ECON 3412 Final

David Chen, dc3451

December 20, 2020

A slight disclaimer before I start: I'm using `R`, which in this second half of the course has differing implementations of some of the features that we've discussed in class. For example, obviously simple OLS lines up, but sometimes the numerical computations (such as MLE iterations) diverge slightly from the STATA values (this divergence also happens to a slight degree with Newey-West estimation, if I recall correctly). I've tried to align `R` and STATA as much as possible, but sometimes even when inputing the correct command, the results are off by a few percent.

Q1

a

Since we have detailed data from *individuals*, we expect that by running a panel regression with entity-fixed-effects should allow us to remove the differing abilities from the question, since we get that this sort of ability ought be internal to the person and shouldn't change drastically. In particular, entity-fixed-effects remove effects that vary accross time but are constant to the person, so the effect of ability should be removed by it and allow for an unbiased estimation of what we want.

b

In particular, this type of fixed regression doesn't tell you the actual affect of ability (or race/gender, which are *more* internal to a person than ability), simply the other regressors. Investigating discrimination would require a different approach: this only gets you the effects of education and experience, with fixed effects (like ability/race/gender) mostly removed: it doesn't give the magnitude or the direction of those removed fixed effects. If we just added race/gender dummies, we expect to see their coefficients vanish in the fixed-effects regression.

c

Time dummies for each period would be the same is time-demeaning in the regression. In particular, this would control for any specific time related phenomenons which vary accross time but not accross each individual, such as if there were a recession at some point in the 5 years which could throw off the data.

Q2

a

```
> apple <- read_dta("./APPLE.dta")
> apple$ecobuy <- apple$ecolbs > 0
> nrow(apple[apple$ecobuy == 1,])/nrow(apple)

[1] 0.6242424
```

so 62% of families in the data buy any ecologically friendly apples.

b

```
> apple.lpm <- lm(ecobuy ~ ecoprc + regprc + faminc + hhsize + educ + age, data = apple)
> coeftest(apple.lpm, vcovCL)

t test of coefficients:

               Estimate  Std. Error t value  Pr(>|t|)
(Intercept)  0.42368654  0.16775289  2.5257  0.011784 *
ecoprc      -0.80262186  0.10566783 -7.5957 1.064e-13 ***
regprc       0.71926754  0.13023174  5.5230 4.816e-08 ***
faminc       0.00055180  0.00052447  1.0521  0.293140
hhsize       0.02382271  0.01246720  1.9108  0.056464 .
educ         0.02478486  0.00845652  2.9309  0.003498 **
age         -0.00050079  0.00126552 -0.3957  0.692442
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

For the two price variables, we have that holding the price of regular apples constant (and the other regressors constant as well), we get that an unit increase in the price of environmentally friendly apples is associated on average with a 80% decrease in likelihood of buying those apples (to be clear, the probability decreases by 0.8, not that it shrinks to a fifth of its original value). Similarly, we have that holding the price of eco-friendly apples constant (and the other regressors constant as well), we get that an unit increase in the price of regular apples is associated on average with a 72% increase in likelihood of buying eco-friendly apples (same as before, the probability increases by 0.72).

Note these prices are only $\approx$ a dollar anyway, so a unit increase in price is *massive*.

c

Testing if nonprice variables are jointly significant:

```
> linearHypothesis(apple.lpm, c("faminc = 0", "hhsize = 0", "educ = 0", "age = 0"), vcov = vcov
Linear hypothesis test

Hypothesis:
faminc = 0
hhsize = 0
educ = 0
age = 0

Model 1: restricted model
Model 2: ecobuy ~ ecoprc + regprc + faminc + hhsize + educ + age

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F   Pr(>F)
1    657
2    653  4 4.2427 0.002133 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

so we see that the nonprice variables are jointly significant at $p = 0.002$.

It seems that age is extremely unimportant, as it is not significantly different from zero, and has a very small coefficient as well; same for family income. However, household size is more important, being significant at $p = 0.056$ and influencing decision making more, with a coefficeint of 0.022. That being said, the max householdsize is ¡ 10, so this effect is still small compared to price effects.

The most salient of the nonprice variables is education, significant at $p = 0.0069$, and with a coefficient of 0.023, so every unit increase of education leads to a 2.3% percentage point increase in the probability of buying eco-friendly apples.

