Homework 4

Chunyu Qu

Oct 18, 2019

Chapter 4 Problems: 4.11, 4.12, 4.13, 4.14.

```
# install.packages('plm')
# install.packages("clubSandwich")
# install.packages('car')
library(plm)
library(car)
```

```
## Loading required package: carData
```

4.11.

a. In Example 4.3, use KWW and IQ simultaneously as proxies for ability in equation (4.29).

$$log(wage) = \beta_0 + \beta_1 exper + \beta_2 tenure + \beta_3 married + \beta_4 south + \beta_5 urban + \beta_6 black + \beta_7 educ + \gamma abil + v$$

Compare the estimated return to education without a proxy for ability and with IQ as the only proxy for ability.

Answer
First we compute the original model without proxy

```
NLS80 = read.csv("D:/Google Drive/Fordham/2019 Spring/AE/TA/PS4/NLS80.csv", header = TRUE)
lm1 = lm(log(wage) ~ exper + tenure + married + south + urban + black + educ , data=NLS80)
summary(lm1)
```

```
##
## Call:
## lm(formula = log(wage) ~ exper + tenure + married + south + urban +
##     black + educ, data = NLS80)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.98069 -0.21996  0.00707  0.24288  1.22822
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.395497   0.113225  47.653  < 2e-16 ***
## exper        0.014043   0.003185   4.409 1.16e-05 ***
## tenure       0.011747   0.002453   4.789 1.95e-06 ***
```

```
## married      0.199417   0.039050   5.107 3.98e-07 ***
## south       -0.090904   0.026249  -3.463 0.000558 ***
## urban        0.183912   0.026958   6.822 1.62e-11 ***
## black       -0.188350   0.037667  -5.000 6.84e-07 ***
## educ         0.065431   0.006250  10.468  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3655 on 927 degrees of freedom
## Multiple R-squared:  0.2526, Adjusted R-squared:  0.2469
## F-statistic: 44.75 on 7 and 927 DF,  p-value: < 2.2e-16
```

Then we include the iq

```
lm2 = lm(log(wage) ~ exper + tenure + married + south + urban + black + educ + iq , data=NLS80)
summary(lm2)
```

```
##
## Call:
## lm(formula = log(wage) ~ exper + tenure + married + south + urban +
##     black + educ + iq, data = NLS80)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.01203 -0.22244  0.01017  0.22951  1.27478
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.1764391  0.1280006  40.441  < 2e-16 ***
## exper        0.0141459  0.0031651   4.469 8.82e-06 ***
## tenure       0.0113951  0.0024394   4.671 3.44e-06 ***
## married      0.1997644  0.0388025   5.148 3.21e-07 ***
## south       -0.0801695  0.0262529  -3.054 0.002325 **
## urban        0.1819463  0.0267929   6.791 1.99e-11 ***
## black       -0.1431253  0.0394925  -3.624 0.000306 ***
## educ         0.0544106  0.0069285   7.853 1.12e-14 ***
## iq           0.0035591  0.0009918   3.589 0.000350 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3632 on 926 degrees of freedom
## Multiple R-squared:  0.2628, Adjusted R-squared:  0.2564
## F-statistic: 41.27 on 8 and 926 DF,  p-value: < 2.2e-16
```

Moreover, we include both iq and kww

```
lm3 = lm(log(wage) ~ exper + tenure + married + south + urban + black + educ + iq + kww, data=NLS80)
summary(lm3)
```

```
##
## Call:
## lm(formula = log(wage) ~ exper + tenure + married + south + urban +
```

```
##     black + educ + iq + kww, data = NLS80)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.05704 -0.21621  0.00824  0.23725  1.24895
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.175644   0.127776  40.506  < 2e-16 ***
## exper        0.012752   0.003231   3.947 8.51e-05 ***
## tenure       0.010925   0.002446   4.467 8.92e-06 ***
## married      0.192145   0.038909   4.938 9.35e-07 ***
## south       -0.082029   0.026222  -3.128  0.00181 **
## urban        0.175823   0.026910   6.534 1.06e-10 ***
## black       -0.130399   0.039901  -3.268  0.00112 **
## educ         0.049837   0.007262   6.863 1.24e-11 ***
## iq           0.003118   0.001013   3.079  0.00214 **
## kww          0.003826   0.001852   2.066  0.03913 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3625 on 925 degrees of freedom
## Multiple R-squared:  0.2662, Adjusted R-squared:  0.2591
## F-statistic: 37.28 on 9 and 925 DF,  p-value: < 2.2e-16
```

