PS5 R Solutions

Matthew Alampay Davis

November 23, 2021

Question 1

```
rental <- read.dta13("rental.dta")
```

Part a: Estimate the equation by pooled OLS and report the results in standard form. Whatdo you make of the estimate on the 1990 dummy variable?

```
mod.pool <- lm(lrent ~ y90 + lpop + lavginc + pctstu, rental)
summary(mod.pool)
```

```
##
## Call:
## lm(formula = lrent ~ y90 + lpop + lavginc + pctstu, data = rental)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.24233 -0.07824 -0.01642  0.04389  0.48082
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.568807   0.534881  -1.063   0.2897
## y90          0.262227   0.034763   7.543 8.78e-12 ***
## lpop         0.040686   0.022515   1.807   0.0732 .
## lavginc      0.571446   0.053098  10.762  < 2e-16 ***
## pctstu       0.005044   0.001019   4.949 2.40e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1259 on 123 degrees of freedom
## Multiple R-squared:  0.8613, Adjusted R-squared:  0.8568
## F-statistic: 190.9 on 4 and 123 DF,  p-value: < 2.2e-16
```

The positive and very significant coefficient on y90 simply means that, other things in the equation fixed, nominal rents grew by over 26% over the 10-year period.

Part b: Interpret the sample coefficient of pctstu

The coefficient on pctstu means that a one percentage point increase in pctstu increases rent by half a percent (.5%). The t statistic of five shows that, at least based on the usual analysis, pctstu is very statistically significant.

Part c: Are the standard errors you report in part (a) valid? Explain.

The standard errors from part (i) are not valid, unless we think ai does not really appear in the equation. If ai is in the error term, the errors across the two time periods for each city are positively correlated, and this invalidates the usual OLS standard errors and t statistics.

Part d: Now, difference the equation and estimate by OLS. Compare your estimate of β_3 with that of part (a). Does the relative size of the student population appear to affect rental prices?

```
# rental %<>% group_by(city) %>% mutate(rent.diff =
# diff(lrent), pop.diff = diff(lpop), inc.diff =
# diff(lavginc), pct.diff = diff(pctstu))
rental %<>%
    group_by(city) %>%
    mutate(rent.diff = lrent - lag(lrent), pop.diff = lpop -
        lag(lpop), inc.diff = lavginc - lag(lavginc), pct.diff = pctstu -
        lag(pctstu))
mod.diff <- lm(rent.diff ~ pop.diff + inc.diff + pct.diff, rental)
summary(mod.diff)
```

```
##
## Call:
## lm(formula = rent.diff ~ pop.diff + inc.diff + pct.diff, data = rental)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.18697 -0.06216 -0.01438  0.05518  0.23783
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.385521   0.036824  10.469 3.66e-15 ***
## pop.diff    0.072246   0.088343   0.818  0.41671
## inc.diff    0.309961   0.066477   4.663 1.79e-05 ***
## pct.diff    0.011203   0.004132   2.711  0.00873 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09013 on 60 degrees of freedom
##   (64 observations deleted due to missingness)
## Multiple R-squared:  0.3223, Adjusted R-squared:  0.2884
## F-statistic:  9.51 on 3 and 60 DF,  p-value: 3.136e-05
```

Interestingly, the effect of pctstu is over twice as large as we estimated in the pooled OLS equation. Now, a one percentage point increase in pctstu is estimated to increase rental rates by about 1.1%. Not surprisingly,

we obtain a much less precise estimate when we difference (although the OLS standard errors from part (i) are likely to be much too small because of the positive serial correlation in the errors within each city). While we have differenced away the individual fixed effe, there may be other unobservables that change over time and are correlated with diff-pctstu.

