PS2

Redmond Xia

June 21, 2020

Libraries

```
library(data.table)
library(ggplot2)
library(lfe)
```

```
## Loading required package: Matrix
```

```
library(foreign)
library(stargazer)
```

```
##
## Please cite as:
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

Question 1: Fixed effects and within transformations

You will find a modified version the imports-85.csv (imports85_modified.csv) file attached to this assignment. Again, make sure that all continuous variables of interest are numeric.

1. Regress fuel efficiency (city.mpg) on horsepower without fixed effects. What would you conclude based on that regression?

```
imports <- as.data.table(read.csv("imports85_modified.csv"))
imports$horsepower <- as.numeric(as.character(imports$horsepower))
```

```
## Warning: NAs introduced by coercion
```

```
imports$city.mpg<-as.numeric(as.character(imports$city))
imports$num.of.cylinders<-as.character(imports$num.of.cylinders)
imports$peak.rpm<-as.numeric(as.character(imports$peak.rpm))
```

```
## Warning: NAs introduced by coercion
```

```
imports <- imports[!is.na(horsepower)&!is.na(city.mpg),]

ols <- lm(city.mpg ~ horsepower, data = imports)
summary(ols)
```

```
##
## Call:
## lm(formula = city.mpg ~ horsepower, data = imports)
##
```

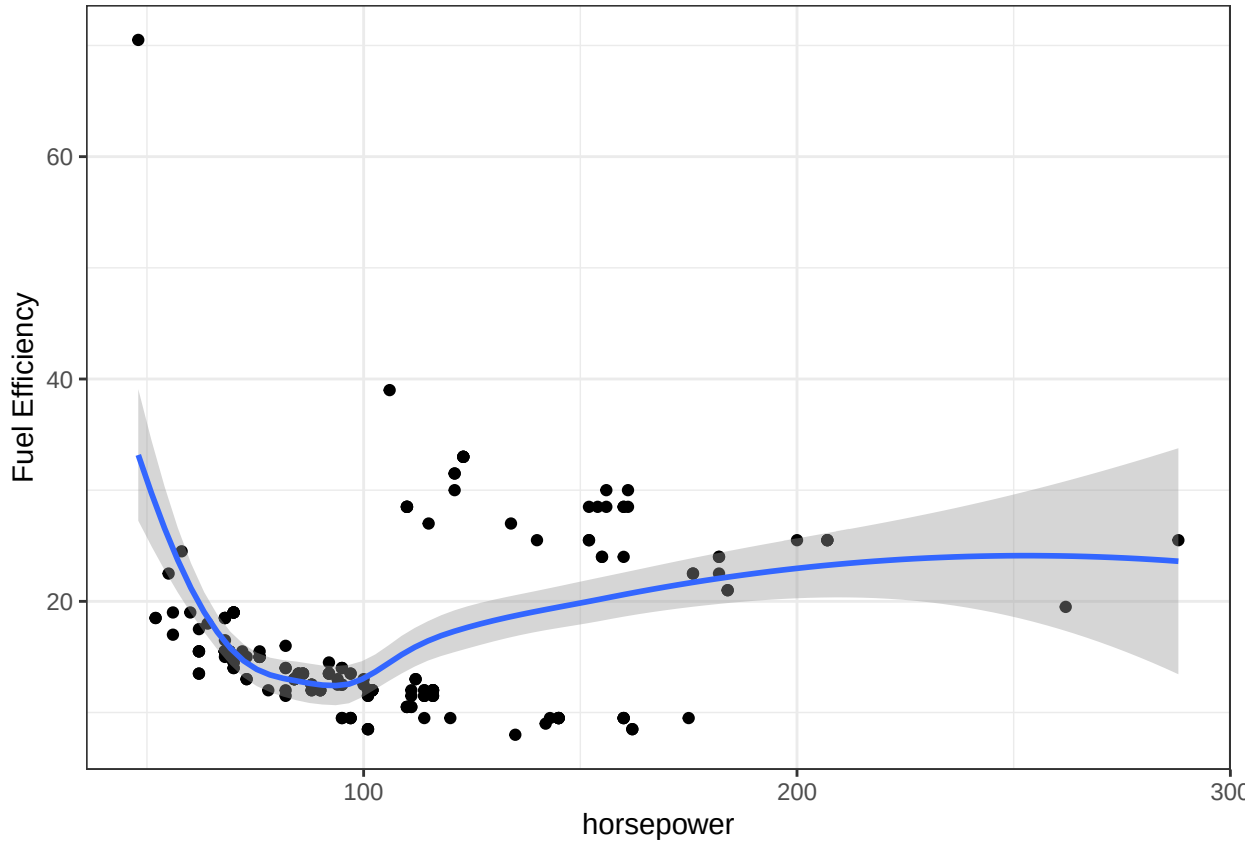
```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.392  -4.424  -1.279   3.085  56.296
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.22949    1.43295   8.534 3.43e-15 ***
## horsepower   0.04113    0.01285   3.201  0.00159 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.252 on 201 degrees of freedom
## Multiple R-squared:  0.0485, Adjusted R-squared:  0.04377
## F-statistic: 10.25 on 1 and 201 DF,  p-value: 0.001592
```

Based on the regression, we would make the mistake to conclude that there is a positive linear association between horsepower and fuel efficiency.

2. Repeat the same regression but this time, add a fixed effect for number of cylinders being "two" or "four". What would you conclude based on this new regression? What do you think drives the results in part 1?