For no proxy the estimated return was about 6.54%, with only IQ as a proxy it was about 5.44%, using both IQ and KWW as proxies for ability is 4.98%. Thus, we have an even lower estimated return to education, but it is still practically nontrivial and statistically very significant

b. Test KWW and IQ for joint significance in the estimated equation from part a.

Answer

```
# For Homoskedasticity
linearHypothesis(lm3, c("iq=0", "kww=0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## iq = 0
## kww = 0
##
## Model 1: restricted model
## Model 2: log(wage) ~ exper + tenure + married + south + urban + black +
##     educ + iq + kww
##
##   Res.Df    RSS Df Sum of Sq     F    Pr(>F)
## 1    927 123.82
## 2    925 121.56  2     2.259 8.595 0.0002002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# For Heteroskedasticity, we adopt HTSKD-robust LM test
```

```r
# (1) Obtain u-tilda from testricted model
lm_restrict = lm(log(wage) ~ exper + tenure + married + south + urban + black + educ, data=NLS80)
u_tilda=resid(lm_restrict)

# (2) Regress each independent v's excluded on all of the included variables and obtain a set of residu
lm_rd1= lm(iq ~ exper + tenure + married + south + urban + black + educ, data=NLS80)
r1_tilda=resid(lm_rd1)
lm_rd2= lm(kww ~ exper + tenure + married + south + urban + black + educ, data=NLS80)
r2_tilda=resid(lm_rd2)

# (3)Find the product between each r_tilda and u_tilda
ru1=r1_tilda*u_tilda
ru2=r2_tilda*u_tilda

# (4) Regress 1 on ru's w.o. intercept
lm_LM= lm (rep(1,length(ru1))~ru1+ru2-1)
LM_test=length(ru1)-sum(resid(lm_LM)^2)

# (5) Which follows a chi-sq(2) test
c("Reject?",LM_test>qchisq(.975, df=2))
```

```
## [1] "Reject?" "TRUE"
```

The F test verifies this, with p-value .0002

c.

Answer
When KWW and IQ are used as proxies for abil, does the wage differential between nonblacks and blacks disappear? What is the estimated differential?
The wage differential between nonblacks and blacks does not disappear. Blacks are estimated to earn about 13% less than nonblacks, holding other factors in the regression fixed.

d. Add the interactions educ(IQ-100) and educ(KWW-$K\bar{W}W$ 100) KWW? to the regression from part a, where KWW is the average score in the sample. Are these terms jointly significant using a standard F test? Does adding them affect any important conclusions?

Answer

```r
educiq0 = NLS80$educ * (NLS80$iq - 100)
educkww0 = NLS80$educ * (NLS80$kww - mean(NLS80$kww))
lm4 = lm(lwage ~ exper + tenure + married + south + urban + black + educ + iq + kww + educiq0 + educkww
summary(lm4)
```

```
##
## Call:
## lm(formula = lwage ~ exper + tenure + married + south + urban +
##     black + educ + iq + kww + educiq0 + educkww0, data = NLS80)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.03083 -0.21490  0.00886  0.23928  1.27862
##
## Coefficients:
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.0800052  0.5610875  10.836  < 2e-16 ***
## exper        0.0121544  0.0032358   3.756 0.000183 ***
## tenure       0.0107206  0.0024383   4.397 1.23e-05 ***
## married      0.1978269  0.0388272   5.095 4.23e-07 ***
## south       -0.0807609  0.0261374  -3.090 0.002063 **
## urban        0.1784310  0.0268710   6.640 5.34e-11 ***
## black       -0.1381481  0.0399615  -3.457 0.000571 ***
## educ         0.0452410  0.0076469   5.916 4.64e-09 ***
## iq           0.0048228  0.0057333   0.841 0.400459
## kww         -0.0248007  0.0107382  -2.310 0.021132 *
## educiq0     -0.0001138  0.0004228  -0.269 0.787971
## educkww0     0.0021610  0.0007957   2.716 0.006735 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3613 on 923 degrees of freedom
## Multiple R-squared:  0.2728, Adjusted R-squared:  0.2641
## F-statistic: 31.48 on 11 and 923 DF,  p-value: < 2.2e-16
```

