   99.15
4     16  0.85  100.00
```

The naïve (cross-sectional) OLS

```
> ols.marhomo <- plm(marhomo ~ married + panelwave, data = pd.sub, index =
c("idnum", "panelwave"), model = "pooling")

> clusterSE(fit = ols.marhomo, cluster.var = "idnum")

t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept)  3.208998   0.054220 59.1853 < 2.2e-16 ***
married      0.354701   0.073520  4.8245 1.468e-06 ***
panelwave2  -0.097465   0.037906 -2.5712   0.01018 *
panelwave3  -0.189778   0.046299 -4.0990 4.252e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The naïve (cross-sectional) OLS

Married people, net of the time trend, score 0.35 points higher on the disapproval of homosexual marriage scale

```
> ols.marhomo <- plm(marhomo ~ married + panelwave, data = pd.sub, index =
c("idnum", "panelwave"), model = "pooling")

> clusterSE(fit = ols.marhomo, cluster.var = "idnum")

t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept)  3.208998   0.054220 59.1853 < 2.2e-16 ***
married      0.354701   0.073520  4.8245 1.468e-06 ***
panelwave2  -0.097465   0.037906 -2.5712   0.01018 *
panelwave3  -0.189778   0.046299 -4.0990 4.252e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The question again

But does becoming married suddenly make someone more disapproving of homosexual marriage?

The fixed effects approach I

```
> summary(fe.marhomo)
Oneway (individual) effect Within Model

Call:
plm(formula = marhomo ~ married + panelwave, data = pd.sub, model = "within",
    index = c("idnum", "panelwave"))

Unbalanced Panel: n=1352, T=1-3, N=3232

Coefficients :
            Estimate Std. Error t-value  Pr(>|t|)
married     0.120369   0.091530  1.3151   0.18865
panelwave2 -0.076299   0.036425 -2.0947   0.03633 *
panelwave3 -0.180273   0.039176 -4.6016 4.473e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    1297
Residual Sum of Squares: 1281.4
R-Squared      :  0.012031
      Adj. R-Squared :  0.0069869
F-statistic: 7.61887 on 3 and 1877 DF, p-value: 4.6088e-05

> sigmaRho(fe.marhomo)
sigma_u =  1.37782
sigma_e =  0.82625
    rho =  0.7355 (fraction of variance due to u_i)
```

The fixed effects approach I

When someone becomes married, their disapproval of homosexual marriage score goes up 0.12 (stat. insig.), net of the time trends

```
> summary(fe.marhomo)
Oneway (individual) effect Within Model

Call:
plm(formula = marhomo ~ married + panelwave, data = pd.sub, model = "within",
    index = c("idnum", "panelwave"))

Unbalanced Panel: n=1352, T=1-3, N=3232

Coefficients :
            Estimate Std. Error t-value  Pr(>|t|)
married     0.120369   0.091530  1.3151   0.18865
panelwave2 -0.076299   0.036425 -2.0947   0.03633 *
panelwave3 -0.180273   0.039176 -4.6016 4.473e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    1297
Residual Sum of Squares: 1281.4
R-Squared      :  0.012031
```

The fixed effects approach II

About 75% of the variance in disapproval of homosexual marriage is attributable to differences between people, not due to the same people changing over time.

```
> sigmaRho(fe.marhomo)
sigma_u =  1.37782
sigma_e =  0.82625
    rho =  0.7355 (fraction of variance due to u_i)
```

Provisional conclusion

- In OLS, marital status appeared to be predictive of disapproval of homosexual marriage
- But then, when we look within the same person over time, the disapproval effect is smaller in magnitude and stat. insig.
- *Should we trust the fixed effects model more?*

About random effects

- Start with this equation again:

$$y_{it} = \beta x_{it} + \alpha_i + u_{it}$$

- Now, let us assume that α_i is an "individual-specific effect" drawn from a random distribution
- If that is so, then α_i would be uncorrelated with the other Xs in the model
- But, α_i is always there over time, and this induces positive serial correlation in our errors (because α_i is a part of our error term, too)

About random effects

- But if we have serial correlation in our errors, we will get incorrect standard errors and t-statistics
- So, we need to correct for this problem
- This is what random effects does, but how?
- Remember what fixed effects does: It subtracts the mean from each year's value
- Instead, random effects subtracts *a fraction* of the mean from each year's value

About random effects

That *fraction* of the mean for each year's value is calculated as lambda (λ)

$$\lambda = 1 - [\sigma_u^2/(\sigma_u^2 + T\sigma_a^2)]^{1/2},$$

Where λ depends on the variances α_i and u_{it} and on the number of time periods (T)

About random effects

Then that lambda (λ) is used to adjust all of the variables in the model, like such:

$$y_{it} - \lambda\bar{y}_i = \beta_0(1 - \lambda) + \beta_1(x_{it1} - \lambda\bar{x}_{i1}) + \dots + \beta_k(x_{itk} - \lambda\bar{x}_{ik}) + (v_{it} - \lambda\bar{v}_i),$$

When λ=0, the estimates from random effects mirror the OLS ones, but if λ=1, the estimates from random effects mimic fixed effects ones

About random effects

- Random effects is a form of Feasible Generalized Least Squares (FGLS) because lambda usually is not known directly, but it can be estimated
- FLGS runs a regression on data has been quasi-differenced, which means that the correlation (lambda-hat) across observations from the same person has been taken into account already

About random effects

- The critical assumption of random effects is that there is no correlation between α_i and the Xs in the model (i.e., there is no omitted variable problem, or self-selection issues)
- This assumption is much closer to OLS than to fixed effects, which is why random effects tend to produce coefficients (and p-values) more in line with OLS than fixed effects

About random effects

- Some scholars like random effects over fixed effects because they can include time-invariant characteristics in the model (like race, geography, sex, etc.)
- But …

The random effects approach I

Here is the set-up