Part e: Obtain the heteroskedasticity-robust standard errors for the first-differenced equation in part (d)

```
mod.diff.hetero <- lm_robust(rent.diff ~ pop.diff + inc.diff +
    pct.diff, rental, se_type = "stata")
summary(mod.diff.hetero)
```

```
##
## Call:
## lm_robust(formula = rent.diff ~ pop.diff + inc.diff + pct.diff,
##     data = rental, se_type = "stata")
##
## Standard error type:  HC1
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper DF
## (Intercept)  0.38552   0.048719   7.913 6.891e-11  0.28807  0.48297 60
## pop.diff     0.07225   0.069680   1.037 3.040e-01 -0.06713  0.21163 60
## inc.diff     0.30996   0.089310   3.471 9.683e-04  0.13131  0.48861 60
## pct.diff     0.01120   0.002936   3.816 3.233e-04  0.00533  0.01708 60
##
## Multiple R-squared:  0.3223 ,    Adjusted R-squared:  0.2884
## F-statistic:  11.3 on 3 and 60 DF,  p-value: 5.638e-06
```

The heteroskedasticity-robust standard error on pctstu is about .0029, which is actually much smaller than the usual OLS standard error (0.0041). This only makes pctstu even more significant (robust t statistic of roughly 4). Note that serial correlation is no longer an issue because we have no time component in the first-differenced equation.

Part f: Estimate the model by fixed effects to verify that you get identical estimates and standard errors to those in part (d) (use areg and xtreg commands and report both results)

```
mod.fe <- lm_robust(lrent ~ y90 + lpop + lavginc + pctstu, rental,
    fixed_effects = city, se_type = "stata")
summary(mod.fe)
```

```
##
## Call:
## lm_robust(formula = lrent ~ y90 + lpop + lavginc + pctstu, data = rental,
##     fixed_effects = city, se_type = "stata")
##
## Standard error type:  HC1
```

```
##
## Coefficients:
##         Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper DF
## y90      0.38552   0.048719   7.913 6.891e-11  0.28807  0.48297 60
## lpop     0.07225   0.069680   1.037 3.040e-01 -0.06713  0.21163 60
## lavginc  0.30996   0.089310   3.471 9.683e-04  0.13131  0.48861 60
## pctstu   0.01120   0.002936   3.816 3.233e-04  0.00533  0.01708 60
##
## Multiple R-squared:  0.9827 ,    Adjusted R-squared:  0.9633
## Multiple R-squared (proj. model):  0.9765 ,  Adjusted R-squared (proj. model):  0.9503
## F-statistic (proj. model): 691.4 on 4 and 60 DF,  p-value: < 2.2e-16
```

Matt: Stata's areg and xtreg actually don't seem to produce the same standard errors according to the official solutions. The command above gives equivalent SEs to the areg implementation of fixed effects.

Question 2

```
murder <- read.dta13("murder.dta")
```

Part a: Consider the unobserved effects model where η_t simply denotes different year intercepts and ai is the unobserved state effect. If past executions of convicted murderers have a deterrent effect, what should be the sign of β_1? What sign do you think β_2 should have? Explain.

If there is a deterrent effect, then $\beta_1 < 0$. The sign of β_2 is not entirely obvious, although one possibility is that a better economy means less crime in general, including violent crime (such as drug dealing) that would lead to fewer murders. This would imply $\beta_2 > 0$.

Part b: Using just the years 1990 and 1993, estimate the equation from part (i) by pooled OLS. Ignore the serial correlation problem in the composite errors. Do you find any evidence for a deterrent effect?

```
murder.90s <- filter(murder, year %in% c(90, 93))
mod.90s <- lm(mrdrte ~ d90 + d93 + exec + unem, murder.90s)
summary(mod.90s)
```

```
##
## Call:
## lm(formula = mrdrte ~ d90 + d93 + exec + unem, data = murder.90s)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.067  -3.356  -1.647   1.607  66.387
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -7.3454     5.0927  -1.442  0.15239
```

```
## d90           2.0674     2.1446   0.964  0.33742
## d93               NA         NA      NA       NA
## exec          0.1277     0.2632   0.485  0.62860
## unem          2.5289     0.7817   3.235  0.00166 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.22 on 98 degrees of freedom
## Multiple R-squared:  0.1016, Adjusted R-squared:  0.07411
## F-statistic: 3.695 on 3 and 98 DF,  p-value: 0.0144
```

There is no evidence of a deterrent effect, as the coefficient on exec is actually positive (though not statistically significant).