The interaction educkww0 is statistically significant, and the two interactions are jointly significant at the 2% signifiance level. The estimated return to education at the average values of IQ and KWW (in the population and sample, respectively) is somewhat smaller now: about 4.5%. Further, as KWW increases above its mean, the return to education increases.

```
linearHypothesis(lm4, c("educiq0=0", "educkww0=0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## educiq0 = 0
## educkww0 = 0
##
## Model 1: restricted model
## Model 2: lwage ~ exper + tenure + married + south + urban + black + educ +
##     iq + kww + educiq0 + educkww0
##
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    925 121.56
## 2    923 120.47  2    1.0949 4.1945 0.01537 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4.12.

Answer

Redo Example 4.4, adding the variable union-a dummy variable indicating whether the workers at the plant are unionized-as an additional explanatory variable.

First, we check the original model.The data in JTRAIN1.RAW are for 157 Michigan manufacturing firms for the years 1987, 1988, and 1989. These data are from Holzer, Block, Cheatham, and Knott (1993). The goal is to determine the effectiveness of job training grants on firm productivity. For this exercise, we use

only the 54 firms in 1988 that reported nonmissing values of the scrap rate (number of items out of 100 that must be scrapped). No firms were awarded grants in 1987; in 1988, 19 of the 54 firms were awarded grants. If the training grant has the intended effect, the average scrap rate should be lower among firms receiving a grant. The problem is that the grants were not randomly assigned: whether or not a firm received a grant could be related to other factors unobservable to the econometrician that affect productivity. In the simplest case, we can write (for the 1988 cross section).

$$log(scrap) = \beta_0 + \beta_1 grant + \gamma q + v$$

where v is orthogonal to grant but q contains unobserved productivity factors that might be correlated with grant, a binary variable equal to unity if the firm received a job training grant. Since we have the scrap rate in the previous year, we can use $log(scrap_{-1})$ as a proxy variable for q where r has zero mean and, by definition, is uncorrelated with it.

```
JTRAIN1 = read.csv("D:/Google Drive/Fordham/2019 Spring/AE/MITDataCSV/JTRAIN1.csv", header = TRUE)
lm5 = lm(lscrap ~ grant + union, data=JTRAIN1, subset=d88==1)
summary(lm5)
```

```
##
## Call:
## lm(formula = lscrap ~ grant + union, data = JTRAIN1, subset = d88 ==
##     1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2265 -0.8563  0.0086  0.9999  2.9473
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.23073    0.26486   0.871    0.388
## grant       -0.02762    0.40436  -0.068    0.946
## union        0.62229    0.40963   1.519    0.135
##
## Residual standard error: 1.406 on 51 degrees of freedom
##   (103 observations deleted due to missingness)
## Multiple R-squared:  0.04365,    Adjusted R-squared:  0.006146
## F-statistic: 1.164 on 2 and 51 DF,  p-value: 0.3204
```

Then we add the dummy variable

```
lm6 = lm(lscrap ~ grant + union + lscrap_1, data = JTRAIN1, subset=d88==1)
summary(lm6)
```

