Data Analytics Machine Learning PS2

Jiaming Huang, Hogun Kim, Yichu Li, An Yang

2019/4/14

Question 1: 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".

```
setwd("/Users/jiaminghuang/Downloads")

library(foreign)
```

```
## Warning: package 'foreign' was built under R version 3.4.4
```

```
library(data.table)
```

```
## Warning: package 'data.table' was built under R version 3.4.4
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.4
```

```
data <-  as.data.table(read.dta("StockRetAcct_insample.dta"))
data$ExRet <- exp(data$lnAnnRet)-exp(data$lnRf)
```

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?

Answer :

```
port_ret = NULL

for (i in 1980:2014) {
  temp <- data[year == i,]
  fit_yr <- lm(temp$ExRet ~ temp$lnInv, data = temp)
  temp <- coefficients(fit_yr)
  port_ret = rbind(port_ret,temp[2])
}

port_ret = -1*port_ret

fm_output = list(
  MeanReturn = mean(port_ret),
  StdReturn = sqrt(var(port_ret)),
  SR_Return = mean(port_ret)/sqrt(var(port_ret))
)

fm_output
```

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

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

$$w_{i,t-1} = \frac{1}{N} \frac{ln(Inv_{i,t-1}) - E[ln(Inv_{i,t-1})]}{var(ln(Inv_{i,t-1}))}$$

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.

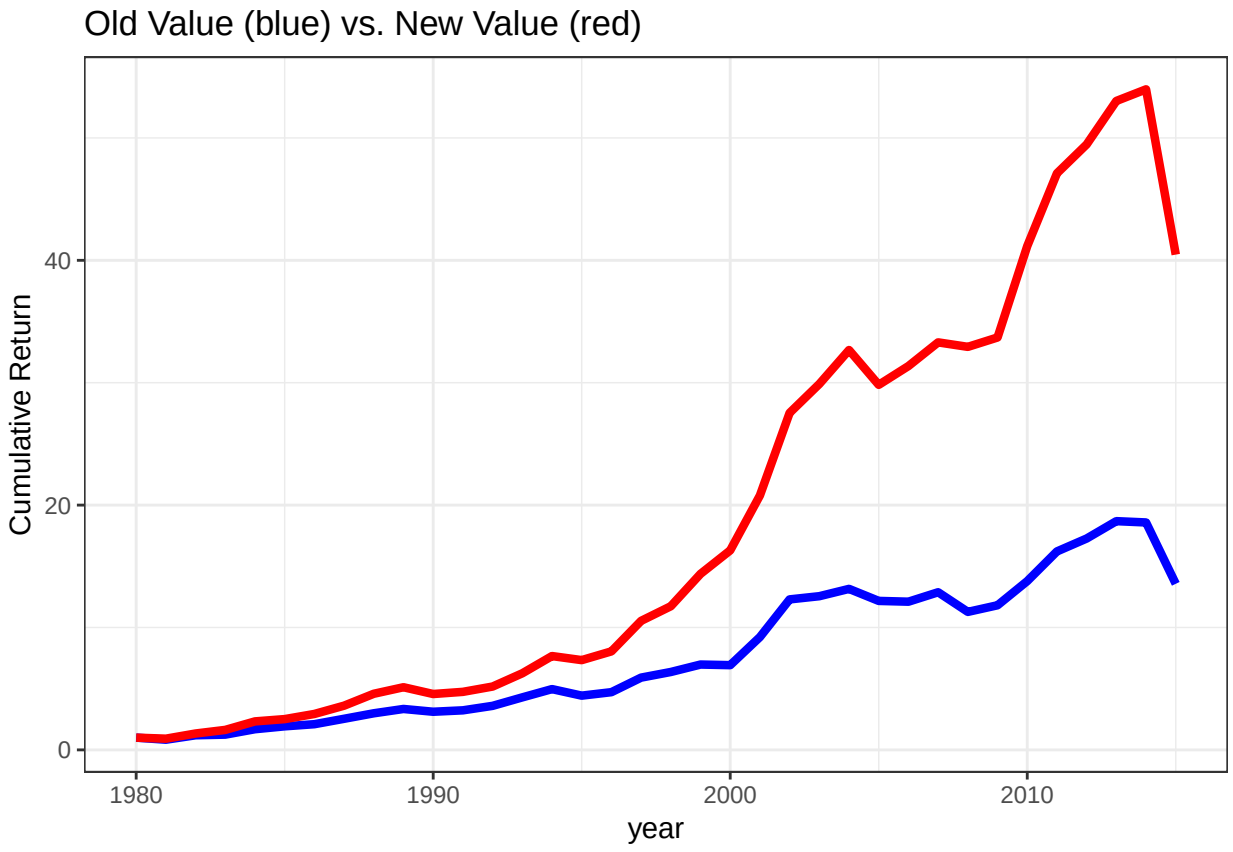
```
port_ret = NULL

for (i in 1980:2014) {
  temp <- data[year == i,]
  fit_yr <- lm(temp$ExRet ~ temp$lnInv+as.factor(temp$ff_ind), data = temp)
  temp <- coefficients(fit_yr)
  port_ret = rbind(port_ret,temp[2])
}

port_ret = -1*port_ret

fm_output = list(
  MeanReturn = mean(port_ret),
  StdReturn = sqrt(var(port_ret)),
  SR_Return = mean(port_ret)/sqrt(var(port_ret))
)

fm_output
```

```
## $MeanReturn
## [1] 0.08257762
##
## $StdReturn
##             temp$lnInv
## temp$lnInv  0.1019642
##
## $SR_Return
##             temp$lnInv
## temp$lnInv  0.8098685
```