Part c: Now, using 1990 and 1993, estimate the equation by fixed effects. You may use first differencing since you are only using two years of data. Is there evidence of a deterrent effect? How strong?

```
mod.90s.fe <- lm_robust(mrdrte ~ d93 + exec + unem, murder.90s,
    fixed_effects = id, se_type = "stata")
summary(mod.90s.fe)
```

```
##
## Call:
## lm_robust(formula = mrdrte ~ d93 + exec + unem, data = murder.90s,
##     fixed_effects = id, se_type = "stata")
##
## Standard error type:  HC1
##
## Coefficients:
##      Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper DF
## d93   0.41327     0.2000  2.0663 4.422e-02  0.01113  0.81540 48
## exec -0.10384     0.0170 -6.1084 1.712e-07 -0.13802 -0.06966 48
## unem -0.06659     0.1469 -0.4532 6.524e-01 -0.36201  0.22883 48
##
## Multiple R-squared:  0.9975 ,    Adjusted R-squared:  0.9948
## Multiple R-squared (proj. model):  0.1653 ,  Adjusted R-squared (proj. model):  -0.7563
## F-statistic (proj. model): 13.52 on 3 and 48 DF,  p-value: 1.606e-06
```

Or by first differencing (note use of non-robust standard errors to match solutions):

```
murder.90s %<>%
    group_by(id) %>%
    mutate(murder.diff = mrdrte - lag(mrdrte), exec.diff = exec -
        lag(exec), unem.diff = unem - lag(unem))
mod.diff <- lm(murder.diff ~ exec.diff + unem.diff, murder.90s)
summary(mod.diff)
```

```
##
## Call:
```

```
## lm(formula = murder.diff ~ exec.diff + unem.diff, data = murder.90s)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.29995 -0.63573 -0.03594  0.83578  2.55896
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.41327    0.20938   1.974   0.0542 .
## exec.diff   -0.10384    0.04341  -2.392   0.0207 *
## unem.diff   -0.06659    0.15869  -0.420   0.6766
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.079 on 48 degrees of freedom
##   (51 observations deleted due to missingness)
## Multiple R-squared:  0.1097, Adjusted R-squared:  0.07266
## F-statistic: 2.959 on 2 and 48 DF,  p-value: 0.06142
```

Now, there is a statistically significant deterrent effect: 10 more executions is estimated to reduce the murder rate by 1.04, or one murder per 100,000 people. Is this a large effect? Executions are relatively rare in most states, but murder rates are relatively low on average, too. In 1993, the average murder rate was about 8.7; a reduction of one would be nontrivial. For the (unknown) people whose lives might be saved via a deterrent effect, it would seem important.

Part d: Compute the heteroskedasticity-robust standard error for the estimation in part (ii)

The heteroskedasticity-robust standard error for exec is .017. Somewhat surprisingly, this is well below the non-robust standard error. If we use the robust standard error, the statistical evidence for the deterrent effect is quite strong (roughly t=-6.1). See also Computer Exercise 13.12.

Part e: Find the state that has the largest number for the execution variable in 1993. (The variable exec is total executions in 1991, 1992, and 1993.) How much bigger is this value than the next highest value?

```
murder.90s %>%
    filter(year == 93) %>%
    arrange(-exec) %>%
    head
```

```
## # A tibble: 6 x 16
## # Groups:   id [6]
##   id    state  year mrdrte  exec  unem   d90   d93 cmrdrte cexec cunem cexec_1
##   <fct> <chr> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>   <dbl> <dbl> <dbl>   <dbl>
## 1 TX    TX       93  11.9     34  7         0     1  -2.20     23 0.800     -11
## 2 VA    VA       93   8.30    11  5         0     1  -0.5       8 0.700      -1
## 3 FL    FL       93   8.90     7  7         0     1  -1.80     -1 1.10        1
## 4 MO    MO       93  11.3      6  6.40      0     1   2.5       1 0.700       5
## 5 AZ    AZ       93   8.60     3  6.20      0     1   0.900     3 0.900       0
```

```
## 6 GA    GA       93  11.4      3  5.80     0     1  -0.400     1 0.400      -7
## # ... with 4 more variables: cunem_1 <dbl>, murder.diff <dbl>, exec.diff <dbl>,
## #   unem.diff <dbl>
```