These results are a little suprising: we might expect age to be more important (young people probably care more about the environment) as well as family income (since eco-firendly apples are likely more expensive) but not that suprising: middle aged families (insert white semi-liberal suburbia) probably also (at least superfically) care about the environment and apples are cheap either way, so its not that important. Education is expected, since we expect more educated people to care more about the environment. I had no real prior re: household size, but its not unsuprising: larger household = more kids = desire for parents to buy healthier food (not directly eco-firendly apples, but related in marketing and perception).

d

```
> apple.lpm <- lm(ecobuy ~ ecoprc + regprc + logfaminc + hhsize + educ + age, data = apple)
> coeftest(apple.lpm, vcovCL)

t test of coefficients:

               Estimate  Std. Error t value  Pr(>|t|)
(Intercept)  0.30375189  0.18178854  1.6709  0.095219 .
ecoprc      -0.80066642  0.10554061 -7.5863 1.137e-13 ***
regprc       0.72137704  0.12989247  5.5536 4.074e-08 ***
logfaminc    0.04451623  0.02927923  1.5204  0.128894
hhsize       0.02270015  0.01249885  1.8162  0.069801 .
educ         0.02309299  0.00852342  2.7094  0.006918 **
age         -0.00038652  0.00126450 -0.3057  0.759952
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

so we have that $logfaminc$ is significant at $p = 0.128$ whereas before, $faminc$ was significant at $p = 0.30$. They're still not significantly different from 0 under conventional standards of $p < 0.05$, but $logfaminc$ does fit the data better.

For every percent increase in family income is an associated increase of 4.5% in the average likelihood of buying ecologically friendly apples (that is, as opposed to the earlier part, here I mean that the probability rises by 4.5% of its old value before the unit change in $logfaminc$).

e

```
> apple[predict(apple.lpm) < 0,]
# A tibble: 0 x 19
# ... with 19 variables: id <dbl>, educ <dbl>, date <chr>, state <chr>,
#   regprc <dbl>, ecoprc <dbl>, inseason <dbl>, hhsize <dbl>, male <dbl>,
#   faminc <dbl>, age <dbl>, reglbs <dbl>, ecolbs <dbl>, numlt5 <dbl>,
#   num5_17 <dbl>, num18_64 <dbl>, numgt64 <dbl>, ecobuy <lgl>, logfaminc <dbl>
> apple[predict(apple.lpm) > 1,]
# A tibble: 2 x 19
     id  educ date  state regprc ecoprc inseason hhsize  male faminc   age
  <dbl> <dbl> <chr> <chr>  <dbl>  <dbl>    <dbl>  <dbl> <dbl>  <dbl> <dbl>
```

```
1 10805    20 1122... VA     0.590  0.590        1      5     0    175    30
2 12592    20 11498 LA     0.890  0.890        0      7     1    105    31
# ... with 8 more variables: reglbs <dbl>, ecolbs <dbl>, numlt5 <dbl>,
#   num5_17 <dbl>, num18_64 <dbl>, numgt64 <dbl>, ecobuy <lgl>, logfaminc <dbl>
```

Suprisingly, we only get that 2 entries have probability > 1 in the model, and no entires have probability < 0 in the model. This is good for a linear probability model, and suggests that probit/logit may be unnecesarry.

f

```
> nrow(apple[(predict(apple.lpm) >= 0.5) & apple$ecobuy,,])
[1] 340
> nrow(apple[apple$ecobuy,])
[1] 412
> nrow(apple[!apple$ecobuy,])
[1] 248
> nrow(apple[(predict(apple.lpm) <= 0.5) & !apple$ecobuy,])
[1] 102
```

so 340/412 of families who do buy ecologically friendly apples are predicted correctly, and 102/248 of families who do not buy ecologically friendly apples are predicted correctly, so we the outcome $ecobuy = 1$ is predicted more accurately by the model.

Q3

a

```
> htv.ols <- lm(lwage ~ educ, data = htv)
> coeftest(htv.ols, vcovCL)

t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept) 1.0923192  0.0991578  11.016 < 2.2e-16 ***
educ        0.1013613  0.0074982  13.518 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 0.1013613 + 1.96 * 0.0074982
[1] 0.1160578
> 0.1013613 - 1.96 * 0.0074982
[1] 0.08666483
```

so we get the confidence interval as $(0.08666483, 0.1160578)$.

b

```
> coeftest(lm(educ ~ ctuit, data = htv), vcovCL)

t test of coefficients:

             Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.038360   0.067147 194.177   <2e-16 ***
ctuit       -0.049447   0.079497  -0.622   0.5341
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

so the coefficient of *ctuit* is not significantly different from 0 in this regression, so we cannot reject the null that they have no bearing on each other. (and note that sample correlation is ≈ 0 at `cor(htv$educ, htv$ctuit) = -0.01689161`)

In particular, we need instrucment relevance, i.e. that $\text{cor}(educ, ctuit) \neq 0$, so *ctuit* would be a bad instrumental variable.

c

```
> htv.olsc <- lm(lwage ~ educ + exper + I(exper^2) + ne + nc + west +
+                   + ne18 + nc18 + west18 + urban + urban18,
+                 data = htv)
> coeftest(htv.olsc, vcovCL)

t test of coefficients:

              Estimate Std. Error t value  Pr(>|t|)
(Intercept) -0.5074853  0.2503133 -2.0274  0.042839 *
educ         0.1371483  0.0100981 13.5816 < 2.2e-16 ***
exper        0.1123554  0.0257243  4.3677 1.362e-05 ***
I(exper^2)  -0.0030404  0.0011594 -2.6225  0.008838 **
ne          -0.0168028  0.0910685 -0.1845  0.853647
nc          -0.0174304  0.0713778 -0.2442  0.807117
west         0.0175429  0.0926163  0.1894  0.849799
ne18         0.1563607  0.0921372  1.6970  0.089944 .
nc18         0.0113699  0.0744449  0.1527  0.878638
west18      -0.0295760  0.0996236 -0.2969  0.766611
urban        0.2046369  0.0414881  4.9324 9.246e-07 ***
urban18      0.1260197  0.0528489  2.3845  0.017253 *
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(note that south and south18 are not in the regression because of the dummy variable trap) so we get that an extra year of education is associated with, on average, an increase of 13% in wages.

d

```
> htv.rfreg <- lm(educ ~ ctuit + exper + I(exper^2) + ne + nc + west +
+                   + ne18 + nc18 + west18 + urban + urban18,
+                 data = htv)
+ coeftest(htv.rfreg, vcovCL)
+
t test of coefficients:

              Estimate Std. Error  t value  Pr(>|t|)    
(Intercept) 21.2424886  0.4142453  51.2800 < 2.2e-16 ***
ctuit       -0.1652227  0.0749731  -2.2038   0.02773 *  
exper       -0.8738310  0.0735128 -11.8868 < 2.2e-16 ***
I(exper^2)   0.0156522  0.0034919   4.4824 8.076e-06 ***
ne          -0.3745971  0.3553196  -1.0543   0.29198    
nc          -0.1415143  0.2752213  -0.5142   0.60722    
west         0.6220028  0.2901246   2.1439   0.03224 *  
ne18         0.6533440  0.3490274   1.8719   0.06146 .  
nc18         0.2322316  0.2773732   0.8373   0.40261    
west18      -0.4480626  0.3078447  -1.4555   0.14579    
urban       -0.0769929  0.1151254  -0.6688   0.50377    
urban18     -0.9885221  0.1254737  -7.8783 7.308e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

so now *ctuit* is significant at $p = 0.027$, as desired.

e

```
> htv.ivregreal <- ivreg(lwage ~ educ + exper + I(exper^2) + ne + nc + west +
+                          + ne18 + nc18 + west18 + urban + urban18 | ctuit + exper +
+                          I(exper^2) + ne + nc + west +
+                          + ne18 + nc18 + west18 + urban + urban18,
+                        data = htv)
> coeftest(htv.ivregreal, vcovCL)
```

```
t test of coefficients:

              Estimate Std. Error t value  Pr(>|t|)
(Intercept) -2.8942321  2.4691598 -1.1722   0.24137
educ         0.2500082  0.1169355  2.1380   0.03272 *
exper        0.2094245  0.1024511  2.0441   0.04115 *
I(exper^2)  -0.0047509  0.0021068 -2.2550   0.02431 *
ne           0.0289029  0.1183882  0.2441   0.80717
nc           0.0028895  0.0874150  0.0331   0.97364
west        -0.0543256  0.1229782 -0.4417   0.65875
ne18         0.0760555  0.1389796  0.5472   0.58431
nc18        -0.0209026  0.0948402 -0.2204   0.82560
west18       0.0234556  0.1178606  0.1990   0.84229
urban        0.2146702  0.0453058  4.7383 2.409e-06 ***
urban18      0.2373826  0.1265110  1.8764   0.06084 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 0.25 + 1.96 * 0.1169355
[1] 0.4791936
> 0.25 - 1.96 * 0.1169355
[1] 0.02080642
```

so the 95% CI for *educ* is $(0.0208, 0.479)$, so we get that the confidence interval in part a is a subset of the CI in this part, which has a much larger standard error, as well as predicting the average affect of educ (25% in wages per unit change in educ) to be much higher than before.