```
##
## Call:
## lm(formula = lscrap ~ grant + union + lscrap_1, data = JTRAIN1,
##     subset = d88 == 1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.79314 -0.22052  0.04252  0.27994  1.62148
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.04778    0.09588  -0.498   0.6205
## grant       -0.28511    0.14526  -1.963   0.0553 .
## union        0.25807    0.14778   1.746   0.0869 .
## lscrap_1     0.82103    0.04396  18.676   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5027 on 50 degrees of freedom
##   (103 observations deleted due to missingness)
## Multiple R-squared:  0.8801, Adjusted R-squared:  0.8729
## F-statistic: 122.3 on 3 and 50 DF,  p-value: < 2.2e-16
```

The basic story does not change: initially, the grant is estimated to have essentially no effect, but adding $log(scrap_{-1})$ gives the grant a strong effect that is marginally statistically significant. Interestingly, unionized firms are estimated to have larger scrap rates; over 25% more in the second equation. The effect is significant at the 10% level.

4.13.

Use the data in CORNWELL.RAW (from Cornwell and Trumball, 1994) to estimate a model of county-level crime rates, using the year 1987 only.

a. Using logarithms of all variables, estimate a model relating the crime rate to the deterrent variables prbarr, prbconv, prbpris, and avgsen.

Answer
Since this is a pooled model, with considering county and year in different levels, we apply plm package in the following parts.

```
CORNWELL = read.csv("D:/Google Drive/Fordham/2019 Spring/AE/MITDataCSV/CORNWELL.csv", header = TRUE)
lm7 = plm(lcrmrte ~ lprbarr + lprbconv + lprbpris + lavgsen, data=CORNWELL, subset=d87==1,index=c("coun
summary(lm7)
```

```
## Pooling Model
##
## Call:
## plm(formula = lcrmrte ~ lprbarr + lprbconv + lprbpris + lavgsen,
##     data = CORNWELL, subset = d87 == 1, model = "pooling", index = c("county",
##         "year"))
##
## Balanced Panel: n = 90, T = 1, N = 90
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -1.361046 -0.191289  0.079389  0.277536  0.868431
##
## Coefficients:
##              Estimate Std. Error  t-value  Pr(>|t|)
## (Intercept) -4.867923   0.431531 -11.2806 < 2.2e-16 ***
## lprbarr     -0.723970   0.115316  -6.2781 1.392e-08 ***
## lprbconv    -0.472511   0.083108  -5.6855 1.798e-07 ***
## lprbpris     0.159670   0.206444   0.7734    0.4414
```

```
## lavgsen      0.076422   0.163473   0.4675    0.6413
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    26.8
## Residual Sum of Squares: 15.645
## R-Squared:      0.41623
## Adj. R-Squared: 0.38876
## F-statistic: 15.1516 on 4 and 85 DF, p-value: 2.1714e-09
```

Because of the log-log functional form, all coefficients are elasticities. The elasticities of crime with respect to the arrest and conviction probabilities are the sign we expect, and both are practically and statistically significant. The elasticities with respect to the probability of serving a prison term and the average sentence length are positive but are statistically insignificant.

b. Add log(crmrte) for 1986 as an additional explanatory variable, and comment on how the estimated elasticities differ from part a.

Answer

```
lm8 = plm(lcrmrte ~ lprbarr + lprbconv + lprbpris + lavgsen + lag(lcrmrte), data=CORNWELL, subset=d87==
summary(lm8)
```