Texas had by far the largest value of exec, 34. The next highest state was Virginia, with 11. These are three-year totals.

Part f: Estimate the equation using first differencing, dropping Texas from the analysis. Compute the usual and heteroskedasticity-robust standard errors. Now, what do you find? What is going on?

```
mod.diff.notexas <- lm(murder.diff ~ exec.diff + unem.diff, filter(murder.90s,
    id != "TX"))
summary(mod.diff.notexas)
```

```
##
## Call:
## lm(formula = murder.diff ~ exec.diff + unem.diff, data = filter(murder.90s,
##     id != "TX"))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2985 -0.5851 -0.1090  0.8045  2.6366
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.41252    0.21128   1.952   0.0569 .
## exec.diff   -0.06747    0.10491  -0.643   0.5233
## unem.diff   -0.07003    0.16037  -0.437   0.6643
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.089 on 47 degrees of freedom
##   (50 observations deleted due to missingness)
## Multiple R-squared:  0.01338,    Adjusted R-squared:  -0.02861
## F-statistic: 0.3186 on 2 and 47 DF,  p-value: 0.7287
```

Now the estimated deterrent effect is smaller. Perhaps more importantly, the standard error on exec has increased by a substantial amount. This happens because when we drop Texas, we lose much of the variation in the key explanatory variable, exec.

Part g: Use all three years of data and estimate the model by fixed effects. Include Texas in the analysis. Discuss the size and statistical significance of the deterrent effect comparedwith only using 1990 and 1993.

```
mod.fe <- lm_robust(mrdrte ~ exec + unem, murder, fixed_effects = ~id +
    year, se_type = "stata")
summary(mod.fe)
```

```
##
## Call:
## lm_robust(formula = mrdrte ~ exec + unem, data = murder, fixed_effects = ~id +
##     year, se_type = "stata")
##
## Standard error type:  HC1
##
## Coefficients:
##      Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper DF
## exec  -0.1383    0.07453  -1.856  0.06648  -0.2862 0.009589 98
## unem   0.2213    0.33736   0.656  0.51335  -0.4482 0.890799 98
##
## Multiple R-squared:  0.9054 ,    Adjusted R-squared:  0.8533
## Multiple R-squared (proj. model):  0.01088 , Adjusted R-squared (proj. model):  -0.5341
## F-statistic (proj. model):  1.88 on 2 and 98 DF,  p-value: 0.1581
```

The size of the deterrent effect is actually slightly larger than when 1987 is not used. However, the coefficient is not significant. Thus, while the magnitude of the effect is similar, the statistical significance is not. It is somewhat odd that adding another year of data causes the standard error on the exec coefficient to increase nontrivially.

