UCLA Anderson
SCHOOL of MANAGEMENT

MGMTMFE 431:

Data Analytics and Machine Learning

Topic 2: Panel regressions

Spring 2020

Professor Lars A. Lochstoer

Last class: Visualization

1. **Reduce noise**
   - E.g., portfolio sorts, as long as possible samples. Realized returns very noisy, never looks good to plot realized versus predicted at the stock level and over short samples

2. **Reduce dimensionality**
   - Plots should typically be 2-D. 3-D almost never looks good
   - Can use multiple lines, multiple panels (facets) to show three dimensional relations
   - Perhaps add different colors, line-types for fourth and fifth dimensions -- this is pushing it.

As data scientists, our job to (**1**) understand which (new) dimensions are important, (**2**) come up with an economic rationale for why something works (a "story"), and (**3**) to show clearly and convincingly (with plots) what is going on in the data.

- Part of "***convincing***" is to show plots of ***implementable strategies*** likely to ***survive transaction costs***
  - E.g., ***value-weighting within portfolio***, calculate approx. ***transaction costs***, ***sorts based on information available at time t*** when considering returns from *t* to *t+1*

Last class: Visualization

Code from last time showed how to:

1. Create quantile sorts (e.g., decile or vingtile) at each *t*
2. Value-weight returns within each portfolio each period from *t* to *t+1*
3. Then take (equal-weighted) average across all years (weight time equally)

Also, showed how to do conditional sorts

- E.g., sort conditional on a size quantile (in this case, just small and large)

You can copy this procedure for ***your signal***

- Interact with size, industry, etc.
- Control for other, known trading strategies
- Today, we will do the latter – i.e., assess ***marginal significance, value added***
- *But, first, let's discuss problem set 1…*

Topic 2: Panel regressions

a. Panel regression overview

b. Multiple Regression vs. Simple Regression

c. Omitted Variables

d. Panel regressions in detail

   1) Predicting firm earnings

   2) Clustered standard errors

   3) Fixed Effects

   4) Predicting firm-level return variance

a. Panel regression overview

The simplest setting for *"**big data**"* type analysis are panel regressions.

- Panel: Typically cross-section ***and*** time series, size N x T
  - **Balanced panel:** There are N observations in cross-section for each *t*
  - **Unbalanced panel**: For each *t* only a subset of the cross-section have data (N(t) < N)

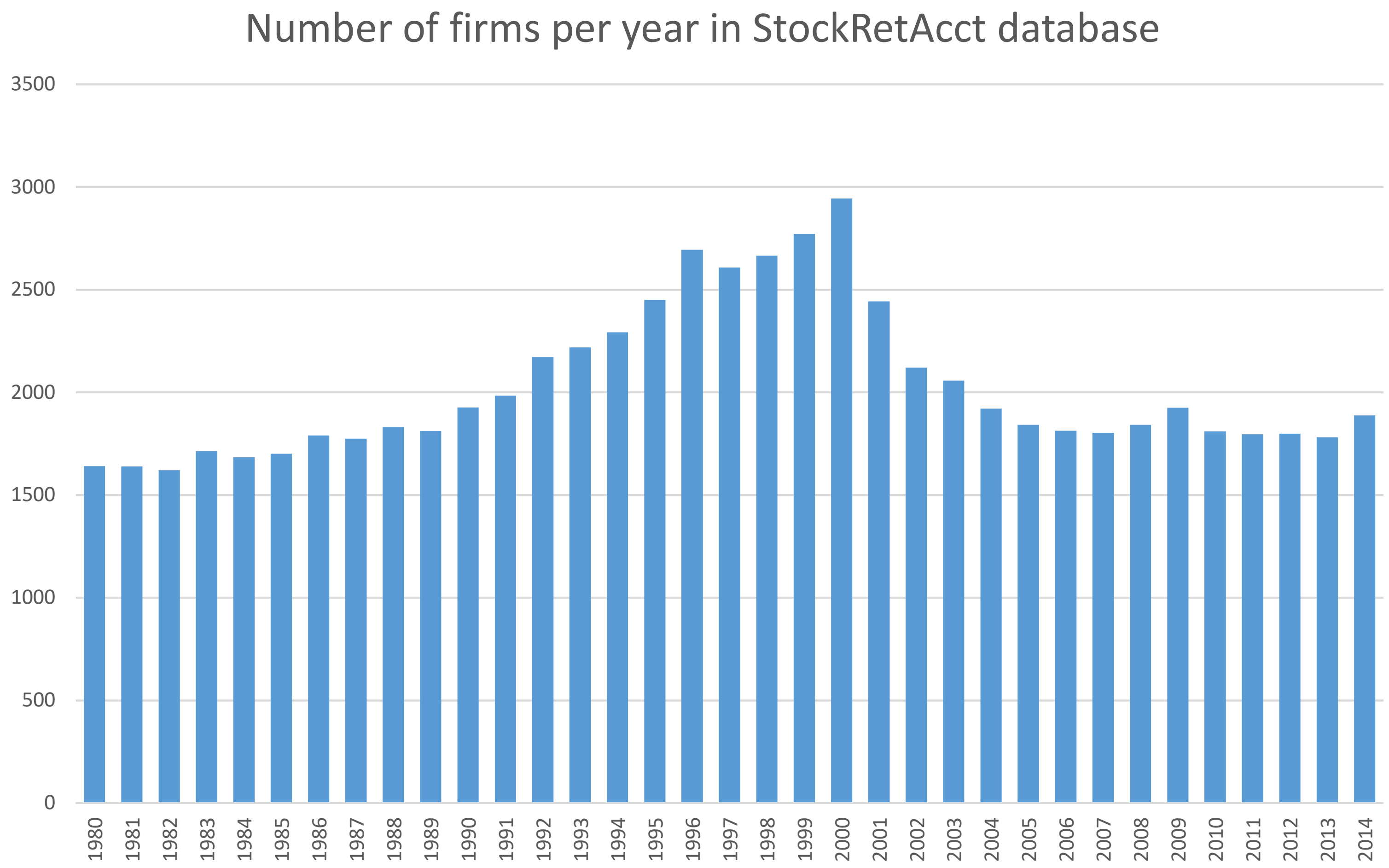
The dataset StockRetAcct_insample.dta that we have been working with is an example of an unbalanced panel

