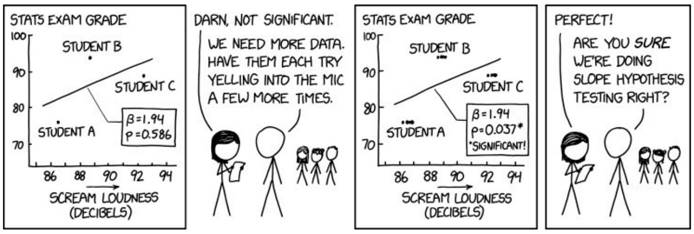
Intro to Econometrics: Pre-Final Review

Matthew Alampay Davis
December 17, 2021

I've prepared some notes that I think may be helpful in guiding and structuring your revision heading into next week's exams. The attached is not meant to be a sufficient summary of each topic; you'll still be best served going through the problem sets and practice problems (my recitation folders sometimes contain additional practice problems, solutions, and notes) and visit topics you are uncertain about in the lecture slides and textbook. That said, students have appreciated similar notes in the past so I hope you'll find them helpful too.
I'll also flag that I have not seen the final exam so the points I address here contain no information about what to expect. Mainly, I just hate when I have to take points off for an error or misunderstanding that could be pre-empted by placing emphasis on certain aspects that may be easy to gloss over or forget. The intention here is to provide more intuition for topics that I think can be confusing and to raise questions that you may find it fruitful to investigate yourelf. Good luck with the exam and thanks for being a great class this year!

1 Panel Data

- Please see the attached notes on fixed effects models for a demonstration of the difference between pooled OLS and fixed effects models, using an extreme case where their respective estimators lead to wildly different conclusions
- Here's an xkcd illustration that describes the importance of clustering standard errors:



By clustering standard errors at the student level, we treat errors as uncorrelated *across* entities while permitting error correlation *within* entities (here, a student). The p-value discussion in the comic shows how failing to account for this can lead us to bad inference.

2 Binary dependent variables

- Interpreting coefficients are different in probit and logit models. For example, a one-unit increase in a regressor in a probit model does not correspond to a β increase in Y, but to a β increase in the z-score
- From Problem Set 6, we also know that the sign of a regressor at least tells you the sign of the corresponding effect on Y *except* if it enters non-linearly as, for example, with age and age-squared

3 Instrumental variables

- There are many estimators that use instrumental variables. The one that we focus on in this course is the two-stage least squares (2SLS or TSLS) because it has the advantage of being able to combine multiple instruments and control variables very easily.
- This is how I explained 2SLS in my office hours. If it is not easy to follow, feel free to forget it.

$$Y_i = \beta_0 + \beta_1 \underbrace{X_i}_{\boxed{\text{Exogenous} \mid \text{Endogenous}}} + e_i$$

First stage: $\hat{X}_i$ = linear function $f(Z_1, \dots, Z_k)$ of instruments $= \hat{\pi}_0 + \hat{\pi}_1 Z_{1i} + \dots + \hat{\pi}_k Z_{ki}$

Second stage:

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + e_i$$

1. The first equation gives the regression we'd like to estimate
2. Endogeneity of X_i (correlation with e_i) biases estimation of β_1
3. We can think of decomposing X_i into its exogenous components (blue) and its endogenous components (red)
4. Ideally, we'll have k instruments $Z_1, ..., Z_k$ that correlate with the exogenous component. This is the first stage: we want to capture as much of the exogenous variation as possible using a linear function of instruments: $\hat{X}_i = f(Z_1, ..., Z_k)$. The greater the proportion of the exogenous variation in X_i it explains (the wider is the blue line), the better will be the approximation $\hat{X}_i$ to the true variation in X_i and thus the greater our identifying variation. This is the relevance condition for a valid instrument.
5. If, however, these instruments capture a portion of the endogenous variation in X_i (the red part), then the proxy $\hat{X}_i$ will be a function of (an) enodogenous variable(s) and thus itself be endogenous. This violates the exogeneity condition for a valid instrument.
6. If the proportion of exogenous variation that the instruments capture is too small (i.e., they are not relevant enough), then we run into the problem of weak instruments. The textbook tells you what the implications are for inference.

- Given the requirements of relevance and exogeneity, it might be important to keep in mind how they affect the resulting estimators. Not just whether they become biased or inconsistent but whether they're biased or inconsistent in a particular direction. Something to look up.

- We have two tests which in turn have their own test statistics: the familiar F statistic to test weak instruments and the J statistic of overidentifying restrictions. We can test whether instruments are relevant but we can't test whether an instrument is exogenous. So then what does the J statistic tell us? Make sure you know precisely what the null hypotheses of these two tests are both in terms of words and in terms of coefficients (and which coefficients of what regression?!

- Here's the basic logic of the overidentifying restrictions test:
  - Suppose we have one endogenous variable and two possible instruments (just for simplicity)
  - We know how to use one instrument to create a "proxy" for an endogenous variable so that we arrive at a 2SLS estimator
  - Let us do this for both instruments to get an estimator for each
  - If the two resulting estimators are different enough from one another, they can't both be unbiased.