```
ggplot(imports, aes(x = horsepower, y = city.mpg)) + geom_point() + ylab("Fuel Efficiency") +
  geom_smooth() + theme_bw()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



```
imports$fol<-ifelse(imports$num.of.cylinders=="four"|imports$num.of.cylinders=="two",1,0)
fit_fe<-felm(city.mpg~horsepower|fol,data=imports) # runs the fix effect linear regression
summary(fit_fe)
```

```
##
## Call:
##    felm(formula = city.mpg ~ horsepower | fol, data = imports)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2857 -1.1657 -0.2762  1.0648 31.4505
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## horsepower -0.100952   0.007334  -13.77   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.132 on 200 degrees of freedom
## Multiple R-squared(full model): 0.8234   Adjusted R-squared: 0.8217
## Multiple R-squared(proj model): 0.4865   Adjusted R-squared: 0.4814
## F-statistic(full model):466.3 on 2 and 200 DF, p-value: < 2.2e-16
## F-statistic(proj model): 189.5 on 1 and 200 DF, p-value: < 2.2e-16
```

From this regression, we see that ther eis a negative beta or association betwen horsepower and fiel efficiency. The results in part 1 doesn't take into account of the clusters of cylinders. THe graph confirms this conclusion.

3. (Within transformation) Now obtain the mean city.mpg and horsepower for each group. Use these group means to demean horsepower and city.mpg. Run the same regression you ran in part 1. Are the results different? Are the results obtained here different from the results in part 2? What does this tell you about the relation between fixed effect regressions and within transformations?

```
#compute means for each of the two groups
mean_by_group_imports<-aggregate(imports[, 22:24], list(imports$fol), mean)

#add the means to the main dataset
imports$hp_mean<-ifelse(imports$fol==1,mean_by_group_imports$horsepower[2],
                        mean_by_group_imports$horsepower[1])

imports$city.mpg_mean<-ifelse(imports$fol==1, mean_by_group_imports$city.mpg[2],
                              mean_by_group_imports$city.mpg[1])

#demean both the dependent and the independent variables
imports$hp_within<-imports$horsepower-imports$hp_mean
imports$city.mpg_within<-imports$city.mpg-imports$city.mpg_mean

fit_within<-felm(city.mpg_within~hp_within,data=imports)
summary(fit_within)
```

```
##
## Call:
##    felm(formula = city.mpg_within ~ hp_within, data = imports)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -5.2857 -1.1657 -0.2762  1.0648 31.4505
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.028e-15  2.193e-01     0.0        1
## hp_within   -1.010e-01  7.316e-03   -13.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.124 on 201 degrees of freedom
## Multiple R-squared(full model): 0.4865   Adjusted R-squared: 0.4839
## Multiple R-squared(proj model): 0.4865   Adjusted R-squared: 0.4839
## F-statistic(full model):190.4 on 1 and 201 DF, p-value: < 2.2e-16
## F-statistic(proj model): 190.4 on 1 and 201 DF, p-value: < 2.2e-16
```

The results produced here are similar to part 2. Fixed effects regressions inituitively is a transformation.