- Total set of firms, N, are 20,314.
  - At each time *t*, there are way fewer firms "alive"
  - See plot on next slide
- 35 years of observations, so T = 35

While *Machine Learning* often refers to nonlinear techniques to analyze data, ***your first analysis will (and should!) typically be linear regressions***

- Robust, easy-to-understand and -communicate

a. Cross-section of firms is an unbalanced panel



a. Panel regression overview

Panel uses variation both in cross-section and over time to identify regression coefficients. That’s BIG DATA!

- ***Implicit assumption: slope coefficients do not vary over time or across firms***
- However, intercept *may* be allowed to vary:
  - Over time: ***Time fixed effects***
  - Across firms: ***Firm fixed effects***

Canonical panel regression:

- delta’s are *time fixed effects*, theta’s are *firm fixed effects*. No *i* subscript on beta
- Note that the error term may be correlated across firms and time, so standard errors typically cannot be found using the classic OLS standard error formula that presumes iid error terms.

$$y_{i,t} = \delta_t + \theta_i + \beta' X_{i,t} + \varepsilon_{i,t}$$

a. Panel regression: increasing power

A simple example of how panel regressions can increase power

- Assume all stocks have the same expected return and return variance = σ^2
- Assume all pairwise ***correlations*** equal ρ ***across stocks***, ***zero over time***
- Simplest model: X's is a vector of ones, i.e., just estimating intercept

$$R_{i,t} = \mu + \varepsilon_{i,t}$$

The estimate is simply $\hat{\mu} = \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N} R_{i,t}$

The variance of this estimate is

$$\mathrm{var}(\hat{\mu}) = E\left[\left(\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N} R_{i,t} - \mu\right)^2\right] = \frac{\sigma^2}{NT}(1 + (N-1)\rho)$$

Notice how if stocks are not perfectly correlated, the variance of the estimate is lower than if we estimated each stock's mean return separately (which has a variance of estimate of σ^2/T)

a. Panel regression: simplest example

- Consider the below simplest panel regression (one regressor plus intercept)

$$y_{i,t} = \alpha + \beta x_{i,t} + \varepsilon_{i,t}$$

- Thus, there are no firm, industry, or year fixed effects.
- Assume a balanced panel where i = 1, …, N; t = 1, …, T.
- How do you estimate this (other than using the *R* routine)?

a. Panel regression: simplest example

- Define the following matrices:

$$\underbrace{X}_{TN\times 2} = \begin{bmatrix} 1 & x_{1,1} \\ 1 & x_{1,2} \\ 1 & \vdots \\ 1 & x_{1,T} \\ 1 & x_{2,1} \\ 1 & \vdots \\ 1 & x_{2,T} \\ 1 & \vdots \\ 1 & x_{N,1} \\ 1 & x_{N,2} \\ 1 & \vdots \\ 1 & x_{N,T} \end{bmatrix} \qquad \underbrace{Y}_{TN\times 1} = \begin{bmatrix} y_{1,1} \\ y_{1,2} \\ \vdots \\ y_{1,T} \\ y_{2,1} \\ \vdots \\ y_{2,T} \\ \vdots \\ y_{N,1} \\ y_{N,2} \\ \vdots \\ y_{N,T} \end{bmatrix}$$

a. Panel regression: simplest example

- Then, we find the regression coefficients as:

$$\begin{bmatrix}\alpha \\ \beta\end{bmatrix} = (X'X)^{-1}X'Y$$

- Note, however, that ***assuming errors are i.i.d. as in standard OLS is not a good idea here. Almost surely, firms' residuals are cross-sectionally correlated*** (e.g., a shock to the state of the economy affects all firms' residuals).
- There could also be autocorrelation over time for each firm
- The variance-covariance matrix of the residuals is huge however:
  - TN x TN
  - In order to decrease the size of the covariance matrix (in terms of estimating it), we typically apply ***clustering*** (see next slide)

a. Clustering

- Assume that firms' shocks are correlated within each year but not across years.
- Assume also that the cross-firm covariance is constant over time.
  - That is: $\mathrm{Cov}(\varepsilon_{i,t}, \varepsilon_{j,t+k}) = \sigma_{ij}$ for all *t* if *k = 0* but zero if *k ≠ 0*.

- Clustering standard errors by ***time*** (e.g., year) imposes these assumptions
  - Note that we now have "only" N(N+1)/2 free coefficients in the variance-covariance matrix
  - However, while still a lot coefficients, we are looking for specific averages that makes up the covariance matrix of the two estimated coefficients (which is only a 2 by 2 matrix), so estimation error in the still big covariance matrix of residuals washes out to a large extent. (See paper by Mitchell Petersen on CCLE if interested in further details).

- Now, assume firm residuals also can be autocorrelated, though only within-firm, not across firms.
  - That is: $\mathrm{Cov}(\varepsilon_{i,t}, \varepsilon_{i,t+k}) = \sigma_{ik}$ for all *t* and $\mathrm{Cov}(\varepsilon_{i,t}, \varepsilon_{j,t+k}) = 0$ for i ≠ j and k ≠ 0.
  - Adding clustering by ***firm*** achieves this.