```
> summary(re.marhomo)
Oneway (individual) effect Random Effect Model
   (Swamy-Arora's transformation)

Call:
plm(formula = marhomo ~ married + panelwave, data = pd.sub, model = "random",
    index = c("idnum", "panelwave"))

Unbalanced Panel: n=1352, T=1-3, N=3232

Effects:
                 var std.dev share
idiosyncratic 0.6827  0.8262 0.307
individual    1.5443  1.2427 0.693
theta  :
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.4463  0.6416  0.6416  0.6146  0.6416  0.6416
```

The random effects approach I

Being married increases someone's disapproval of homosexual marriage score by .27 points (stat. sig.), net of the time trends

```
> summary(re.marhomo)
Oneway (individual) effect Random Effect Model
   (Swamy-Arora's transformation)

Call:
plm(formula = marhomo ~ married + panelwave, data = pd.sub, model = "random",
    index = c("idnum", "panelwave"))

Unbalanced Panel: n=1352, T=1-3, N=3232

Coefficients :
             Estimate Std. Error t-value  Pr(>|t|)
(Intercept)  3.244885   0.049141 66.0319 < 2.2e-16 ***
married      0.275917   0.059462  4.6402 3.619e-06 ***
panelwave2  -0.082272   0.035747 -2.3015   0.02143 *
panelwave3  -0.183785   0.038378 -4.7888 1.753e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    2375.6
```

The random effects approach II

Being married increases someone's disapproval of homosexual marriage score by .27 points (stat. sig.), net of the time trends

```
> summary(re.marhomo)
Oneway (individual) effect Random Effect Model
   (Swamy-Arora's transformation)

Call:
plm(formula = marhomo ~ married + panelwave, data = pd.sub, model = "random",
    index = c("idnum", "panelwave"))

Unbalanced Panel: n=1352, T=1-3, N=3232

Total Sum of Squares:    2375.6
Residual Sum of Squares: 2202
R-Squared      :  0.073122
      Adj. R-Squared :  0.073031
F-statistic: 84.8322 on 3 and 3228 DF, p-value: < 2.22e-16
```

The random effects approach III

Around 70% of the variance in disapproval of homosexual marriage is attributable to differences between people, not due to the same people changing over time.

```
> # can also use sigmaRho function in QMSS package for random effects models
> sigmaRho(re.marhomo)
sigma_u =  1.24268
sigma_e =  0.82625
    rho =  0.69344 (fraction of variance due to u_i)
```

How do we choose between random and fixed effects?

- If we have a strong suspicion that there is a correlation between α_i and our Xs in the model (i.e., there is an omitted variable problem, or self-selection issues), then you should think about fixed effects
- *That is about it, except for the Hausman test …*

The Hausman test

- Hausman developed a seemingly simple test to decide whether to use fixed effects or random effects
- First, you run the fixed effects model, then the random effects one – and you compare the coefficients between them. If they are equivalent, then use random effects (because it is more efficient), but if they are different, then use fixed effects

The Hausman test

```
> phtest(fe.marhomo, re.marhomo)

        Hausman Test

data:  marhomo ~ married + panelwave
chisq = 5.9066, df = 3, p-value = 0.1162
alternative hypothesis: one model is inconsistent
```

We cannot reject the null that they are essentially the same coefficients

The Hausman test - results

- Random effects it is!
- That was easy, right?
- The Hausman test makes pretty strong assumptions; I wouldn't put too much weight on it, but if you really want to use random effects, that gives you the chance
- But I almost always prefer fixed effects because it tries to deal with omitted variables explicitly. (But what do we want the β to be, really?)

A bigger random effects model

You can add additional controls to these random effects models too