Question 2: On marginal significance and trading strategy improvements

You come up with a signal of stock outperformance: log total asset growth. You realize that your professor has conveniently already coded up this variable for you in the dataset StockRetAcct_insample.dta. The variable is called "lnInv".

1. Using the Fama-MacBeth regression approach, what are the average return, standard deviation and Sharpe ratio of the trading strategy implied by using only an intercept and lnInv on the right hand side in the regressions?

```
# Import data and set as data.table
StockRetAcct_DT = as.data.table(read.dta("StockRetAcct_insample.dta"))

# create excess returns in levels
StockRetAcct_DT[, ExRet := exp(lnAnnRet) - exp(lnRf)]

# Hindsight is 20/20, flip the sign on lnInv to get positive returns
StockRetAcct_DT[, neg_lnInv := -1*lnInv]

# Fama-MacBeth Regressions
port_ret = StockRetAcct_DT[, .(lambda = felm(ExRet ~ neg_lnInv, na.action = na.omit)$coef[2]),by = year]
#IMPORTANT: We need to order lambdas by year for part 4, as mentioned during the lecture.
port_ret<-port_ret[order(year)]

fm_output = list(MeanReturn = mean(port_ret$lambda), StdReturn = sqrt(var(port_ret$lambda)),
                 SR_Return = mean(port_ret$lambda)/sqrt(var(port_ret$lambda)),
                 tstat_MeanRet = sqrt(1 + 2014 - 1980) * mean(port_ret$lambda)/sqrt(var(port_ret$lambda
fm_output
```

```
## $MeanReturn
## [1] 0.08679146
##
## $StdReturn
## [1] 0.1486441
##
## $SR_Return
## [1] 0.5838877
```

```
##
## $tstat_MeanRet
## [1] 3.454326
```

2. What is the analytical expression for the portfolio weights in this case? (I'm looking for a formula)

$w_{i,t-1} = -\frac{1}{N_t}\frac{ln(Inv_{i,t-1} - E_i[Inv_{i,t-1}]}{var_i(ln(Inv_{i,t-1}))}$

3. You worry that there is industry-related noise associated with the characteristic lnInv and want to clean up your trading strategy with the goal of reducing exposure to unpriced industry risks. What regressions to you run? Report mean, standard deviation, and Sharpe ratio of the 'cleaned-up' trading strategy.

```
# Fama-MacBeth Regressions
port_ret_ind_FE = StockRetAcct_DT[, .(lambda = felm(ExRet ~ neg_lnInv | ff_ind | 0 | 0 ,
                                                    na.action = na.omit)$coef[1]), by = year]

#IMPORTANT: We need to order lambdas by year for part 4, as mentioned during the lecture.
port_ret_ind_FE<-port_ret_ind_FE[order(year)]
fm_output_ind_FE = list(MeanReturn = mean(port_ret_ind_FE$lambda),
                        StdReturn = sqrt(var(port_ret_ind_FE$lambda)),
                        SR_Return = mean(port_ret_ind_FE$lambda)/sqrt(var(port_ret_ind_FE$lambda)),

tstat_MeanRet = sqrt(1+2014-1980)*mean(port_ret_ind_FE$lambda) /
  sqrt(var(port_ret_ind_FE$lambda)))
fm_output_ind_FE
```