- Allowing for heteroskedasticity in addition to clustering can be achieved using so-called Rogers standard errors.

a. Fixed effects prelude

- Before getting into fixed effects, consider the standard regression

$$y_{i,t} = \alpha + \beta x_{i,t} + \varepsilon_{i,t}$$

- Recall, from standard OLS results, that if we define

$$\tilde{x}_{i,t} \equiv x_{i,t} - \frac{1}{NT} \sum_{t=1}^{T} \sum_{i=1}^{N} x_{i,t} \quad and \quad \tilde{y}_{i,t} \equiv y_{i,t} - \frac{1}{NT} \sum_{t=1}^{T} \sum_{i=1}^{N} y_{i,t}$$

the β and $\varepsilon_{i,t}$ in the below regression are identical to the ones at the top

$$\tilde{y}_{i,t} = \beta \tilde{x}_{i,t} + \varepsilon_{i,t}$$

- ***In sum, the intercept is ‘taking out the means’***

a. Fixed effects

- Now, allow for firm fixed effects:

$$y_{i,t} = \alpha_i + \beta x_{i,t} + \varepsilon_{i,t}$$

- For simplicity, assume N = 2.
- Then Y is the same as before, but X has firm dummies and becomes:

$$\underbrace{X}_{TN\times 3} = \begin{bmatrix} 1 & 0 & x_{1,1} \\ 1 & 0 & x_{1,2} \\ \vdots & \vdots & \vdots \\ 1 & 0 & x_{1,T} \\ 0 & 1 & x_{2,1} \\ 0 & 1 & x_{2,2} \\ \vdots & \vdots & \vdots \\ 0 & 1 & x_{2,T} \end{bmatrix} \quad \textit{and we have} \quad \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \beta \end{bmatrix} = (X'X)^{-1}X'Y$$

UCLA Anderson
SCHOOL of MANAGEMENT

a. Fixed effects

- What does this firm fixed effect achieve?
- Calculate beta explicitly using the usual (X'X)^(-1)X'Y formula

$$\beta = \frac{Cov\left(y_{i,t} - \overline{y_i}, x_{i,t} - \overline{x_i}\right)}{Var\left(x_{i,t} - \overline{x_i}\right)} \quad where \quad \overline{y_i} \equiv \frac{1}{T}\sum_{t=1}^{T} y_{i,t}\,, \quad \overline{x_i} \equiv \frac{1}{T}\sum_{t=1}^{T} x_{i,t}$$

- In words: with firm fixed effects, we are capturing variation in the independent and dependent variable as deviations from each firms' mean of these variables, not from the grand mean as we would in a typical regression
- It is as if we are demeaning all variables at the firm-level. I.e., we are not explaining variation in means across firms with beta.
- Below, for comparison, is the beta from a regression ***without*** the fixed effect:

$$\beta = \frac{Cov\left(y_{i,t} - \bar{y}, x_{i,t} - \bar{x}\right)}{Var\left(x_{i,t} - \bar{x}\right)} \quad where\ \bar{y} \equiv \frac{1}{TN}\sum_{t=1}^{T}\sum_{i=1}^{N} y_{i,t}\,, \ \bar{x} \equiv \frac{1}{TN}\sum_{t=1}^{T}\sum_{i=1}^{N} x_{i,t}$$

a. Panel regression and Fama-MacBeth

The Fama-MacBeth looks a lot like the panel regression on the previous slides

- *After all: we are trying to estimate δ_0 and δ_1, for instance:*

(Note: the deltas do not have time or firm subscripts)

$$R_{i,t+1} = \delta_0 + \delta_1 lnBM_{i,t} + \varepsilon_{i,t+1}$$

However:

- While Fama-MacBeth is a kind of panel approach, it cross-sectionally demeans and effectively standardizes the predictor variable at each time *t*, and in addition weighs each time *t* coefficient the same when taking the average
    - Thus, portfolio returns from years where there are few firms are weighted as much as portfolio returns from years where there are many firms
    - A standard panel regression does not do this. It weighs observations equally (unless you specify a weighted panel regression).

- We will next continue with Fama-MacBeth to discuss how to assess the marginal contribution of a given trading signal as the Fama-MacBeth regression has a very nice portfolio return interpretation
    - We get back to the panel with its fixed effects and standard error clustering after this segment

b. MR (multiple regression) vs. SR (simple regression)

We ended Topic 1 with a review of Fama-MacBeth regressions, relating the FM regression coefficients to returns to implementable trading strategies.

Here, we will consider multiple regressions and the well-known *Omitted Variable Bias*

The overall concept: How to establish that your proposed trading signal (e.g., a sentiment index from social media sites) is ***valuable above and beyond*** other well-known trading signals

- Value in economics is about ***marginal contribution***

First, let's run some simple Fama-MacBeth regressions

- Returns regressed only on lnBM first, then add other signals

UCLAAnderson
SCHOOL of MANAGEMENT

b. Fama-MacBeth using lnBM (value)

Hard-code Fama-MacBeth procedure:

```
➤ # Fama-MacBeth Regressions
➤ # loop through the years in the data base
➤ port_ret = NULL > for (i in 1980:2014) +
➤ {
➤ temp <- StockRetAcct_DT[year==i,]
➤ fit_yr <- lm(temp$ExRet ~ temp$lnBM, data=temp)
➤ temp <- coefficients(fit_yr)
➤ port_ret = rbind(port_ret,temp[2])
➤ }

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

➤ fm_output
```

```
$MeanReturn
[1] 0.0146509

$StdReturn
           temp$lnBM
temp$lnBM 0.09157593

$SR_Return
          temp$lnBM
temp$lnBM 0.1599864

$tstat_MeanRet
          temp$lnBM
temp$lnBM 0.9464924
```

B/M sorted portfolio not statistically significant in this sample!