```
## Pooling Model
##
## Call:
## plm(formula = lcrmrte ~ lprbarr + lprbconv + lprbpris + lavgsen +
##     lag(lcrmrte), data = CORNWELL, subset = d87 == 1, model = "pooling",
##     index = c("county", "year"))
##
## Balanced Panel: n = 90, T = 1, N = 90
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -1.280226 -0.089311  0.030551  0.110188  0.324217
##
## Coefficients:
##               Estimate Std. Error t-value  Pr(>|t|)
## (Intercept)  -0.766626   0.313099 -2.4485  0.016425 *
## lprbarr      -0.185042   0.062762 -2.9483  0.004137 **
## lprbconv     -0.038677   0.046600 -0.8300  0.408903
## lprbpris     -0.126688   0.098851 -1.2816  0.203507
## lavgsen      -0.152023   0.078292 -1.9418  0.055519 .
## lag(lcrmrte)  0.779813   0.045211 17.2481 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    26.8
## Residual Sum of Squares: 3.4447
## R-Squared:      0.87146
## Adj. R-Squared: 0.86381
## F-statistic: 113.903 on 5 and 84 DF, p-value: < 2.22e-16
```

There are some notable changes in the coefficients on the original variables. The elasticities with respect to prbarr and prbconv are much smaller now, but still have signs predicted by a deterrent-effect story. The

conviction probability is no longer statistically significant. Adding the lagged crime rate changes the signs of the elasticities with respect to prbpris and avgsen, and the latter is almost statistically significant at the 5% level against a two-sided alternative (p-value = .056). Not surprisingly, the elasticity with respect to the lagged crime rate is large and very statistically significant. (The elasticity is also statistically less than unity.)

c. Compute the F statistic for joint significance of all of the wage variables (again in logs), using the restricted model from part b.

Answer
Note there is a little difference between our answer with the solution, due to little different pooled modedling algorithms, which is not a problem at all this stage. Additionally, we conduct a chi-square test first, then followed by a F-test, note that they are equivalent.

```
lm9 = plm(lcrmrte ~ lprbarr + lprbconv + lprbpris + lavgsen + lag(lcrmrte) + lwcon  + lwtuc + lwtrd + l
summary(lm9)
```

```
## Pooling Model
##
## Call:
## plm(formula = lcrmrte ~ lprbarr + lprbconv + lprbpris + lavgsen +
##     lag(lcrmrte) + lwcon + lwtuc + lwtrd + lwfir + lwser + lwmfg +
##     lwfed + lwsta + lwloc, data = CORNWELL, subset = d87 == 1,
##     model = "pooling", index = c("county", "year"))
##
## Balanced Panel: n = 90, T = 1, N = 90
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -1.085507 -0.088927  0.012477  0.108601  0.351013
##
## Coefficients:
##               Estimate Std. Error t-value Pr(>|t|)
## (Intercept)  -3.792531   1.957473 -1.9375  0.05645 .
## lprbarr      -0.172512   0.065953 -2.6157  0.01076 *
## lprbconv     -0.068364   0.049728 -1.3748  0.17330
## lprbpris     -0.215556   0.102401 -2.1050  0.03864 *
## lavgsen      -0.196055   0.084465 -2.3211  0.02300 *
## lag(lcrmrte)  0.745341   0.053033 14.0543  < 2e-16 ***
## lwcon        -0.285001   0.177518 -1.6055  0.11259
## lwtuc         0.064131   0.134327  0.4774  0.63445
## lwtrd         0.253708   0.231745  1.0948  0.27712
## lwfir        -0.083527   0.196497 -0.4251  0.67199
## lwser         0.112754   0.084743  1.3305  0.18737
## lwmfg         0.098737   0.118610  0.8325  0.40780
## lwfed         0.336129   0.245313  1.3702  0.17471
## lwsta         0.039510   0.207211  0.1907  0.84930
## lwloc        -0.036986   0.329155 -0.1124  0.91083
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    26.8
## Residual Sum of Squares: 2.9198
## R-Squared:      0.89105
## Adj. R-Squared: 0.87071
```

```
## F-statistic: 43.8136 on 14 and 75 DF, p-value: < 2.22e-16
```

```
linearHypothesis(lm9, c("lwcon=0","lwtuc=0","lwtrd=0","lwfir=0","lwser=0","lwmfg=0","lwfed=0","lwsta=0"
```