4 Experiments and quasi-experiments

- The difference-in-differences with repeated cross-section is given by the following equation

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 T_i + \beta_3 D_t + u_{it} \qquad (1)$$

where

  - T_i is an individual binary indicating whether observation i is assigned to the treatment group (very important: if an individual i is in the treatment group, they will still have a value of $T_i = 1$ even before the treatment period)
  - D_t is a time binary indicating whether the treatment has been administered (very important: if an individual i is in the control group and doesn't receive treatment, they will still have a value of $D_t = 1$ during the treatment period)
  - X_{it} is the interaction $T_i \times D_i$. It only equals one if individual i is in the treatment group *and* has received treatment

The coefficient β_1 is the desired difference-in-differences estimate. Why? First consider the treatment group ($T_i = 1$):

  - Post treatment: $\mathbb{E}[Y_{it}|T_i = 1, D_t = 1] = \beta_0 + \beta_1 + \beta_2 + \beta_3$
  - Pre treatment: $\mathbb{E}[Y_{it}|T_i - 1, D_t = 0] = \beta_0 + \beta_2$
  - Their difference: $\mathbb{E}[\Delta Y^{treatment}] = \beta_1 + \beta_3$

Then consider the control group ($T_i = 0$):

  - Post treatment: $\mathbb{E}[Y_{it}|T_i = 0, D_t = 1] = \beta_0 + \beta_3$
  - Pre treatment: $\mathbb{E}[Y_{it}|T_i - 0, D_t = 0] = \beta_0$
  - Their difference: $\mathbb{E}[\Delta Y^{control}] = \beta_3$

Then the differences in their differences:

$$\mathbb{E}[\Delta Y^{treatment}] - \mathbb{E}[\Delta Y^{control}] = \beta_1 + \beta_3 - \beta_3 = \beta_1 \qquad (2)$$

5 Big data

- So this is the chapter I ran out of time for (this guide took a long time to make!) but the thing I wanted to flag is that between regular regression, this big data section, and the time series section, we've come across three different notions of "prediction" that are often confused for one another. These predictions are evaluated by how they minimize the following:

  1. In-sample regression: the mean squared error
  2. Big data: the mean squared prediction error
  3. Time series: the mean squared forecast error

  What are the differences between these? What are their objectives? What are the implications for inference and interpretation? Why do they not lead to the same estimates? In all cases we produce some prediction $\hat{Y}$ to be evaluated against some true value Y: what data is used to produce $\hat{Y}$? Are they in-sample? Out of sample?

- You should be able to explain the conceptual differences between lasso, ridge, and principal component analysis in words. How do these methods affect inference/interpretability compared to regular unpenalized regressions? What are the benefits?

6 Time series and dynamic causal effects

- We can estimate dynamic multipliers by running regressions of the following form (for simplicity, we are here assuming just one independent variable of interest X):

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_{t-1} + ... + \beta_p X_{t-p} + u_t \qquad (3)$$

or we can estimate it through the differenced regression:

$$Y_t = \gamma_0 + \gamma_1 \Delta X_i + \gamma_2 \Delta X_{t-1} + ... + \gamma_p X_{t-p} + u_t \qquad (4)$$

- In the first regression, β_1 is the estimated contemporaneous effect of an increase in X by one unit. β_2 is the estimated effect of a one-unit increase in $X_{i,t-1}$. Since we cannot change the past value of X today, you can interpret this as the estimated effect of a one-unit increase in the previous period's level of X.

- Suppose we only experience a one-unit impulse shock in X at time t and otherwise X is zero in all other periods. If the model is correctly specified, this would imply this one-off shock increases Y_t by β_1, increases Y_{t+1} by β_2, ..., increases Y_{t+p} by β_p

- This motivates consideration of the cumulative effect of this one-off shock, captured by the cumulative multiplier. Two years after this one-off shock, the total effect is a β_1 increase in Y_t and a β_2 increase in Y_{t+1}. This results in a two-period cumulative multiplier of $\beta_1 + \beta_2$, the total impact on Y that the one-off shock has over two years. Over p periods, the cumulative effect is of course $\beta_1 + \beta_2 + ... + \beta_p$. According to our model, a one-off shock does not have a measurable impact after p periods so this sum represents the long-run cumulative multiplier of a one-off shock.

- These cumulative effects are immediately given by estimating the second regression by the following relations:

  - $\gamma_0 = \beta_0$
  - $\gamma_1 = \beta_1$
  - $\gamma_2 = \beta_1 + \beta_2$
  - $\gamma_3 = \beta_1 + \beta_2 + \beta_3$
  - ...
  - $\gamma_p = \sum_i^p \beta_i$

  So you can back out the coefficients of the first regression from the coefficients of the second regression and vice versa. However, only the second regression can give you the appropriate standard errors for cumulative multipliers and only the first regression can give you the appropriate standard errors for the dynamic multipliers

- Importantly: this second equation contains differenced regressors *except for the non-differenced term representing the pth lag!*. Not realizing this is a very common mistake.

7 General

- Read the whole question closely. Seriously, don't speed through reading because of the time constraint; you'll just end up spending more time re-reading it trying to find the one throwaway sentence that enables you to answer a particular question.

- Subquestions often ask you for multiple things. Two missteps arise when people lose points for getting the econometrics right but fail to notice the "Discuss" follow-up question or they lose time by answering questions that aren't asked just to be overly safe.