- Mean return not the same as in portfolio sort as not the same portfolio and different scale (leverage)

Sharpe ratio is low

- not affected by leverage

b. Fama-MacBeth and Portfolio Weights

- Here, I derive the portfolio weight expression to make the connection between Fama-MacBeth regressions and portfolio sorts 100% clear.

- Consider the simple cross-sectional for a particular time *t*:

$$R_{i,t} = \lambda_{0,t} + \lambda_{1,t} x_{i,t-1} + \varepsilon_{i,t}$$

- Define:

$$\underbrace{X_{t-1}}_{N\times 2} = \begin{bmatrix} 1 & x_{1,t-1} \\ \vdots & \vdots \\ 1 & x_{N,t-1} \end{bmatrix}, \quad \underbrace{R_t}_{N\times 1} = \begin{bmatrix} R_{1,t} \\ \vdots \\ R_{N,t} \end{bmatrix}, \quad then \qquad \begin{bmatrix} \lambda_{0,t} \\ \lambda_{1,t} \end{bmatrix} = (X_{t-1}{}' X_{t-1})^{-1} X_{t-1}{}' R_t$$

b. Fama-MacBeth and Portfolio Weights

- Let’s look inside the expression: $(X_{t-1}{}'X_{t-1})^{-1}X_{t-1}{}'R_t$
- First:

$$X_{t-1}{}'X_{t-1}\text{=N}\begin{bmatrix} 1 & \frac{1}{N}\sum_{i=1}^{N} x_{i,t-1} \\ \frac{1}{N}\sum_{i=1}^{N} x_{i,t-1} & \frac{1}{N}\sum_{i=1}^{N} x_{i,t-1}^2 \end{bmatrix},\quad \text{(check this yourself!)}$$

Then (again, check this yourself):

$$(X_{t-1}{}'X_{t-1})^{-1}$$

$$= \frac{1}{N}\frac{1}{\frac{1}{N}\sum_{i=1}^{N} x_{i,t-1}^2 - \left(\frac{1}{N}\sum_{i=1}^{N} x_{i,t-1}\right)^2}\begin{bmatrix} \frac{1}{N}\sum_{i=1}^{N} x_{i,t-1}^2 & -\frac{1}{N}\sum_{i=1}^{N} x_{i,t-1} \\ -\frac{1}{N}\sum_{i=1}^{N} x_{i,t-1} & 1 \end{bmatrix}$$

UCLAAnderson
SCHOOL of MANAGEMENT

b. Fama-MacBeth and Portfolio Weights

- Define:

$$E_N\left[x_{i,t-1}\right] = \frac{1}{N}\sum_{i=1}^{N} x_{i,t-1}, \quad Var_N\left[x_{i,t-1}\right] = \frac{1}{N}\sum_{i=1}^{N} x_{i,t-1}^2 - \left(\frac{1}{N}\sum_{i=1}^{N} x_{i,t-1}\right)^2$$

- Then, we can write:

$$(X_{t-1}{}'X_{t-1})^{-1} = \frac{1}{N}\frac{1}{Var_N\left[x_{i,t-1}\right]}\begin{bmatrix} E_N\left[x_{i,t-1}^2\right] & -E_N\left[x_{i,t-1}\right] \\ -E_N\left[x_{i,t-1}\right] & 1 \end{bmatrix}$$

- Then:

$$(X_{t-1}{}'X_{t-1})^{-1}X_{t-1}{}'=\frac{1}{N}\frac{1}{Var_N\left[x_{i,t-1}\right]}\begin{bmatrix} E_N\left[x_{i,t-1}^2\right] & -E_N\left[x_{i,t-1}\right] \\ -E_N\left[x_{i,t-1}\right] & 1 \end{bmatrix}\begin{bmatrix} 1 & \cdots & 1 \\ x_{1,t-1} & \cdots & x_{N,t-1} \end{bmatrix}$$

b. Fama-MacBeth and Portfolio Weights

- Let’s focus on the second row (an 1 by N vector) of the last expression as we are mainly interested in understanding the second regression coefficient, $\lambda_{1,t}$:

$$\frac{1}{N}\frac{\left[x_{1,t-1}-E_N\left[x_{i,t-1}\right] \quad \cdots \quad x_{N,t-1}-E_N\left[x_{i,t-1}\right]\right]}{Var_N\left[x_{i,t-1}\right]}$$

The final step is to multiply this 1 by N vector by the N by 1 vector R_t.

$$\lambda_{1,t}=\sum_{i=1}^{N}\frac{1}{N}\frac{\left(x_{i,t-1}-E_N\left[x_{i,t-1}\right]\right)}{Var_N\left[x_{i,t-1}\right]}R_{i,t}$$

Define $w_{i,t-1}=\frac{1}{N}\frac{\left(x_{i,t-1}-E_N\left[x_{i,t-1}\right]\right)}{Var_N\left[x_{i,t-1}\right]}$ and we have $\lambda_{1,t}=\sum_{i=1}^{N}w_{i,t-1}\,R_{i,t}$

b. Fama-MacBeth using lnProf (profitability)

Only change from previous FMB code: lnProf instead of lnBM in the below

```
➢ # Fama-MacBeth Regressions
➢ # loop through the years in the data base
➢ port_ret = NULL > for (i in 1980:2014) +
➢ {
➢ temp <- StockRetAcct_DT[year==i,]
➢ fit_yr <- lm(temp$ExRet ~ temp$lnProf, data=temp)
➢ temp <- coefficients(fit_yr)
➢ port_ret = rbind(port_ret,temp[2])
➢ }

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

➢ fm_output
```

```
$MeanReturn
[1] 0.1117721

$StdReturn
            temp$lnProf
temp$lnProf   0.2046914

$SR_Return
            temp$lnProf
temp$lnProf   0.5460517

$tstat_MeanRet
            temp$lnProf
temp$lnProf    3.230485
```

Profitability-sorted portfolio *IS* statistically significant in this sample!