f

I think that it is relatively convincing, once the controls are added. After the introduction of controls, we see that there is a significant relationship between *educ* and *ctuit*, so we get instrument relavance, and I expect that the correlation between *ctuit* which seems to be semi-random flucuations (mostly 0) is plausibly 0, so we get exogenity as well (most of *ctuit* is likely determined by factors internal to colleges themselves, and as a result isn't really reflected in general wages.)

There is also a reasonable case that we don't have instrument exogenity: suppose we have some deep recession and stagflation, which winds up deflating wages, such that the model overestimates the relationship between educ and lwage. Then, we might also expect that ctuit might increase as well because of this recession since prices are increasing; in this case, we have nonzero correlation between the insturment and the error of the model.

Q4

a

```
> mean(supas[supas$young & supas$high,]$yeduc)
[1] 8.914207
> mean(supas[supas$young & !supas$high,]$yeduc)
[1] 10.11892
> mean(supas[supas$old & supas$high,]$yeduc)
[1] 8.539234
> mean(supas[supas$old & !supas$high,]$yeduc)
[1] 9.861114
```

so the table (dropping the labels, same as on final) is

8.914	10.1189	-1.2049
8.539	9.8611	-1.3221
0.375	0.2578	0.1172

so we get that $(educ_{high,young} - educ_{high,old}) - (educ_{low,young} - educ_{low,old}) = 0.1172$, so the interpretation is that the difference between and after the school building policy in years of education is higher by 0.1172 compared to those in areas where the policy did not take place; thus, this predicts that the policy increased education years by 0.1172 on average than the counterfactual where no new schools were built.

b

```
> supas$treated <- supas$young * supas$high
+ filter <- supas[supas$young | supas$old,]
+ filter.did <- lm(yeduc ~ high + young + treated, data = filter)
+ coeftest(filter.did, vcovCL)
+
> >
t test of coefficients:

             Estimate Std. Error  t value  Pr(>|t|)
(Intercept)  9.861114   0.042117 234.1359 < 2.2e-16 ***
high        -1.321880   0.068617 -19.2646 < 2.2e-16 ***
youngTRUE    0.257810   0.055627   4.6346  3.59e-06 ***
treated      0.117164   0.089352   1.3113    0.1898
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We need the assumptions that the treatment and the sample are both representative and effectively random, i.e. that the correlation between *yeduc* and high as well as *yeduc* and *young* and *yeduc* and *treated* are all 0. The difference-in-difference estimator is not significant at $p = 0.1898$.

c

```
filter.fe <- plm(yeduc ~ treated, data = supas, index = c("ROB", "YOB"), model = "within", effe
```

doesn't correctly handle the fixed-effects, since the indexes are nonunique. I could manually fix this by writing a macro to generate all the dummies, so I did.

```
> coef_test(filter.fe, vcov = "CR1S", cluster = filter$ROB)

             Coef. Estimate     SE    t-stat d.f. p-val (Satt) Sig.
1      (Intercept) 10.83665 0.0792  136.7769  192      < 0.001  ***
2   filter$treated  0.17694 0.1177    1.5028  183      0.13461
3      rob3516TRUE -2.74260 0.0582  -47.0988  183      < 0.001  ***
4      rob3219TRUE -3.64662 0.0688  -52.9701  167      < 0.001  ***
5      rob3578TRUE  0.72416 0.0572   12.6709  183      < 0.001  ***
6      rob5309TRUE -2.26996 0.0578  -39.2475  184      < 0.001  ***
7      rob1671TRUE  0.49790 0.0533    9.3341  184      < 0.001  ***
8      rob3518TRUE -1.88666 0.0555  -33.9641  184      < 0.001  ***
9      rob1275TRUE  0.60917 0.0555   10.9741  184      < 0.001  ***
10     rob3172TRUE  0.21007 0.0647    3.2490  174      0.00139   **
...
```

which is truncated. This increases the magnitude of the effect of being treated from earlier, the dif-in-dif regression.

Fixed effects regression here controls for things like large natrual disasters which stop education uniformly accross entities but not aacross time (this is time-fixed-effects) or for things like differentt cultural/material demands in regions which are constant accross time (for example a rural region is less likely to see large amounts of education) in entity-fixed effects.

Q5

a

The unfilled portion is as follows:

0.3	2.34
9.2	-5.05
-4.2	-35.95

b