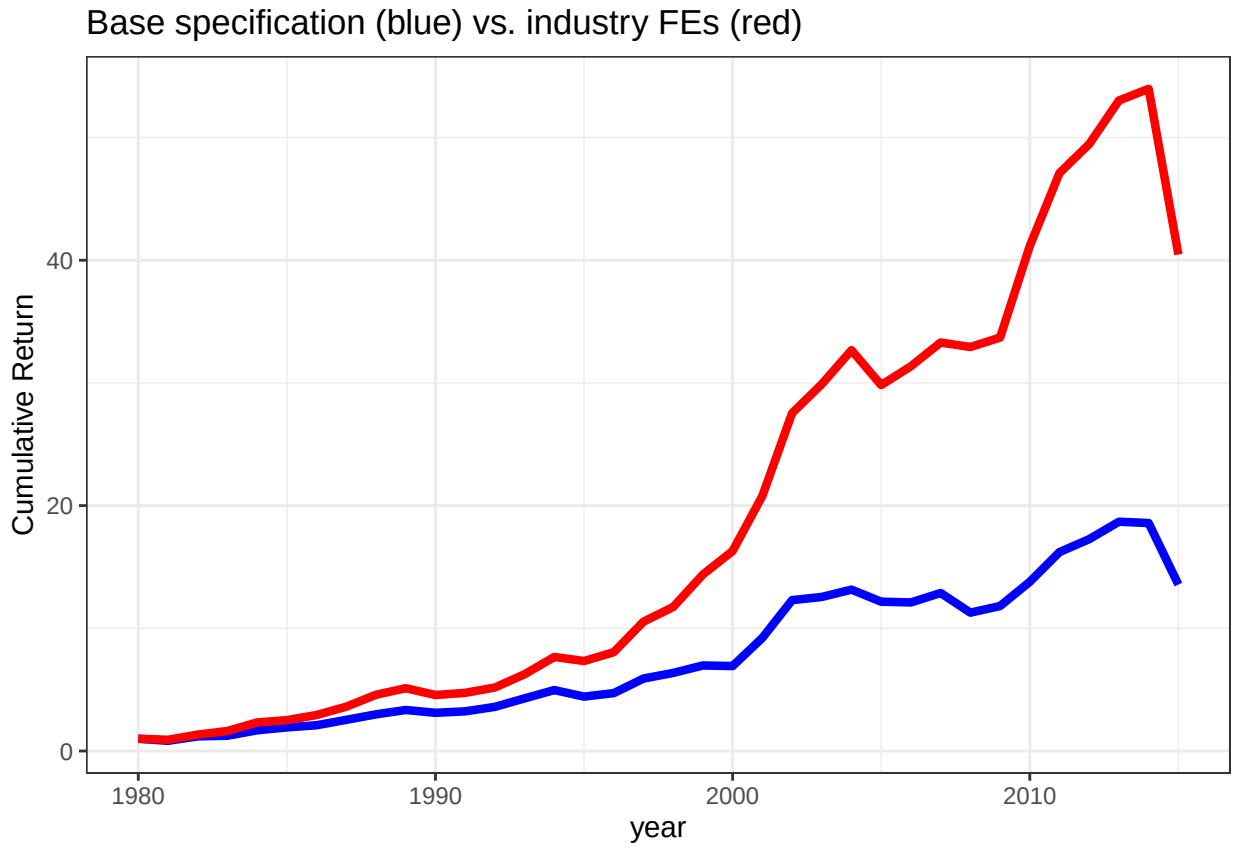
```
## $MeanReturn
## [1] 0.08257762
##
## $StdReturn
## [1] 0.1019642
##
## $SR_Return
## [1] 0.8098685
##
## $tstat_MeanRet
## [1] 4.791247
```