Sharpe ratio is high

b. Let’s do both: Multiple Regression

```
> port_ret = NULL
➢ for (i in 1980:2014)
➢ + {
➢ + temp <- StockRetAcct_DT[year==i,]
➢ + fit_yr <- lm(temp$ExRet ~ temp$lnBM + temp$lnProf, data=temp)
➢ + temp <- coefficients(fit_yr)
➢ + port_ret = rbind(port_ret,temp[2:length(temp)])
➢ + }

➢ fm_output = list(MeanReturn = colMeans(port_ret), StdReturn = sqrt(diag(var(port_ret))),
➢ + SR_Return = colMeans(port_ret)/sqrt(diag(var(port_ret))),
➢ + tstat_MeanRet = sqrt(1+2014-1980)*colMeans(port_ret)/sqrt(diag(var(port_ret))))
```

```
$MeanReturn
  temp$lnBM temp$lnProf
 0.01943783  0.12311199

$StdReturn
  temp$lnBM temp$lnProf
 0.09093362  0.24158392

$SR_Return
  temp$lnBM temp$lnProf
  0.2137585   0.5096034

$tstat_MeanRet
  temp$lnBM temp$lnProf
   1.264612    3.014854
```

B/M-sorted portfolio is slightly more statistically significant, with higher average return

Profitability-sorted portfolio also has slightly higher average return than in univariate regression case

Both of these facts can be traced back to a negative correlation between b/m and profitability

b. Let’s add industry dummies

Now, both profitability and value are significant with much higher Sharpe ratios. ***Why?***

```
> port_ret = NULL
➢ for (i in 1980:2014)
➢ + {
➢ + temp <- StockRetAcct_DT[year==i,]
➢ + fit_yr <- lm(temp$ExRet ~ temp$lnBM + temp$lnProf + as.factor(ff_ind), data=temp)
➢ + temp <- coefficients(fit_yr)
➢ + port_ret = rbind(port_ret,temp[2:length(temp)])
➢ + }

➢ fm_output = list(MeanReturn = colMeans(port_ret), StdReturn = sqrt(diag(var(port_ret))),
➢ + SR_Return = colMeans(port_ret)/sqrt(diag(var(port_ret))),
➢ + tstat_MeanRet = sqrt(1+2014-1980)*colMeans(port_ret)/sqrt(diag(var(port_ret))))
```

```
$SR_Return
             temp$lnProf                temp$lnBM  as.factor(temp$ff_ind)2  as.factor(temp$ff_ind)3 
              0.63320324               0.34942536              -0.08357964              -0.11511659 
 as.factor(temp$ff_ind)4  as.factor(temp$ff_ind)5  as.factor(temp$ff_ind)6  as.factor(temp$ff_ind)7 
             -0.10875525               0.03191245               0.03696328               0.10031592 
 as.factor(temp$ff_ind)8  as.factor(temp$ff_ind)9 as.factor(temp$ff_ind)10 as.factor(temp$ff_ind)11 
             -0.15321847               0.05589172               0.30738341               0.05385884 
as.factor(temp$ff_ind)12 
             -0.26645630 

$tstat_MeanRet
             temp$lnProf                temp$lnBM  as.factor(temp$ff_ind)2  as.factor(temp$ff_ind)3 
               3.7460809                2.0672283               -0.4944638               -0.6810390 
 as.factor(temp$ff_ind)4  as.factor(temp$ff_ind)5  as.factor(temp$ff_ind)6  as.factor(temp$ff_ind)7 
              -0.6434048                0.1887966                0.2186777                0.5934770 
 as.factor(temp$ff_ind)8  as.factor(temp$ff_ind)9 as.factor(temp$ff_ind)10 as.factor(temp$ff_ind)11 
              -0.9064527                0.3306599                1.8185048                0.3186332 
as.factor(temp$ff_ind)12 
              -1.5763767 
```

b. Why did Sharpe ratios of value and profitability strategies increase?

Because, profitability and book-to-market ratios vary across industries

- Thus, an industry component in these characteristics
- But, industry exposure do not carry a risk premium (empirical statement)
  - Thus, the industry exposure just adds volatility, noise
  - So, control for this by adding industry fixed effect!

The multiple regression gets at the ***marginal effect*** of changing b/m and profitability

- I.e., holding industry and profitability exposures constant, what is the effect of varying the b/m ratio of the portfolio?
- I.e., holding industry and b/m exposures constant, what is the effect of varying the profitability characteristic of the portfolio?

Well, that’s wonderful, but how do I trade these strategies?

UCLA Anderson
SCHOOL of MANAGEMENT

b. Enhanced trading strategies

The previous Fama-MacBeth regressions give us the portfolio weights of a long-short portfolio (which leverage you can adjust as you please)

We already saw how this works in the 1-factor regression. On the previous slide, there are 14 factors (12 industry dummies (one is the intercept), lnBM, and lnProf).

So, if we want to trade the marginal book-to-market effect, which is the third regressor after the intercept and the lnProf factor, we use the N_t portfolio weights given by the 3rd row of $(X_t'X_t)^{-1}X_t'$, where

$$X_t = \begin{bmatrix} 1 & lnProf_{1,t} & lnBM_{1,t} & indDum2_{1,t} & \cdots & indDum12_{1,t} \\ 1 & lnProf_{2,t} & lnBM_{2,t} & indDum2_{2,t} & \cdots & indDum12_{2,t} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & lnProf_{N_t,t} & lnBM_{N_t,t} & indDum2_{N_t,t} & \cdots & indDum12_{N_t,t} \end{bmatrix}$$

b. Enhanced trading strategies: Recipe

This exercise is what you should do when you want to pitch your new strategy!