```
## Linear hypothesis test
##
## Hypothesis:
## lwcon = 0
## lwtuc = 0
## lwtrd = 0
## lwfir = 0
## lwser = 0
## lwmfg = 0
## lwfed = 0
## lwsta = 0
## lwloc = 0
##
## Model 1: restricted model
## Model 2: lcrmrte ~ lprbarr + lprbconv + lprbpris + lavgsen + lag(lcrmrte) +
##     lwcon + lwtuc + lwtrd + lwfir + lwser + lwmfg + lwfed + lwsta +
##     lwloc
##
##   Res.Df Df  Chisq Pr(>Chisq)
## 1     84
## 2     75  9 13.483     0.1419
```

d. Redo part c, but make the test robust to heteroskedasticity of unknown form.

Answer

We Adopt Hetero-robust LM test

```
lm9 = plm(lcrmrte ~ lprbarr + lprbconv + lprbpris + lavgsen + lag(lcrmrte) + lwcon  + lwtuc + lwtrd + lw

# 1. Compute u_tilda from restricted model
# In our model, lwcon, lwtuc, lwtrd, lwfir, lwser, lwmfg, lwfed, lwsta, lwloc are the tested variables,
lm_restrict2 = plm(lcrmrte ~ lprbarr + lprbconv + lprbpris + lavgsen + lag(lcrmrte), data=CORNWELL, sub
u_tilda2=resid(lm_restrict2)

# 2. Regress each of the independent v's excluded on all of the included v's, and obtain 9 sets of resi
lm_lwcon = plm(lcrmrte ~ lprbarr + lprbconv + lprbpris + lavgsen + lag(lcrmrte), data=CORNWELL, subset=
r_lwcon=resid(lm_lwcon)

lm_lwtuc = plm(lcrmrte ~ lprbarr + lprbconv + lprbpris + lavgsen + lag(lcrmrte), data=CORNWELL, subset=
r_lwtuc=resid(lm_lwtuc)

lm_lwtrd = plm(lcrmrte ~ lprbarr + lprbconv + lprbpris + lavgsen + lag(lcrmrte), data=CORNWELL, subset=
r_lwtrd=resid(lm_lwtrd)

lm_lwfir = plm(lcrmrte ~ lprbarr + lprbconv + lprbpris + lavgsen + lag(lcrmrte), data=CORNWELL, subset=
r_lwfir=resid(lm_lwfir)

lm_lwser = plm(lcrmrte ~ lprbarr + lprbconv + lprbpris + lavgsen + lag(lcrmrte), data=CORNWELL, subset=
r_lwser=resid(lm_lwser)
```

```
lm_lwmfg = plm(lcrmrte ~ lprbarr + lprbconv + lprbpris + lavgsen + lag(lcrmrte), data=CORNWELL, subset=c
r_lwmfg=resid(lm_lwmfg)

lm_lwfed = plm(lcrmrte ~ lprbarr + lprbconv + lprbpris + lavgsen + lag(lcrmrte), data=CORNWELL, subset=c
r_lwfed=resid(lm_lwfed)

lm_lwsta = plm(lcrmrte ~ lprbarr + lprbconv + lprbpris + lavgsen + lag(lcrmrte), data=CORNWELL, subset=c
r_lwsta=resid(lm_lwsta)

lm_lwloc = plm(lcrmrte ~ lprbarr + lprbconv + lprbpris + lavgsen + lag(lcrmrte), data=CORNWELL, subset=c
r_lwloc=resid(lm_lwloc)

# 3. Find the product of each r and u_tilda without intercept
lm_LM2=lm(rep(1,length(u_tilda2)) ~ r_lwcon*u_tilda2    + r_lwtuc*u_tilda2  + r_lwtrd*u_tilda2  + r_lwfi
LM_test2=length(u_tilda2)-sum(resid(lm_LM2)^2)

# 4. Conduct chi-square(9) test
c("Reject?",LM_test>qchisq(.975, df=9))
```

```
## [1] "Reject?" "FALSE"
```

4.14. Use the data in ATTEND.RAW to answer this question.

a. To determine the effects of attending lecture on final exam performance, estimate a model relating stndfnl (the standardized final exam score) to atndrte (the percent of lectures attended). Include the binary variables frosh and soph as explanatory variables. Interpret the coefficient on atndrte, and discuss its significance.

Answer