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.

```
lnInv_ret = port_ret[,1] * 0.15 / sqrt(var(port_ret[,1]))
```

```
cum_ret_lnInv = 0

for (ii in 1:35) {
  cum_ret_lnInv = rbind(cum_ret_lnInv,cum_ret_lnInv[ii]+log(1+lnInv_ret[ii]))
}

port_ret = NULL

for (i in 1980:2014) {
  temp <- data[year == i,]
  fit_yr <- lm(temp$ExRet ~ temp$lnInv, data = temp)
  temp <- coefficients(fit_yr)
  port_ret = rbind(port_ret,temp[2])
}

port_ret = -1 * port_ret

lnInv_old_ret = port_ret[,1] * 0.15 / sqrt(var(port_ret[,1]))

cum_ret_OldlnInv = 0

for (ii in 1:35) {
  cum_ret_OldlnInv = rbind(cum_ret_OldlnInv,cum_ret_OldlnInv[ii]+log(1+lnInv_old_ret[ii]))
}

qplot(c(1980:2015), exp(cum_ret_OldlnInv), geom="line", xlab="year",ylab="Cumulative Return",color = I(
```



Question 2: 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 firm level 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?

Pls see R code for the regression and results are as below: To see what predicts next year's rv, the table below are without fixed effects and clustering of standard errors

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

```
## Warning: package 'lfe' was built under R version 3.4.4
```

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

```
library(stargazer)
```

```
## Warning: package 'stargazer' was built under R version 3.4.4
```

```
##
## 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
```

```
rm(list = ls())

# Download data and set as data.table
StockRetAcct_DT <- as.data.table(read.dta("StockRetAcct_insample.dta"))

# set keys for StockRetAcct_DT In particular, it will be useful to sort on FirmID and year
setkey(StockRetAcct_DT, FirmID, year)

# Get mean and standard deviation of rv across firms and time
mean_rv = mean(na.omit(StockRetAcct_DT$rv))
mean_rv
```

```
## [1] 0.1770095
```

```
std_rv = sqrt(var(na.omit(StockRetAcct_DT$rv)))
std_rv
```

```
## [1] 0.1893437
```

```
# create excess returns (what we really care about)
StockRetAcct_DT[,ExRet:=exp(lnAnnRet) - exp(lnRf)]

# What predicts next year's rv?
setorder(StockRetAcct_DT, FirmID, year) # Set order of data so that shift does what we want it to
```

```
StockRetAcct_DT[, lead_rv := shift(rv, type = 'lead'), by = FirmID]  # Define next year's rv
rv_panel = felm(lead_rv ~ rv, StockRetAcct_DT) # Regression with no FE or SE
stargazer(rv_panel, type = 'text', report = 'vc*t') # Output regressions as text, report t-stats
```