1. Construct your trading signal (a characteristic, $z_{i,t}$)

2. Run a Fama-MacBeth regression adding other characteristics that you want to be robust to (e.g., market beta, value, profitability), and take out noise (e.g., industry exposure)

3. Get the trading strategy based on your signal by extracting the portfolio weights from the FMB regression per the previous slide. These portfolio weights ensure that the strategy (a) does not overlap with value or profitability trading and (b) is rid of unwanted noise (industry)

4. If you only want to take out noise, only add the industry dummies in the Fama-MacBeth regression

5. May want to run a value-weighted FMB regression to avoid to much trading (see solution to Problem Set 1)

b. Trading Strategy and Leverage

As explained earlier, the trading strategy from the Fama-MacBeth regressions yield portfolio weights that sum to zero.

- That is, the portfolios are long-short portfolios that require zero net capital (ignoring margin requirements and transaction costs)
- Also, recall that long-short portfolios are excess returns

You can thus regard this long-short portfolio as an overlay on any base portfolio you may have.

- The simplest baseline is a portfolio invested 100% in the risk-free rate
- The leverage of the portfolio can be adjusted as one pleases. As usual, changing leverage would not affect Sharpe ratios.
- Let *k* be a choice variable for the investor that multiplies all the Fama-MacBeth portfolio weights. Setting *k* = 0 means no exposure (obviously)
  - Then:

$$E[R_{TotPort}] = E\left[R_{risk-free}\right] + kE[R_{Fama-MacBeth}]$$

$$\sigma[R_{TotPort}] = k\sigma[R_{Fama-MacBeth}]$$

$$SR[R_{TotPort}] = \frac{E[R_{TotPort}] - E\left[R_{risk-free}\right]}{k\sigma[R_{Fama-MacBeth}]} = SR[R_{Fama-MacBeth}]$$

b. Trading Strategy and Leverage

OK, let’s implement this for the value strategy, controlling for profitability and industry exposure.

- I normalize the standard deviation to be 15% p.a. and compare to original simple value strategy with same standard deviation
- First show how to get portfolio weights

```
➢ # getting portfolio weights for "new" value factor implied by the preceding FMB regressions
➢ # choose the current date (end of sample)

➢ LastDate <- na.omit(StockRetAcct_DT[year==2014,])

➢ Nt = length(LastDate$ExRet)

➢ Xmat <- cbind(rep(1,Nt),LastDate$lnProf,LastDate$lnBM)

➢ for (ii in 1:11) # note drop last industry dummy as we have intercept in regression
   { + Xmat <- cbind(Xmat,(LastDate$ff_ind==ii)) + }

➢ portweights_lnBM = solve(t(Xmat)%*%Xmat)%*%t(Xmat)

➢ # lnBM is third row (first is intercept, second is profitability, fourth and on are industry)
➢ portweights_lnBM = c(0,0,1,0,0,0,0,0,0,0,0,0,0,0) %*% portweights_lnBM

➢ # scale portfolio weights to get 15% standard deviation of returns
➢ portweights_lnBM = portweights_lnBM * 0.15 / sqrt(var(port_ret[,2]))
```