```
ATTEND = read.csv("D:/Google Drive/Fordham/2019 Spring/AE/MITDataCSV/ATTEND.csv", header = TRUE)
lm11 = lm(stndfnl ~ atndrte + frosh + soph, data=ATTEND)
summary(lm11)
```

```
##
## Call:
## lm(formula = stndfnl ~ atndrte + frosh + soph, data = ATTEND)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2825 -0.6719 -0.0295  0.6791  2.5455
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.501731   0.196314  -2.556 0.010813 *
## atndrte      0.008163   0.002203   3.705 0.000228 ***
## frosh       -0.289894   0.115724  -2.505 0.012478 *
## soph        -0.118446   0.099027  -1.196 0.232078
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9772 on 676 degrees of freedom
## Multiple R-squared:  0.02904,    Adjusted R-squared:  0.02473
## F-statistic: 6.739 on 3 and 676 DF,  p-value: 0.0001752
```

Because the final exam score has been standardized, it has close to a zero mean and its standard deviation is close to one. The values are not closer to zero and one, respectively, because the standardization was done with a larger data set that included students with missing values on other key variables. It might make sense to redefine the standardized test score using the mean and standard deviation in the sample of 680, but the effect should be minor.
If atndrte increases by 10 percentage points (say, from 75 to 85), the standardized test score is estimated to increase by about .082 standard deviations.

b. How confident are you that the OLS estimates from part a are estimating the causal effect of attendence? Explain.

Answer

Certainly there is a potential for self-selection. The better students may also be the ones attending lecture more regularly. So the positive effect of the attendance rate simply might capture the fact that better students tend to do better on exams. It is unlikely that controlling just for year in college (frosh and soph) solves the endogeneity of atndrete.

c.As proxy variables for student ability, add to the regression priGPA (prior cumulative GPA) and ACT (achievement test score). Now what is the effect of atndrte? Discuss how the effect differs from that in part a.

Answer

```
lm12 = lm(stndfnl ~ atndrte + frosh + soph + priGPA + ACT, data=ATTEND)
summary(lm12)
```

```
##
## Call:
## lm(formula = stndfnl ~ atndrte + frosh + soph + priGPA + ACT,
##     data = ATTEND)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1904 -0.5668 -0.0252  0.5887  2.2915
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.297342   0.308831 -10.677  < 2e-16 ***
## atndrte      0.005225   0.002384   2.191   0.0288 *
## frosh       -0.049469   0.107890  -0.459   0.6467
## soph        -0.159648   0.089772  -1.778   0.0758 .
## priGPA       0.426584   0.081920   5.207 2.55e-07 ***
## ACT          0.084412   0.011168   7.559 1.33e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8851 on 674 degrees of freedom
## Multiple R-squared:  0.2058, Adjusted R-squared:  0.1999
## F-statistic: 34.93 on 5 and 674 DF,  p-value: < 2.2e-16
```

The effect of atndrte has fallen, which is what we expect if we think better, smarter students also attend lectures more frequently. The estimate now is that a 10 percentage point increase in atndrte increases the standardized test score by .052 standard deviations; the effect is statistically significant at the usual 5% level against a two-sided alternative, but the t statistic is much lower than in part a. The strong positive effects of prior GPA and ACT score are also expected.

d.

Answer

Controlling for priGPA and ACT causes the sophomore effect (relative to students in year three and beyond) to get slightly larger in magnitude and more statistically significant. These data are for a course taught in the second term, so each frosh student does have a prior GPA - his or her GPA for the first semester in college. Adding priGPA in particular causes the "freshman effect" to essentially disappear. This is not too surprising because the average prior GPA for first-year students is notably less than the overall average priGPA.

e. Add the squares of priGPA and ACT to the equation. What happens to the coefficient on atndrte? Are the quadratics jointly significant?

Answer