4. As in the class notes, plot the cumulative returns to the simple and the 'cleaned-up' trading strategies based on your new signal, lnInv. Make sure both trading strategies result in portfolios with a 15% return standard deviation.

```
#Note that this graph is slightly different from the one you saw during the lecture.
#This is the final version of the proposed solution.
# For both strategies

for (i in c('', '_ind_FE')) {
  # Scale returns to have a 15% standard deviation
  assign(paste0('scaled_ret', i), get(paste0('port_ret', i))$lambda*0.15/sqrt(var(get(paste0('port_ret'
  assign(paste0('Cum_ret', i), 0)
  for (j in 1:35) { # Calculate cumulative returns
    assign(paste0('Cum_ret', i), c(get(paste0('Cum_ret', i)),
                                   get(paste0('Cum_ret', i))[j] + log(1 + get(paste0('scaled_ret', i))[
  }
}
```

```r
# Plot cumulative returns
qplot(c(1980:2015), exp(Cum_ret), geom = "line", xlab = "year",
      ylab = "Cumulative Return", color = I("blue"), size = I(1.5),
      main = "Base specification (blue) vs. industry FEs (red)") +
  geom_line(aes(y = exp(Cum_ret_ind_FE)), color = I("red"), size = I(1.5)) + theme_bw()
```



Question 3: Predicting medium to long-run firm-level return variance

There are many return volatility models, such as GARCH. These work best at shorter horizons. As an alternative, we will explore a panel regression approach to predicting firmlevel return variance. The data set StockRetAcct_insample.dta has annual realized variance (rv), calculated as the sum of squared daily returns to each firm, each year. Run panel forecasting regressions to forecast firm-level one-year ahead rv along the lines of what we did with lnROE in class.

1. Try with and without industry and year fixed effects, with and without clustering of standard errors. Discuss which specification makes most sense to you. In particular, discuss the effect of a year fixed effect. What is the intuition for the impact of this fixed effect?

```r
# Define next year's rv. Make sure next observation is actually next year.
setorder(StockRetAcct_DT, FirmID, year)
StockRetAcct_DT[, next_rv := shift(rv, type = 'lead'), by = FirmID]
StockRetAcct_DT[, next_year := shift(year, type = 'lead'), by = FirmID]
StockRetAcct_DT[next_year != (year + 1), next_rv := NA]
```

```r
# Define previous year's return
StockRetAcct_DT[, prev_ExRet := shift(ExRet), by = FirmID]
StockRetAcct_DT[, prev_year := shift(year), by = FirmID]
StockRetAcct_DT[prev_year != (year - 1), prev_ExRet := NA]

# Run some panel regressions
r1 = felm(next_rv ~ rv + lnBM + lnProf + lnLever + lnIssue +
            lnInv + prev_ExRet, data = StockRetAcct_DT, na.action = na.omit)

r2 = felm(next_rv ~ rv + lnBM + lnProf + lnLever + lnIssue + lnInv +
            prev_ExRet | 0 | 0 | year + FirmID, data = StockRetAcct_DT, na.action = na.omit)
r3 = felm(next_rv ~ rv + lnBM + lnProf + lnLever + lnIssue + lnInv +
            prev_ExRet | year, data = StockRetAcct_DT, na.action = na.omit)
r4 = felm(next_rv ~ rv + lnBM + lnProf + lnLever + lnIssue + lnInv +
            prev_ExRet | year | 0 | year + FirmID, data = StockRetAcct_DT, na.action = na.omit)
r5 = felm(next_rv ~ rv + lnBM + lnProf + lnLever + lnIssue + lnInv +
            prev_ExRet | year + ff_ind, data = StockRetAcct_DT, na.action = na.omit)