Matt: the standard errors here are different from what are in the Stata solutions because they apparently did not use robust standard errors. You can recover their standard errors running the following:

```
summary(lm(mrdrte ~ exec + unem + factor(year) + factor(id),
    murder))
```

```
##
## Call:
## lm(formula = mrdrte ~ exec + unem + factor(year) + factor(id),
##     data = murder)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.6858  -0.6584  -0.0657   0.6747  13.3941
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.9038     3.3143   1.781  0.07796 .
## exec            -0.1383     0.1770  -0.781  0.43642
## unem             0.2213     0.2964   0.747  0.45701
## factor(year)90   1.5562     0.7453   2.088  0.03939 *
## factor(year)93   1.7332     0.7004   2.475  0.01506 *
## factor(id)AL     2.6177     2.9457   0.889  0.37636
## factor(id)AR     0.9869     2.9159   0.338  0.73575
## factor(id)AZ    -0.2343     2.9806  -0.079  0.93749
## factor(id)CA     3.4389     2.9164   1.179  0.24119
## factor(id)CO    -3.0467     2.9692  -1.026  0.30736
## factor(id)CT    -2.6440     3.0634  -0.863  0.39020
## factor(id)DC    55.5877     2.9009  19.162  < 2e-16 ***
## factor(id)DE    -2.9702     3.0989  -0.958  0.34018
## factor(id)FL     3.0048     3.2552   0.923  0.35823
## factor(id)GA     4.0799     3.1287   1.304  0.19527
## factor(id)HI    -3.5970     3.2123  -1.120  0.26556
```

```
## factor(id)IA    -5.9109     3.0952  -1.910  0.05910 .
## factor(id)ID    -5.5683     2.9243  -1.904  0.05982 .
## factor(id)IL     1.4506     2.9064   0.499  0.61884
## factor(id)IN    -1.7749     2.9918  -0.593  0.55437
## factor(id)KS    -3.1218     3.0737  -1.016  0.31230
## factor(id)KY    -1.4347     2.9094  -0.493  0.62303
## factor(id)LA     8.0028     3.0078   2.661  0.00911 **
## factor(id)MA    -4.5546     3.0155  -1.510  0.13415
## factor(id)MD     3.1599     3.0500   1.036  0.30274
## factor(id)ME    -6.1172     2.9793  -2.053  0.04271 *
## factor(id)MI     2.1546     2.8905   0.745  0.45779
## factor(id)MN    -5.2289     3.0402  -1.720  0.08860 .
## factor(id)MO     1.6162     3.0376   0.532  0.59588
## factor(id)MS     3.3342     2.8842   1.156  0.25048
## factor(id)MT    -4.4167     2.9377  -1.503  0.13594
## factor(id)NC     2.1755     3.1132   0.699  0.48633
## factor(id)ND    -6.6554     3.1063  -2.143  0.03462 *
## factor(id)NE    -4.3492     3.2622  -1.333  0.18556
## factor(id)NH    -5.7847     3.0600  -1.890  0.06166 .
## factor(id)NJ    -3.0434     3.0067  -1.012  0.31392
## factor(id)NM     0.4251     2.8867   0.147  0.88322
## factor(id)NV     1.4006     2.9938   0.468  0.64095
## factor(id)NY     4.7200     2.9692   1.590  0.11513
## factor(id)OH    -2.4500     2.9377  -0.834  0.40632
## factor(id)OK    -0.2969     2.9503  -0.101  0.92004
## factor(id)OR    -3.7279     2.9439  -1.266  0.20841
## factor(id)PA    -2.0355     2.9619  -0.687  0.49356
## factor(id)RI    -4.2762     2.9596  -1.445  0.15168
## factor(id)SC     2.1377     2.9834   0.717  0.47537
## factor(id)SD    -5.4412     3.1863  -1.708  0.09086 .
## factor(id)TN     1.6421     2.9767   0.552  0.58245
## factor(id)TX     7.0622     4.9264   1.434  0.15489
## factor(id)UT    -4.8057     3.0736  -1.564  0.12115
## factor(id)VA     1.0004     3.3107   0.302  0.76315
## factor(id)VT    -5.1664     3.0843  -1.675  0.09712 .
## factor(id)WA    -3.1962     2.9240  -1.093  0.27703
## factor(id)WI    -3.9549     3.0434  -1.299  0.19683
## factor(id)WV    -3.4060     2.9109  -1.170  0.24481
## factor(id)WY    -4.9520     2.9353  -1.687  0.09478 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.521 on 98 degrees of freedom
## Multiple R-squared:  0.9054, Adjusted R-squared:  0.8533
## F-statistic: 17.37 on 54 and 98 DF,  p-value: < 2.2e-16
```