```
priGPAsq = ATTEND$priGPA^2
ACTsq = ATTEND$ACT^2
lm13 = lm(stndfnl ~ atndrte + frosh + soph + priGPA + ACT + priGPAsq + ACTsq, data=ATTEND)
summary(lm13)
```

```
##
## Call:
## lm(formula = stndfnl ~ atndrte + frosh + soph + priGPA + ACT +
##     priGPAsq + ACTsq, data = ATTEND)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.14241 -0.54902 -0.02163  0.56155  2.36547
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.384811   1.239361   1.117  0.26424
## atndrte      0.006232   0.002358   2.642  0.00842 **
## frosh       -0.105337   0.106975  -0.985  0.32513
## soph        -0.180729   0.088635  -2.039  0.04184 *
## priGPA      -1.526139   0.473971  -3.220  0.00134 **
## ACT         -0.112433   0.098172  -1.145  0.25251
## priGPAsq     0.368218   0.088985   4.138 3.95e-05 ***
## ACTsq        0.004182   0.002169   1.928  0.05425 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8718 on 672 degrees of freedom
## Multiple R-squared:  0.2316, Adjusted R-squared:  0.2236
## F-statistic: 28.94 on 7 and 672 DF,  p-value: < 2.2e-16
```

```
linearHypothesis(lm13, c("priGPAsq=0","ACTsq=0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## priGPAsq = 0
## ACTsq = 0
##
```

```
## Model 1: restricted model
## Model 2: stndfnl ~ atndrte + frosh + soph + priGPA + ACT + priGPAsq +
##     ACTsq
##
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    674 527.96
## 2    672 510.79  2    17.172 11.296 1.496e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Adding the squared terms - one of which is very significant, the other of which is marginally significant - actually increases the attendance rate effect. And it does so while slightly reducing the standard error on atndrte, resulting in a t statistic that is notably more significant than in part c.

f. To test for a nonlinear effect of atndrte, add its square to the equation from part e. What do you conclude?

Answer

Adding the squared attendance rate is not warranted, as it is very insignificant:

```
atndrtesq = ATTEND$atndrte^2
lm14 = lm(stndfnl ~ atndrte + frosh + soph + priGPA + ACT + priGPAsq + ACTsq + atndrtesq, data=ATTEND)
summary(lm14)
```

```
##
## Call:
## lm(formula = stndfnl ~ atndrte + frosh + soph + priGPA + ACT +
##     priGPAsq + ACTsq + atndrtesq, data = ATTEND)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.14165 -0.54816 -0.02205  0.55940  2.36637
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.394e+00  1.267e+00   1.100  0.27159
## atndrte      5.843e-03  1.092e-02   0.535  0.59282
## frosh       -1.054e-01  1.071e-01  -0.984  0.32537
## soph        -1.808e-01  8.875e-02  -2.038  0.04199 *
## priGPA      -1.525e+00  4.757e-01  -3.205  0.00141 **
## ACT         -1.123e-01  9.828e-02  -1.143  0.25339
## priGPAsq     3.679e-01  8.944e-02   4.113 4.38e-05 ***
## ACTsq        4.180e-03  2.171e-03   1.925  0.05461 .
## atndrtesq    2.873e-06  7.871e-05   0.036  0.97089
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8725 on 671 degrees of freedom
## Multiple R-squared:  0.2316, Adjusted R-squared:  0.2225
## F-statistic: 25.28 on 8 and 671 DF,  p-value: < 2.2e-16
```

The very large increase in the standard error on atndrte suggest that atndrte and atndrte2 are highly collinear. In fact, their sample correlation is about .983. Importantly, the coefficient on atndrte now has an uninteresting interpretation: it measures the partial effect of atndrte starting from atndrte = 0. The lowest

attendance rate in the sample is 6.25, with the vast majority of students (94.3%) attending 50 percent or more of the lectures. If the quadratic term were significant, we might want to center atndrte about its mean or median before creating the square. Or, a more sophisticated functional form might be called for. It may be better to define several intervals for atndrrte and include dummy variables for those intervals.