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                                lead_rv
## ------------------------------------------------
## rv                            0.507***
##                              t = 140.892
##
## Constant                      0.080***
##                              t = 89.991
##
## ------------------------------------------------
## Observations                   55,654
## R2                              0.263
## Adjusted R2                     0.263
## Residual Std. Error      0.152 (df = 55652)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

R square=0.263 and all t-stats are significant.

(1) Clustering of standard errors If cluster standard errors at the firm level:

```
# cluster standard errors at the firm level
rv_panel2 = felm(lead_rv ~ rv | 0 | 0 | FirmID, StockRetAcct_DT) # Regression with no FE, standard erro
stargazer(rv_panel2, type = 'text', report = 'vc*t') # Output regressions as text, report t-stats
```

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                                lead_rv
## ------------------------------------------------
## rv                            0.507***
##                              t = 66.107
##
## Constant                      0.080***
##                              t = 56.034
##
## ------------------------------------------------
## Observations                   55,654
## R2                              0.263
## Adjusted R2                     0.263
## Residual Std. Error      0.152 (df = 55652)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

R square no change, but t_stats much smaller.

If cluster standard errors at the firm level and year level, result in (3) compared with no clustering(1) and clustering at firm level:

```
# clustered standard errors at the firm and year level
rv_panel3 = felm(lead_rv ~ rv | 0 | 0 | year + FirmID, StockRetAcct_DT) # Regression with no FE, standa
stargazer(rv_panel,rv_panel2, rv_panel3, type = 'text', report = 'vc*t') # Output regressions as text,
```

```
##
## ===============================================================
##                                       Dependent variable:
##                                   -----------------------------
##                                              lead_rv
##                                      (1)        (2)        (3)
## ---------------------------------------------------------------
## rv                                 0.507***    0.507***  0.507***
##                                   t = 140.892 t = 66.107 t = 2.875
##
## Constant                           0.080***    0.080***  0.080***
##                                   t = 89.991  t = 56.034 t = 3.277
##
## ---------------------------------------------------------------
## Observations                        55,654      55,654    55,654
## R2                                   0.263      0.263      0.263
## Adjusted R2                          0.263      0.263      0.263
## Residual Std. Error (df = 55652)     0.152      0.152      0.152
## ===============================================================
## Note:                                  *p<0.1; **p<0.05; ***p<0.01
```

R square no change, but t_stats are much smaller. We have evidence of both firm and time dependences in the errors.

(2) Fixed effects If industry component in accounting variables, include industry fixed effects:

```
# industry component in accounting variables, include industry fixed effects
rv_panel4 = felm(lead_rv ~ rv | ff_ind | 0 | year + FirmID, StockRetAcct_DT) # Regression with FE, stan
stargazer(rv_panel4, type = 'text', report = 'vc*t') # Output regressions as text, report t-stats
```

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                                lead_rv
## ------------------------------------------------
## rv                            0.461***
##                               t = 2.651
##
## ------------------------------------------------
## Observations                   55,654
## R2                              0.284
## Adjusted R2                     0.284
## Residual Std. Error     0.150 (df = 55641)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

R square increase and t-stats are smaller.

If industry component in accounting variables, include industry and year fixed effects(2):

```
# industry component in accounting variables, include industry and year fixed effects
rv_panel5 = felm(lead_rv ~ rv| year + ff_ind | 0 | year + FirmID, StockRetAcct_DT)
```

```
stargazer(rv_panel4, rv_panel5, type = 'text', report = 'vc*t') # Output regressions as text, report t-
```

```
##
## ============================================================
##                               Dependent variable:
##                     -----------------------------------
##                                      lead_rv
##                             (1)                (2)
## ------------------------------------------------------------
## rv                        0.461***           0.575***
##                          t = 2.651          t = 4.320
##
## ------------------------------------------------------------
## Observations               55,654             55,654
## R2                         0.284              0.622
## Adjusted R2                0.284              0.622
## Residual Std. Error 0.150 (df = 55641) 0.109 (df = 55608)
## ============================================================
## Note:                           *p<0.1; **p<0.05; ***p<0.01
```

R square increase dramatically, it is a big effect.