Question 3

```
pension <- read.dta13("pension.dta")
```

Part a: Ignoring the clustering by family, use OLS to estimate the model where the variables are defined in the data set. The variable of most interest is choice, which is a dummy variable equal to one if the worker has a choice in how to allocate pension funds among different investments. What is the estimated effect of choice? Is it statistically significant?

```
mod.ols <- lm(pctstck ~ choice + prftshr + female + age + educ +
    finc25 + finc35 + finc50 + finc75 + finc100 + finc101 + wealth89 +
    stckin89 + irain89, pension)
summary(mod.ols)
```

```
##
## Call:
## lm(formula = pctstck ~ choice + prftshr + female + age + educ +
##     finc25 + finc35 + finc50 + finc75 + finc100 + finc101 + wealth89 +
##     stckin89 + irain89, data = pension)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -73.062 -34.797  -0.385  34.235  73.032
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 128.544172  55.169906   2.330   0.0209 *
## choice       11.744324   6.232065   1.884   0.0611 .
## prftshr      14.336095   7.231378   1.982   0.0490 *
## female        1.452231   6.765598   0.215   0.8303
## age          -1.500617   0.776582  -1.932   0.0549 .
## educ          0.703626   1.196754   0.588   0.5573
## finc25      -15.288670  14.229461  -1.074   0.2841
## finc35        0.188015  14.692879   0.013   0.9898
## finc50       -3.861741  14.551245  -0.265   0.7910
## finc75      -13.748071  16.021946  -0.858   0.3920
## finc100      -2.686100  15.718504  -0.171   0.8645
## finc101     -25.050361  17.800382  -1.407   0.1611
## wealth89     -0.002559   0.012776  -0.200   0.8415
## stckin89      6.674152   6.682981   0.999   0.3193
## irain89      -7.497800   6.377805  -1.176   0.2413
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38.96 on 179 degrees of freedom
## Multiple R-squared:  0.1076, Adjusted R-squared:  0.0378
## F-statistic: 1.542 on 14 and 179 DF,  p-value: 0.1004
```

Investment choice is associated with about 11.7 percentage points more in stocks. The t-statistic is 1.88 so it is marginally significant.

Part b: Are the income, wealth, stock holding, and IRA holding control variables important? Explain.

```
linearHypothesis(mod.ols, c("finc25 = 0", "finc35 = 0", "finc50 = 0",
    "finc75 = 0", "finc100 = 0", "finc101 = 0", "wealth89 = 0",
    "stckin89 = 0", "irain89 = 0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## finc25 = 0
## finc35 = 0
## finc50 = 0
## finc75 = 0
## finc100 = 0
## finc101 = 0
## wealth89 = 0
## stckin89 = 0
## irain89 = 0
##
## Model 1: restricted model
## Model 2: pctstck ~ choice + prftshr + female + age + educ + finc25 + finc35 +
##     finc50 + finc75 + finc100 + finc101 + wealth89 + stckin89 +
##     irain89
##
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    188 285854
## 2    179 271768  9     14086 1.0309 0.4172
```

These variables are not very important. The F test for joint significant is 1.03. With 9 and 179 df, this gives p-value = .42. Plus, when these variables are dropped from the regression, the coefficient on choice only falls to 11.15.

Part c: Determine how many different families there are in the data set.

```
length(unique(pension$id))
```

```
## [1] 171
```