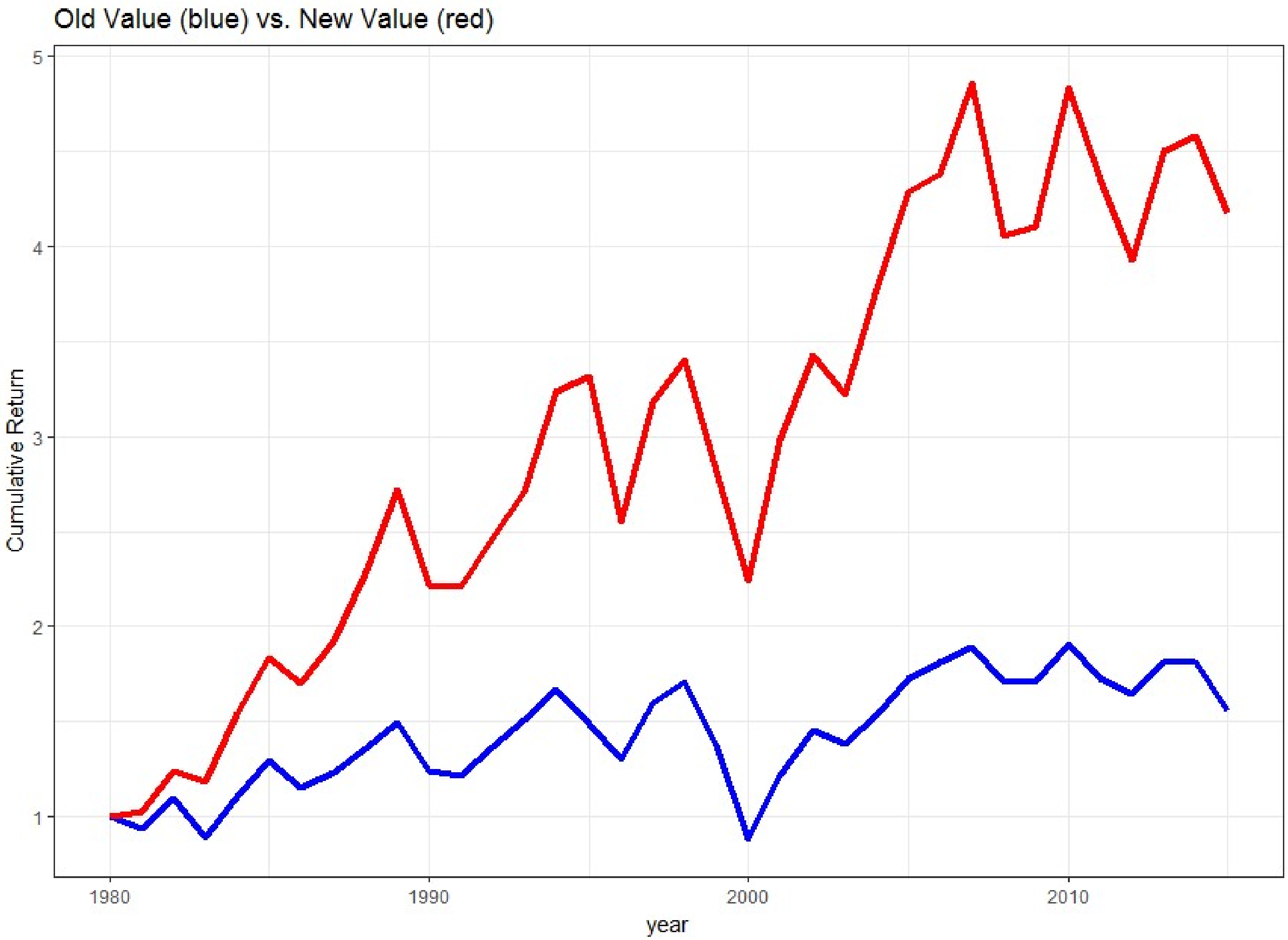
b. Trading Strategy and Leverage

Next, get return series and plot

```
➢ # for plotting, get the scaled excess portfolio return
➢ lnBM_ret = port_ret[,2] * 0.15 / sqrt(var(port_ret[,2]))

➢ # create cumulative log return series
➢ cum_ret_lnBM = 0 > for (ii in 1:35)
 + { + cum_ret_lnBM = rbind(cum_ret_lnBM,cum_ret_lnBM[ii]+log(1+lnBM_ret[ii])) + }

➢ # get "old" simple value strategy returns
➢ port_ret = NULL

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

➢ lnBM_old_ret = port_ret[,1] * 0.15 / sqrt(var(port_ret[,1]))\

➢ cum_ret_oldlnBM = 0

➢ for (ii in 1:35) + { + cum_ret_oldlnBM = rbind(cum_ret_oldlnBM,cum_ret_oldlnBM[ii]+log(1+lnBM_old_ret[ii])) + }

➢ # plot exponential of cumulative log return to get to regular cumulative returns
➢ # shows pretty convincingly how the new cleaned-up value strategy performs better than the old

➢ qplot(c(1980:2015), exp(cum_ret_oldlnBM), geom="line", xlab="year",ylab="Cumulative Return",color = I("blue"),
 size=I(1.5), main = "Old Value (blue) vs. New Value (red)")
 + geom_line(aes(y=exp(cum_ret_lnBM)),color = I("red"),size=I(1.5)) + theme_bw()
```

UCLAAnderson
SCHOOL of MANAGEMENT

b. Old Value vs. New (improved) Value



b. MR (multiple regression) vs. SR (simple regression)

Another verbalization of what we just discussed:

As book-to-market increases across firms, there is a “pure” effect of increasing discount rates (higher future returns).

But there are also strong industry component in book-to-market. Industry risk has historically not been priced (marginally).

Thus, the book-to-market coefficient in the simple regression reflects two effects:

1. “direct” effect which is the return predictor and it is positive
2. “indirect” effects from industry and profitability which are both correlated with book-to-market

The Simple Regression coefficient is the sum of these effects which is smaller than the “pure” book-to-market effect.

c. Omitted Variable Bias

Econometricians call what we just saw "omitted" variables bias.

Suppose the "true" effect of a variable X is from a regression of Y on X,Z.

$$y_i = \alpha + \beta x_i + \gamma z_i + \varepsilon_i$$

But we actually only run Y on X (that is, we "omit" Z). Then the least squares coefficient on X will no longer be a pure estimate even if we have an infinite amount of data.

$$y_i = \alpha + \beta^* x_i + \varepsilon_i^*$$

c. Omitted Variable Bias

For large samples the difference between the univariate regression estimate and the multivariate regression estimate will be approximately

$$\hat{\beta}_{SR} - \hat{\beta}_{MR} \approx \gamma \frac{\operatorname{cov}(X,Z)}{\operatorname{var}(Z)}$$

This quantity is often called the “omitted variable bias.” Obviously, there is no omitted variable bias when Z is not correlated with X. The larger the indirect effect of Z on X, the larger the bias.

BUT, even if Z has a large effect on Y, omitted variable bias is only present if X and Z are correlated!!!

c. Omitted Variable Bias

How can we eliminate Omitted Variable Bias?

In theory, omitted variable bias may always be present.

We must think carefully about what variables need to be measured and controlled for. If there are obvious omitted variables that are not available and can plausibly be correlated with the variable of interest, X, we are in trouble.

The only way to insure omitted variable bias is not present is to vary X in an experimental fashion using randomization. By definition, random variation in X is not correlated with anything as long as we have a large enough sample.

d. Panel regressions: case study

We will use the package “lfe” (*Linear group Fixed Effects*)

- Also “stargazer” which helps tabulate results and can export to .tex

Our application will be firm-level earnings forecasting

- Idea: using large set of historical data may beat analyst forecasts
  - Actually, analysts have been shown to have a strong upwards bias in their earnings forecasts and poor overall ability to be better than very simple models
- Forecasting earnings or other cash flow related quantities is important for valuations (possible input to trading strategies), as well as capital budgeting within firms
- Methodology is general – could also be used for forecasting realized variances, covariances, etc.

We will continue using the “StockAcct”-dataset, which contains many accounting variables as well as returns and market values.

d. Firm Earnings

Our dependent variable will be *Return on Equity* (ROE):

$$ROE_{i,t} = \frac{Net\ Income_{i,t}}{BookEquity_{i,t-1}}$$

We will continue using the “StockRetAcct”-dataset, which contains many accounting variables as well as returns and market values.