- If a question asks you to interpret a coefficient, write an interpretation in words describing the implied relationship between the relevant variables, including units for all. You can almost never go wrong with "a one-(unit? percent? percentage point? standard deviation?) increase in X is associated with a $\hat{\beta}$ (unit? percent? percentage point? standard deviation?) increase in Y"

- Cite one of the p-value, confidence interval, or test statistic when arguing for (non-)significance. Saying "it is (in)significant" is not enough.

- On units, the textbook tells you how to interpret a log-log, log-linear, linear-log, and linear-linear regression. What about the case in problem set 8 where we were regressing the first difference of log GDP against the first difference of log money supply? For example

$$\Delta \log Y_t = \beta_0 + \beta_1 \Delta \log M_t + u_t \tag{5}$$

  - First note that when you're taking differences in logs of some unit, the result is still in log units. So $\log(GDP_{2021}) - \log(GDP_{2020})$ is still in units of log GDP. Thus, we're still talking about log variables and thus we can interpret the coefficients in terms of percentages
  - Also note that the difference in logs is a growth rate. So you are regressing a percentage against a percentage and from our binary dependent variables chapter, we know we can thus interpret the coefficients in terms of percentage points (or we should have; lots of people made this mistake in problem set 6)
  - Both interpretations are permissible because both of the following are equivalent, just note the difference in units used:
    1. "A one percent increase in the money supply is associated with a $\hat{\beta}_1$ percent increase in GDP."
    2. "A one percentage point increase *in the growth rate* of the money supply is associated with a $\hat{\beta}_1$ percentage increase in *the growth rate* of GDP."

- Standard errors. We've encountered several different standard errors, keep in mind what problem they intend to address, i.e. what are they robust to? what models do they correspond to? what assumption is violated if we don't use them?

  1. heteroskedasticity-robust standard errors
  2. cluster-robust standard errors
  3. heteroskedasticity and autocorrelation-robust standard errors (aka Newey-West standard errors)

- Joint hypothesis testing

  - Testing the joint significance of regressors amounts to testing the following null hypothesis:

$$H_0 : \beta_1 = ... = \beta_k = 0 \tag{6}$$

  - The following is NOT:

$$H_0 : \beta_1 = ... = \beta_k \tag{7}$$

- Revise your logarithm rules

- At least two topics lend themselves to essay/understanding-based questions: discussion of whether an analysis has internal/external validity and discussion of whether a particular variable is exogenous or endogenous. These often require some creativity to apply it to a new context and thus a good understanding of what those concepts entail.