There are 171 different families in the sample

Part d: Now, obtain the standard errors for OLS that are robust to cluster correlation within a family. Do they differ much from the usual OLS standard errors? Are you surprised?

```
mod.ols.cluster <- lm_robust(pctstck ~ choice + prftshr + female +
    age + educ + finc25 + finc35 + finc50 + finc75 + finc100 +
    finc101 + wealth89 + stckin89 + irain89, pension, clusters = id,
    se_type = "stata")
summary(mod.ols.cluster)
```

```
##
## Call:
## lm_robust(formula = pctstck ~ choice + prftshr + female + age +
##     educ + finc25 + finc35 + finc50 + finc75 + finc100 + finc101 +
##     wealth89 + stckin89 + irain89, data = pension, clusters = id,
##     se_type = "stata")
##
## Standard error type:  stata
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)  CI Lower  CI Upper  DF
## (Intercept) 128.544172   56.96060  2.25672   0.0253  16.10300 240.98534 170
## choice       11.744324    6.19771  1.89494   0.0598  -0.49007  23.97872 170
## prftshr      14.336095    8.20762  1.74668   0.0825  -1.86588  30.53807 170
## female        1.452231    6.64278  0.21862   0.8272 -11.66072  14.56519 170
## age          -1.500617    0.80980 -1.85308   0.0656  -3.09917   0.09794 170
## educ          0.703626    1.17690  0.59786   0.5507  -1.61960   3.02685 170
## finc25      -15.288670   16.44093 -0.92992   0.3537 -47.74333  17.16599 170
## finc35        0.188015   16.31252  0.01153   0.9908 -32.01318  32.38920 170
## finc50       -3.861741   15.96283 -0.24192   0.8091 -35.37263  27.64915 170
## finc75      -13.748071   16.92901 -0.81210   0.4179 -47.16622  19.67007 170
## finc100      -2.686100   17.06689 -0.15739   0.8751 -36.37642  31.00422 170
## finc101     -25.050361   17.20684 -1.45584   0.1473 -59.01695   8.91623 170
## wealth89     -0.002559    0.01194 -0.21434   0.8305  -0.02612   0.02101 170
## stckin89      6.674152    7.25819  0.91953   0.3591  -7.65363  21.00193 170
## irain89      -7.497800    6.26496 -1.19678   0.2331 -19.86494   4.86934 170
##
## Multiple R-squared:  0.1076 ,    Adjusted R-squared:  0.0378
## F-statistic: 2.248 on 14 and 170 DF,  p-value: 0.007955
```

The instructor reported only the cluster-robust standard error for choice: 6.20. Therefore, it is essentially the same as the usual OLS standard error. This is not very surprising, because at least 171 of the 194 observations can be assumed independent of one another. The explanatory variables may adequately capture the within-family correlation.

Part e: Estimate the equation by differencing across only the spouses within a family. Why do the explanatory variables asked about in part (ii) drop out in the first-differenced estimation?

```
pension %<>%
    group_by(id) %>%
    mutate(pctstck.diff = pctstck - lag(pctstck), choice.diff = choice -
        lag(choice), prftshr.diff = prftshr - lag(prftshr), female.diff = female -
        lag(female), age.diff = age - lag(age), educ.diff = educ -
```

```
        lag(educ))
mod.diff <- lm(pctstck.diff ~ choice.diff + prftshr.diff + female.diff +
    age.diff + educ.diff, pension)
summary(mod.diff)
```

```
##
## Call:
## lm(formula = pctstck.diff ~ choice.diff + prftshr.diff + female.diff +
##     age.diff + educ.diff, data = pension)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -84.625 -15.927   4.404  10.373  81.798
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    15.927     10.938   1.456    0.164
## choice.diff     2.276     15.000   0.152    0.881
## prftshr.diff   -9.267     16.924  -0.548    0.591
## female.diff    21.551     21.485   1.003    0.330
## age.diff       -3.573      8.999  -0.397    0.696
## educ.diff      -1.220      3.429  -0.356    0.726
##
## Residual standard error: 34.17 on 17 degrees of freedom
##   (171 observations deleted due to missingness)
## Multiple R-squared:  0.206,  Adjusted R-squared:  -0.02754
## F-statistic: 0.8821 on 5 and 17 DF,  p-value: 0.514
```

All of the income and wealth variables, and the stock and IRA indicators, drop out, as these are defined at the family level (and therefore are the same for the husband and wife)

Part f: Are any of the remaining explanatory variables in part (v) significant? Are you surprised?

None of the explanatory variables is significant in part (v), and this is not too surprising. We have only 23 observations, and we are removing much of the variation in the explanatory variables (except the gender variable) by using within-family differences.