Since annual realized variance (rv) has large expose to time (year), so by adding the year fixed effect, we can eliminate the risk and have a better fit in model.

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?

If clustered standard errors at the firm and year level, industry and year effects: big model

```
# clustered standard errors at the firm and year level, industry effects: big model
rv_panel6 = felm(lead_rv ~ rv+lnROE + lnBM +lnProf + lnLever + lnIssue + lnInv | year + ff_ind | 0 | ye
stargazer(rv_panel6, type = 'text', report = 'vc*t') # Output regressions as text, report t-stats
```

```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                 lead_rv
## ------------------------------------------------
## rv                             0.520***
##                                t = 3.957
##
## lnROE                           -0.011
##                               t = -1.463
##
## lnBM                            -0.012*
##                               t = -1.957
##
## lnProf                          -0.039***
##                               t = -6.613
##
## lnLever                         -0.003
##                               t = -0.690
##
## lnIssue                         0.036***
```

```
##                                t = 3.786
##
## lnInv                           0.027**
##                                t = 2.520
##
## ------------------------------------------------
## Observations                    51,707
## R2                               0.627
## Adjusted R2                      0.627
## Residual Std. Error     0.104 (df = 51655)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

R square increases a little, except for rv, InProf and InIssue are significant at 0.01 level of significance.

Predicts rv in five years, compare results with dependent variable as lead5_rv (2):

```
#What predicts rv in five  years?
StockRetAcct_DT[, lead2_rv := shift(lead_rv, type = 'lead'), by = FirmID]  # Define next year's rv
StockRetAcct_DT[, lead3_rv := shift(lead2_rv, type = 'lead'), by = FirmID]  # Define next year's rv
StockRetAcct_DT[, lead4_rv := shift(lead3_rv, type = 'lead'), by = FirmID]  # Define next year's rv
StockRetAcct_DT[, lead5_rv := shift(lead4_rv, type = 'lead'), by = FirmID]  # Define next year's rv
roe_panel7 = felm(lead5_rv ~ rv+lnROE + lnBM +lnProf + lnLever + lnIssue + lnInv | year+ff_ind | 0 | ye
stargazer(rv_panel6, roe_panel7, type = 'text', report = 'vc*t') # Output regressions as text, report t
```

```
##
## ============================================================
##                                Dependent variable:
##                           ------------------------------------
##                                lead_rv            lead5_rv
##                                  (1)                (2)
## ------------------------------------------------------------
## rv                             0.520***           0.167***
##                               t = 3.957          t = 2.651
##
## lnROE                           -0.011             -0.013
##                               t = -1.463         t = -1.480
##
## lnBM                            -0.012*            -0.006
##                               t = -1.957         t = -1.208
##
## lnProf                        -0.039***          -0.028***
##                               t = -6.613         t = -4.517
##
## lnLever                         -0.003             0.0003
##                               t = -0.690         t = 0.077
##
## lnIssue                         0.036***           0.032***
##                               t = 3.786          t = 4.459
##
## lnInv                           0.027**            0.029***
##                               t = 2.520          t = 3.174
##
## ------------------------------------------------------------
## Observations                    51,707             34,011
## R2                              0.627              0.418
```

```
## Adjusted R2               0.627              0.417
## Residual Std. Error 0.104 (df = 51655) 0.119 (df = 33963)
## ==========================================================
## Note:                           *p<0.1; **p<0.05; ***p<0.01
```

R square becomes smaller as we expected.

The lagged 5 year rv is still significant and so are the InProf and InIssue. InInv has became more significant.

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

Benefits of panel regression vs running regression one regression for each firm:

a. Having more data to run, Impossible to run individual model at the firm level given only 10 year median firm survival in data.

b. get access to both time series and cross sectional data Cost of panel regression: impact of unobserved heterogeneity