A graphical exploration of fixed effects panel models

Matthew Alampay Davis

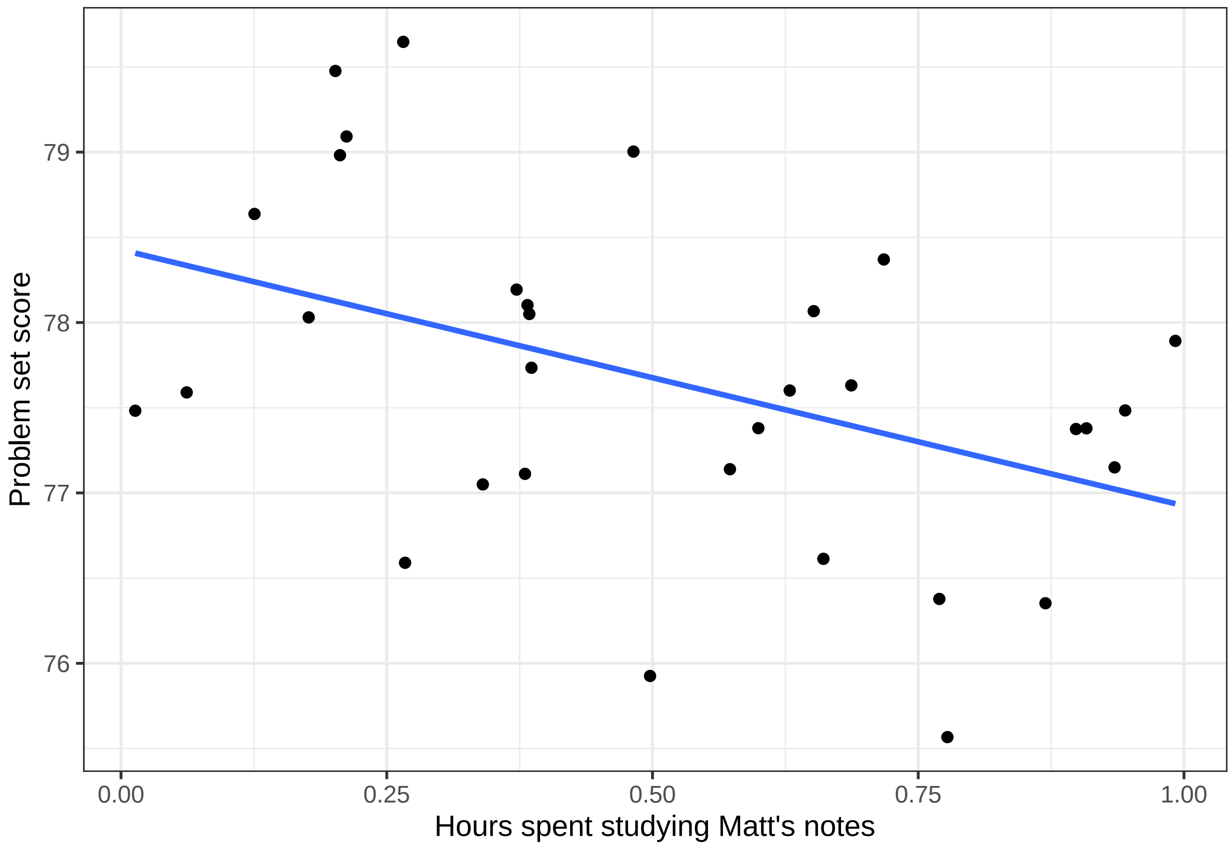
December 17, 2021

This note is just to demonstrate what fixed effects panel models do and how they differ from pooled OLS estimation. If you find it just adds to your confusion, then that's my fault and feel free to disregard. I've omitted most code but what is visible doesn't matter; you only need to follow the discussion and the graphs.

Suppose I have 32 datapoints. This (fake) dataset has two variables: $Notes_i$ is the number of hours student i spent studying my recitation notes and $Score_i$ is the score they received in the corresponding problem set:

```
ggplot(test, aes(x = study, y = score)) +
  theme_bw() +
  geom_smooth(method = 'lm', se = F) +
  ylab('Problem set score') + xlab("Hours spent studying Matt's notes") +
  geom_point()
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



This plot seems to suggest that studying my notes actually makes students perform worse. Indeed, when we run a basic regression, we get a negative effect significant at the 5% level:

```
lm(score ~ study, test) %>% summary
```

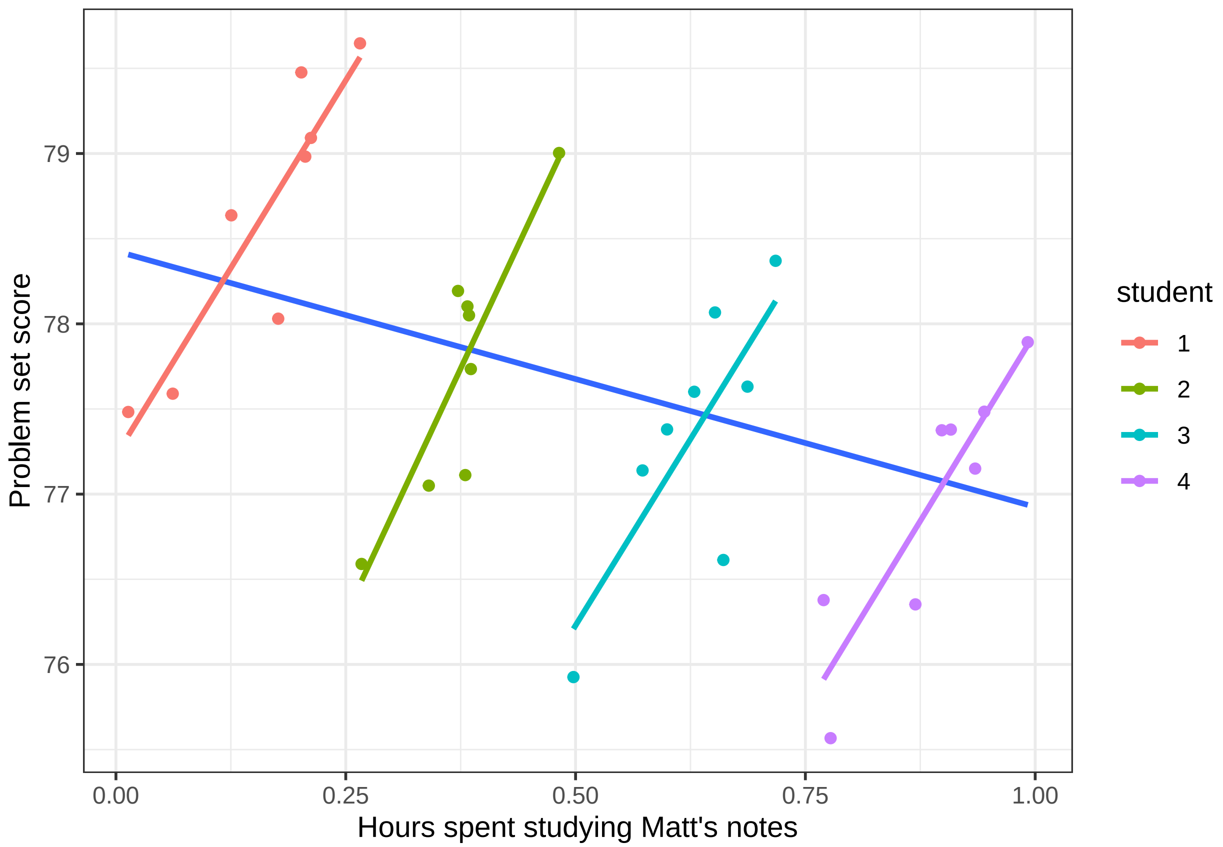
```
##
## Call:
## lm(formula = score ~ study, data = test)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7537 -0.7504  0.1636  0.5120  1.6187
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  78.4273     0.3283 238.867   <2e-16 ***
## study        -1.5029     0.5623  -2.673    0.012 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.8957 on 30 degrees of freedom
## Multiple R-squared:  0.1923, Adjusted R-squared:  0.1654
## F-statistic: 7.143 on 1 and 30 DF,  p-value: 0.01205
```

This regression suggests that for every additional hour a student spends studying my notes, their expected problem set score falls by 1.5 points. Obviously, this interpretation doesn't make sense given that I'm such a good TA. How did this happen?

It turns out that our data isn't cross-sectional after all. The 32 observations actually correspond to just four students and their performances on the eight problem sets. Thus we can say the dataset has a panel structure where we have $T = 8$ periods corresponding to the eight problem sets and $N = 4$ students as different entities. If we color the points by student, we get the following:

```
ggplot() +
  theme_bw() +
  geom_smooth(data = test, aes(x = study, y = score), method = 'lm', se = F) +
  ylab('Problem set score') + xlab("Hours spent studying Matt's notes") +
  geom_point(data = test, aes(x = study, y = score, color = student)) +
  geom_smooth(data = test, aes(x = study, y = score, color = student), se = F, method = 'lm')
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
## ‘geom_smooth()‘ using formula ’y ~ x’
```



Now the data is starting to make some sense: each student individually is seeing positive performance gains from reading my notes (as seen in the similarly positive sloped student-specific lines of best fit) but if we regarded all the points as independently drawn as before, we'd infer a negative effect (as seen in the blue line of best fit) because we'd be actually be running a pooled OLS regression of panel data. This gives misleading estimates if there are student-specific omitted factors that need to be accounted for. Fixed effects panel models account for these even without having exhaustive data on those potentially relevant variables. Formally, these individual-specific omitted factors are contained in the α_i term here:

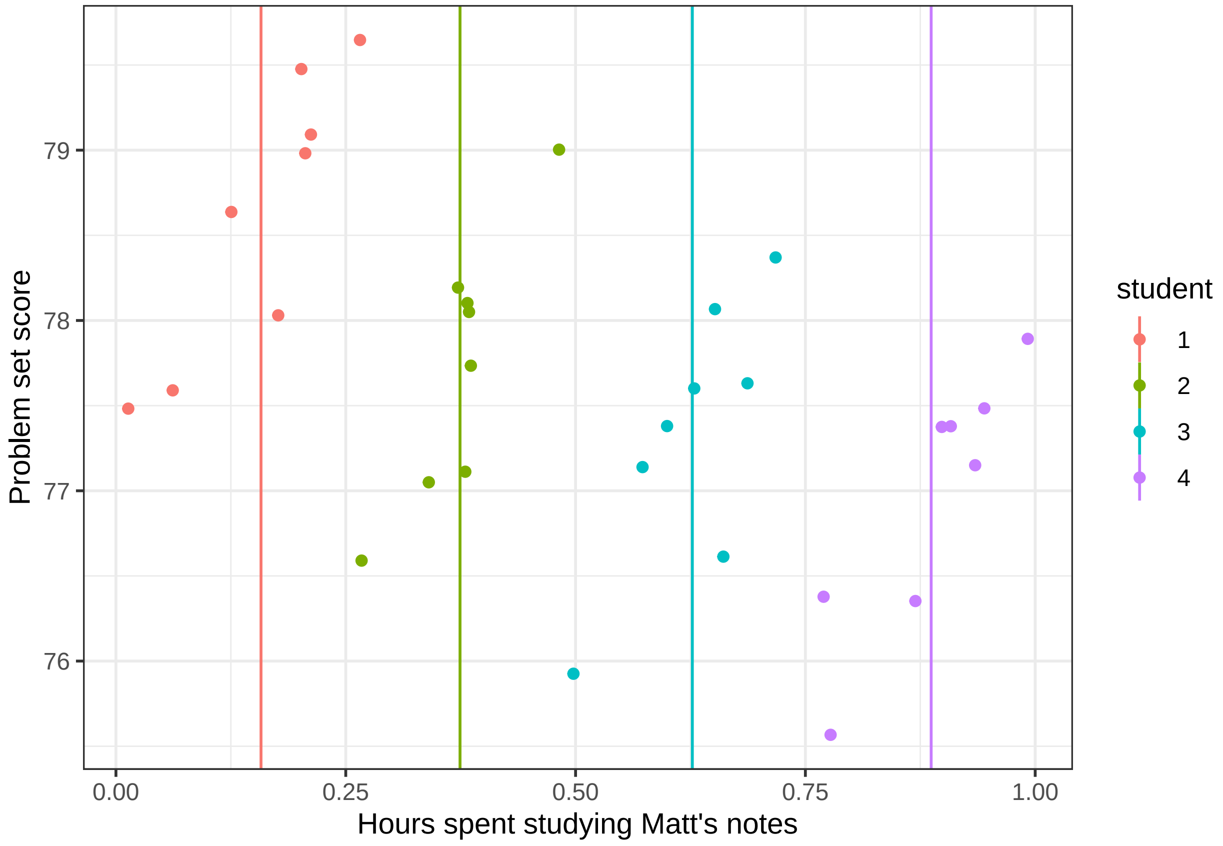
$$Score_{it} = \alpha_i + \beta Notes_{it} + u_{it}$$

For example, α_4 might absorb the negative effect of Student 4 having a particularly annoying roommate who prevents Student 4 from focusing on his work. Even if different students experience similar benefits to studying my excellent notes (meaning that reading my notes have the same per-hour positive effect β on their problem set score), systematic differences in student-level characteristics could give each linear relationship a different intercept α_i for each student i.

Fixed effects models allow us to account for these unobservable characteristics simply by entity demeaning. The lecture and textbook guides us through the algebraic justification for this, but basically the individual effects are differenced away in the demeaning process so that the identifying variation comes from deviations from entity means: this is why it's often called the "within" estimator. We can visualize what demeaning variables accomplishes graphically here for intuition.

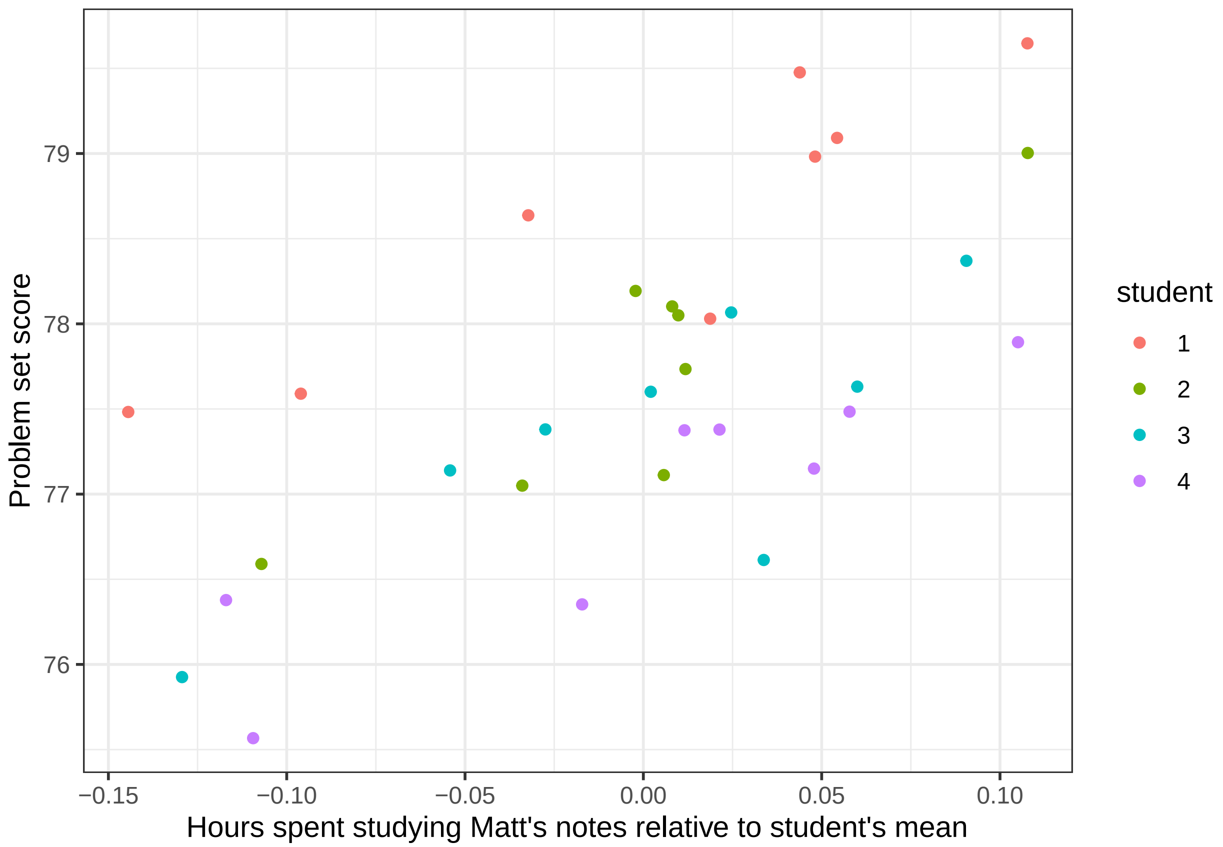
First, let us mark the mean time spent studying my $Notes$ for each student:

```
ggplot(test, aes(x = study, y = score, color = student)) +
  theme_bw() +
  ylab('Problem set score') + xlab("Hours spent studying Matt's notes") +
  geom_point() + geom_vline(aes(xintercept = mean.study, color = student))
```



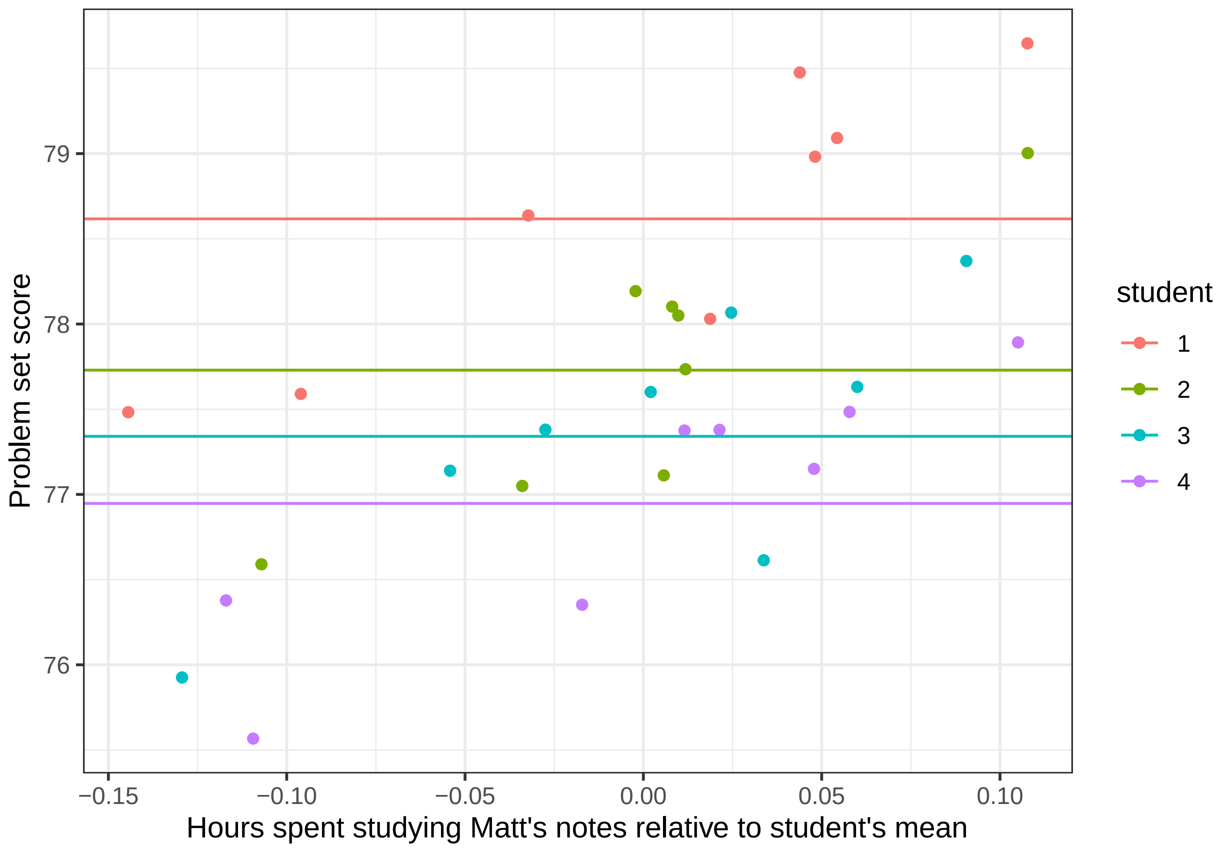
Then, for each observation, subtract their mean hours from their student-specific mean hours so that each student's hours are mean zero:

```
ggplot(test, aes(x = study-mean.study, y = score, color = student)) +
  theme_bw() +
  ylab('Problem set score') + xlab("Hours spent studying Matt's notes relative to student's mean") +
  geom_point()
```



Now for the dependent variable. Let's mark the mean problem set score for each student:

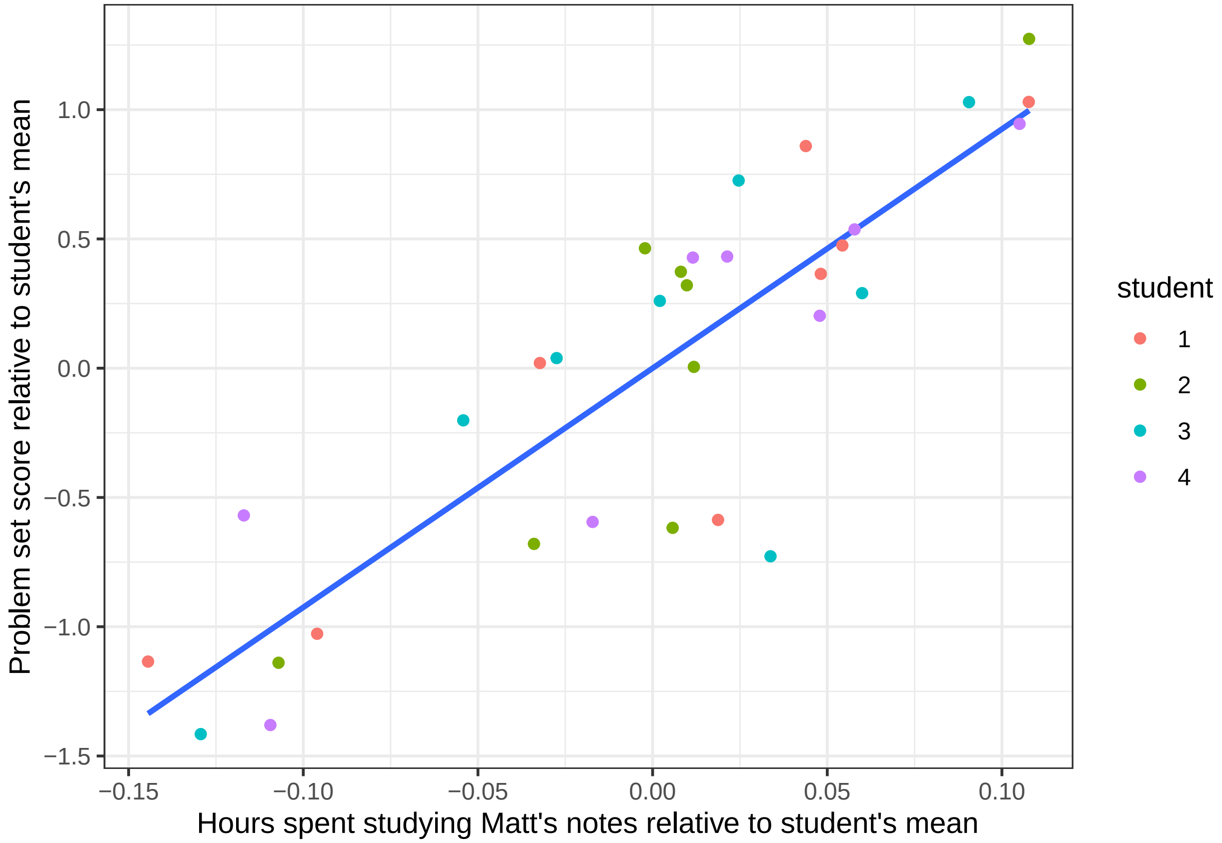
```
ggplot(test, aes(x = study-mean.study, y = score, color = student)) +
  theme_bw() +
  ylab('Problem set score') + xlab("Hours spent studying Matt's notes relative to student's mean") +
  geom_point() + geom_hline(aes(yintercept = mean.score, color = student))
```



Then as before, for each observation, subtract their student-specific mean score so the vertical axis is now relative to their student-specific means:

```
ggplot(test, aes(x = study-mean.study, y = score-mean.score)) +
  theme_bw() +
  ylab("Problem set score relative to student's mean") + xlab("Hours spent studying Matt's notes relati
  geom_smooth(method = 'lm', se = F) +
  geom_point(aes(color = student))
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



Aha. Now when we run the OLS regression on this demeaned data, we can get a sensical estimate of the effect of studying my notes.

```
test %<>% mutate(score.demeaned = score-mean.score,
                 study.demeaned = study-mean.study)
lm(score.demeaned ~ study.demeaned, data = test) %>% summary
```

```
##
## Call:
## lm(formula = score.demeaned ~ study.demeaned, data = test)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.03981 -0.22407  0.01793  0.29445  0.51302
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -8.304e-16  6.880e-02   0.000        1
## study.demeaned  9.250e+00  9.982e-01   9.267 2.61e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3892 on 30 degrees of freedom
## Multiple R-squared:  0.7411, Adjusted R-squared:  0.7325
## F-statistic: 85.87 on 1 and 30 DF,  p-value: 2.614e-10
```

We might also prefer to run this regression as a panel model so that we can also estimate the fixed effects themselves:

```
lm(score ~ study + factor(student), data = test) %>% summary
```

```
##
## Call:
## lm(formula = score ~ study + factor(student), data = test)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.03981 -0.22407  0.01793  0.29445  0.51302
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       77.1571     0.2205 349.929  < 2e-16 ***
## study              9.2503     1.0522   8.791 2.09e-09 ***
## factor(student)2  -2.8902     0.3065  -9.429 4.94e-10 ***
## factor(student)3  -5.6166     0.5346 -10.505 4.88e-11 ***
## factor(student)4  -8.4136     0.7941 -10.596 4.04e-11 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.4103 on 27 degrees of freedom
## Multiple R-squared:  0.8475, Adjusted R-squared:  0.8249
## F-statistic: 37.51 on 4 and 27 DF,  p-value: 1.173e-10
```

The coefficients in the two regressions agree: for every hour spent studying my notes, a student’s expected problem set score increases by 9.25 points and that this estimate is very significant. Note however that the first regression has standard errors that are a bit too small because when we manually demean the variables ourselves, we are exploiting information given by the student variable: effectively, we are manually using the student’s identity as a control variable.

Also note the coefficients on the fixed effects. Student 1, the highest performing student, is the omitted case (who has an expected problem set score of 77.2). The fixed effect estimates give the relative expected problem set scores of the other students with students 2-4 decreasing in expected problem set score, corresponding to systematic differences in implied intercepts. The p-values indicate these intercepts are significant, meaning there is strong evidence to reject the null hypothesis that the other students are expected to receive the same problem set score as Student 1 for the same level of studying.

Note that all the above analysis describes entity fixed effects. We could perform a similar exercise to capture time fixed effects μ_t if we suspect the presence of time-specific (i.e., problem-set specific) characteristics of each observation. For example, a particularly difficult problem set would cause trouble for all students relative to the average problem set while still increasing with exposure to my notes. And of course we could include both entity and time fixed effects together to account for both sources of systematic difference.

Hopefully this was helpful but if not, feel free to disregard. The main point is that if anyone does poorly on the exam, even if it might seem like it’s my fault it’s really not. Best of luck!