```
> growth.ar2 <- lm(growth ~ growthlag1 + growthlag2, data = growth)
+ coeftest(growth.ar2, vcovCL)
+
>
t test of coefficients:

             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.146417   0.838486  2.5599   0.0124 *
growthlag1  -0.363991   0.361718 -1.0063   0.3174
growthlag2   0.071781   0.241908  0.2967   0.7675
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

c

```
> growth.adl11 <- lm(growth ~ growthlag1 + unratelag1, data = growth)
> coeftest(growth.adl11, vcovCL)

t test of coefficients:

            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.84792    3.61652 -0.5110   0.6108
growthlag1  -0.24333    0.34701 -0.7012   0.4852
unratelag1   0.67800    0.48190  1.4069   0.1634
```

d

(I assume that there is no intermediate rounding? Unclear what it means to use 2 decimal places. This one is without.)

The prediction is $(-0.2433 * 29.6139717 + 0.678 * 8.8) = -1.24$ for Q4 2020's growth rate.

With intermediate rounding, $(-0.24 * 29.61 + 0.67 * 8.8) = -1.21$ for Q4 2020's growth rate.

Q6

a

```
> oo.dyn <- lm(pc_price ~ gdd + gddlag1 + gddlag2 + gddlag3 + gddlag4, data = oo)
> coeftest(oo.dyn, vcov = NeweyWest(oo.dyn, lag = 4))
```

```
t test of coefficients:

            Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.79298   27.75026 -0.2808   0.7795
gdd          0.34052    0.28408  1.1987   0.2340
gddlag1     -0.53489    0.37243 -1.4362   0.1546
gddlag2      0.13198    0.18737  0.7044   0.4831
gddlag3      0.37086    0.24010  1.5446   0.1261
gddlag4     -0.21974    0.22181 -0.9906   0.3246
```

so I pick the rule of thumb for Newey-West, which is $0.75 * 96^{1/3} = 3.4$ rounded up to 4.

b

Impact effect of change is an increase (on average) of $10 * 0.34052 = 3.405$ in *pc_price*, such that the percent change in prices goes up by 3.405 percentage points.

c

```
> oo$delta_0 <- oo$gdd - oo$gddlag1
+ oo$delta_1 <- oo$gddlag1 - oo$gddlag2
+ oo$delta_2 <- oo$gddlag2 - oo$gddlag3
+ oo$delta_3 <- oo$gddlag3 - oo$gddlag4
> oo.cumdyn <- lm(pc_price ~ delta_0 + delta_1 + delta_2 + delta_3 + gddlag4, data = oo)
+ coeftest(oo.cumdyn, vcov = NeweyWest(oo.cumdyn, lag = 4))
+
>
t test of coefficients:

            Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.79298   27.75026 -0.2808  0.77952
delta_0      0.34052    0.28408  1.1987  0.23395
delta_1     -0.19437    0.20100 -0.9670  0.33624
delta_2     -0.06239    0.11835 -0.5272  0.59943
delta_3      0.30846    0.15407  2.0021  0.04843 *
gddlag4      0.08873    0.19840  0.4472  0.65584
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

d

This is the coefficient on δ_2, since we want the impact effect and the effect over the next two months, so we have that this is -0.0623, so an unit change in *gdd* will over the next two months associate on average with a decrease in *pc_price* of 0.0623 percentage points.

e

Strictly exogenous means that $E(u_t \mid \dots, X_{t+1}, X_t, X_{t-1}, \dots) = 0$, i.e. that the future and the past *gdd* is uncorrelated with the error of the model, whereas normal exogenity is just $E(u_t \mid X_t, X_{t-1}, \dots) = 0$, in which case we only care about the past *gdd*.

This data is potentially not strictly exogenous, since we might have something like olive farmers are adapt to grow in better places as time moves on, so if we see that olive production is slowing due to climate changes (and thus changes *gdd*; for example, climate change may cause a steady rise in *gdd*) compared to the past (and the model is thus overpredicting), farmers may move elsewhere, meaning that present model error and future *gdd* can be correlated.