r6 = felm(next_rv ~ rv + lnBM + lnProf + lnLever + lnIssue + lnInv +
            prev_ExRet | year + ff_ind | 0 | year + FirmID,
          data = StockRetAcct_DT, na.action = na.omit)
stargazer(r1, r2, r3, r4, r5, r6, type = 'text', report = 'vc*t',
          add.lines = list(c('Year FE', 'N', 'N', 'Y', 'Y', 'Y', 'Y'),
                           c('Ind FE', 'N', 'N', 'N', 'N', 'Y', 'Y'),
                           c('Firm, Year Clustering', 'N', 'Y', 'N', 'Y', 'N', 'Y')),
          omit.stat = 'ser')
```

```
##
## ================================================================================================
##                                              Dependent variable:
##                      ------------------------------------------------------------------------
##                                                      next_rv
##                          (1)          (2)          (3)          (4)          (5)          (6)
## ------------------------------------------------------------------------------------------------
## rv                     0.426***     0.426**      0.554***     0.554***     0.516***     0.516***
##                       t = 103.445 t = 2.545   t = 146.543 t = 4.009   t = 132.257 t = 3.784
##
## lnBM                   -0.024***    -0.024***    -0.009***     -0.009      -0.004***     -0.004
##                       t = -25.707 t = -3.219 t = -13.553 t = -1.380 t = -5.270   t = -0.786
##
## lnProf                 -0.055***    -0.055***    -0.036***    -0.036***    -0.039***    -0.039***
##                       t = -17.379 t = -4.322 t = -16.184 t = -5.236 t = -17.280 t = -6.327
##
## lnLever                -0.006***      -0.006     -0.005***      -0.005       0.0003       0.0003
##                       t = -6.045   t = -0.944 t = -8.080   t = -0.816   t = 0.337   t = 0.076
##
## lnIssue                0.030***       0.030*     0.031***     0.031***     0.034***     0.034***
##                        t = 9.525    t = 1.745   t = 14.105   t = 3.125   t = 15.055   t = 3.202
##
## lnInv                  0.069***     0.069***     0.034***     0.034**      0.032***     0.032**
##                       t = 22.351   t = 3.026   t = 15.508   t = 2.590   t = 15.063   t = 2.713
##
## prev_ExRet              -0.0004      -0.0004      0.019***       0.019       0.021***       0.021
##                       t = -0.260   t = -0.016 t = 16.829   t = 1.399   t = 19.018   t = 1.589
##
```

```
## Constant               0.073***    0.073***
##                       t = 47.373  t = 3.352
##
## ---------------------------------------------------------------------------------------
## Year FE                    N          N          Y          Y          Y          Y
## Ind FE                     N          N          N          N          Y          Y
## Firm, Year Clustering      N          Y          N          Y          N          Y
## Observations            46,685     46,685     46,685     46,685     46,685     46,685
## R2                       0.268      0.268      0.653      0.653      0.661      0.661
## Adjusted R2              0.268      0.268      0.652      0.652      0.660      0.660
## =======================================================================================
## Note:                                                        *p<0.1; **p<0.05; ***p<0.01
```

I test four specifications - with and without industry and year fixed effects, and with and without firm and year clustering. We can see that clustering significantly decreases the t-statistic of all coefficients, and that fixed effects increase the predictive power of rv. This likely is due to the year fixed effects accounting for changes in systematic risk driving differences in observed realized variance across time.

2. Also try forecasting at the 5-year horizon (rv in 5 years). How do the results change? Can we predict return variance 5-years ahead? Is the 5-year lagged rv significant, or are other variables more important?

```
# Define five year forward rv
five_year_rv = copy(StockRetAcct_DT[, .(FirmID, year, rv)])
setorder(five_year_rv, FirmID, year)
five_year_rv[, year := year - 5]
setnames(five_year_rv, 'rv', 'next_5_rv')
StockRetAcct_DT = merge(StockRetAcct_DT, five_year_rv, by = c('FirmID', 'year'), all.x = T)
setkey(StockRetAcct_DT)

# Run some panel regressions
r1 = felm(next_5_rv ~ rv + lnBM + lnProf + lnLever + lnIssue + lnInv +
            prev_ExRet, data = StockRetAcct_DT, na.action = na.omit)
r2 = felm(next_5_rv ~ rv + lnBM + lnProf + lnLever + lnIssue + lnInv +
            prev_ExRet | 0 | 0 | year + FirmID, data = StockRetAcct_DT, na.action = na.omit)
r3 = felm(next_5_rv ~ rv + lnBM + lnProf + lnLever + lnIssue + lnInv +
            prev_ExRet | year, data = StockRetAcct_DT, na.action = na.omit)
r4 = felm(next_5_rv ~ rv + lnBM + lnProf + lnLever + lnIssue + lnInv +
            prev_ExRet | year | 0 | year + FirmID, data = StockRetAcct_DT, na.action = na.omit)
r5 = felm(next_5_rv ~ rv + lnBM + lnProf + lnLever + lnIssue + lnInv +
            prev_ExRet | year + ff_ind, data = StockRetAcct_DT, na.action = na.omit)

r6 = felm(next_5_rv ~ rv + lnBM + lnProf + lnLever + lnIssue + lnInv +
            prev_ExRet | year + ff_ind | 0 | year + FirmID,
          data = StockRetAcct_DT, na.action = na.omit)
stargazer(r1, r2, r3, r4, r5, r6, type = 'text', report = 'vc*t',
          add.lines = list(c('Year FE', 'N', 'N', 'Y', 'Y', 'Y', 'Y'),
                           c('Ind FE', 'N', 'N', 'N', 'N', 'Y', 'Y'),
                           c('Firm, Year Clustering', 'N', 'Y', 'N', 'Y', 'N', 'Y')),
          omit.stat = 'ser')
```

```
##
## =======================================================================================
##                                          Dependent variable:
##                      ---------------------------------------------------------------
##                                                  next_5_rv
##                          (1)        (2)         (3)        (4)         (5)        (6)
```

```
## ------------------------------------------------------------------------------------------
## rv                        -0.001      -0.001    0.217***    0.217**    0.156***    0.156**
##                         t = -0.273  t = -0.031 t = 41.127  t = 2.698  t = 28.931  t = 2.431
##
## lnBM                     -0.025***  -0.025***   -0.012***   -0.012*    -0.002**     -0.002
##                         t = -19.639 t = -3.270 t = -12.045 t = -1.797 t = -2.083  t = -0.456
##
## lnProf                   -0.079***  -0.079***   -0.036***  -0.036***   -0.039***  -0.039***
##                         t = -16.501 t = -2.898 t = -10.048 t = -3.158 t = -10.934 t = -3.600
##
## lnLever                  -0.007***    -0.007    -0.003***    -0.003     0.002**     0.002
##                         t = -5.339  t = -0.849 t = -2.886  t = -0.342  t = 2.015  t = 0.522
##
## lnIssue                   0.053***    0.053***   0.024***    0.024***   0.029***    0.029***
##                         t = 12.114  t = 4.141   t = 7.510  t = 3.334   t = 9.184  t = 4.646
##
## lnInv                     0.035***    0.035**    0.034***    0.034***   0.030***    0.030***
##                          t = 8.045  t = 2.183  t = 10.488  t = 2.901   t = 9.598  t = 3.143
##
## prev_ExRet                0.033***     0.033*     0.004**     0.004     0.007***     0.007
##                         t = 16.488  t = 1.711   t = 2.257  t = 0.425   t = 4.720  t = 0.976
##
## Constant                  0.133***    0.133***
##                         t = 64.125  t = 6.491
##
## ------------------------------------------------------------------------------------------
## Year FE                      N           N           Y           Y           Y           Y
## Ind FE                       N           N           N           N           Y           Y
## Firm, Year Clustering        N           Y           N           Y           N           Y
## Observations            31,220      31,220      31,220      31,220      31,220      31,220
## R2                       0.057       0.057       0.487       0.487       0.511       0.511
## Adjusted R2              0.057       0.057       0.487       0.487       0.511       0.511
## ==========================================================================================
## Note:                                                          *p<0.1; **p<0.05; ***p<0.01
```

Regression coefficients are less significant when predicting realized variance at a five-year horizon. With firm and year clustering, and industry and year fixed effects, the t-stat of rv is only 2.429. While other variables have higher explanatory power, they are harder to motivate (lnIssue as an example).

3. What are the benefits of the panel approach, versus simply running one regression for each firm? What are the potential costs?

The primary issue with running a regression for each firm, is that the time-series for each firm is not long enough to get statistically significant estimates. The panel approach allows you to estimate covariates across many firms, allowing for more observations to be used in the regression. However, the panel specification will identify the aggregate affect of the covariates, which may not be directly applicable to any specific firm.