In particular, we will consider the variable *lnROE*, which is the log of 1 + ROE minus log inflation (I.e., log real ROE).

For now, we will ignore out-of-sample vs. in-sample issues

d. Firm Earnings

First, some summary statistics (across firms and time):

1. Mean lnROE = 8.8%
2. St.dev. lnROE = 27.4%

Consider the simple panel forecasting model:

$$lnROE_{i,t+1} = \delta + \beta' X_{i,t} + \varepsilon_{i,t+1}$$

Let’s start with lagged lnROE and lagged log book-to-market as forecasting variables (obvious ones).

d. Firm Earnings: Panel Forecasting

```
➢ # What predicts next year's ROE?
➢ setorder(StockRetAcct_DT, FirmID, year) # Set order of data so that shift does what we want it to
➢ StockRetAcct_DT[, lead_lnROE := shift(lnROE, type = 'lead'), by = FirmID] # Define next year's lnROE
➢ roe_panel1 = felm(lead_lnROE ~ lnROE, StockRetAcct_DT) # Regression with no fixed effects or clustering
➢ stargazer(roe_panel1, type = 'text', report = 'vc*t') # Output regressions as text, report t-stats
```

```
=====================================
                 Dependent variable:
               ---------------------------
                   lead_lnROE
-----------------------------------------------
lnROE                 0.522***
                   t = 150.845

Constant               0.041***
                   t = 45.864


-----------------------------------------------
Observations              60,473
R2                    0.273
Adjusted R2               0.273
Residual Std. Error     0.201 (df = 60471)
=====================================
Note:          *p<0.1; **p<0.05; ***p<0.01
```

Decent predictability: R2 = 30%

Positive autocorrelation

Careful about those *t*-statistics!

- Created under (bad) assumption of i.i.d. error terms

d. Standard Errors: Clustering

Firms’ shocks to earnings are correlated within firms

- E.g., a firm’s positive earnings shock is autocorrelated
  - To account for this add standard error clustering at the firm level

Clustering! (The first 0 is for fixed effects, the second for IV approach)

```
➢ # cluster standard errors at the firm level
➢ roe_panel2 = felm(lead_lnROE ~ lnROE | 0 | 0 | FirmID, StockRetAcct_DT)
➢ stargazer(roe_panel2, type = 'text', report = 'vc*t') # Output regressions as text, report t-stats
```

```
                    Dependent variable:
                  ---------------------------
                        lead_lnROE
-----------------------------------------------
lnROE                    0.522***
                       t = 35.359

Constant                  0.041***
                       t = 19.646

-----------------------------------------------
Observations              60,473
R2                         0.273
Adjusted R2                0.273
Residual Std. Error    0.201 (df = 60471)
==========================================
Note:            *p<0.1; **p<0.05; ***p<0.01
```

Note: R2 didn’t change

But, *t*-stats much smaller

d. Standard Errors: Clustering

- Cluster on time
- Simplest example: economy enters expansion, all firms make more money
- Note “stargazer” package really nice feature below

Clustering! (The first 0 is for fixed effects, the second for IV approach)

```
➢ roe_panel3 = felm(lead_lnROE ~ lnROE | 0 | 0 | year + FirmID, StockRetAcct_DT)
➢ stargazer(roe_panel1, roe_panel2, roe_panel3, type = 'text', report = 'vc*t')
```

```
==========================================================
                         Dependent variable:
                    ---------------------------------
                             lead_lnROE
                       (1)          (2)          (3)
----------------------------------------------------------------
lnROE                  0.522***    0.522***    0.522***
                    t = 150.845 t = 35.359 t = 16.072

Constant               0.041***    0.041***    0.041***
                    t = 45.864  t = 19.646 t = 6.992

----------------------------------------------------------------
Observations              60,473     60,473     60,473
R2                        0.273      0.273      0.273
Adjusted R2               0.273      0.273      0.273
Residual Std. Error (df = 60471)   0.201      0.201      0.201
==========================================================
Note:                     *p<0.1; **p<0.05; ***p<0.01
```

Note: Again R2 didn’t change

But, *t*-stats even smaller, evidence of both firm and time dependencies in the errors

d. Fixed Effects

Well-known permanent industry effects in accounting variables

- Due to different production opportunities as well as conventions
- Add an industry fixed effect (δ_j), like adding industry dummies:

$$lnROE_{i,j,t+1} = \delta_j + \beta' X_{i,t} + \varepsilon_{i,t+1}$$

Here *j* denotes industry and *i* is firm.

Be careful using firm-level fixed effects!! (That's like adding a dummy variable for each firm)

- Small sample issues are severe since median firm life in sample is only 10 years
- Thus, firm-level average is badly estimated, which affects other regression coefficients as well.
- Try to avoid. If you can't, run extensive Monte-Carlo simulations to assess likely magnitude of biases

d. Fixed Effects

```
➢ roe_panel4 = felm(lead_lnROE ~ lnROE | ff_ind | 0 | year + FirmID, StockRetAcct_DT)
➢ stargazer(roe_panel4, type = 'text', report = 'vc*t') # Output regressions as text, report t-stats
```

Fixed effect, at the industry level as defined by the ff_ind variable

```
================================
              Dependent variable:
          ---------------------------
               lead_lnROE
-----------------------------------------------
lnROE                  0.505***
                  t = 15.965

-----------------------------------------------
Observations              60,467
R2                        0.282
Adjusted R2               0.282
Residual Std. Error    0.200 (df = 60454)
===================================
Note:            *p<0.1; **p<0.05; ***p<0.01
```

Now, the R2 increased slightly

- Note: industry dummies are not reported in output

Not a huge effect in this case.

d. Fixed Effects

Let’s also add year fixed effects:

$$lnROE_{i,j,t+1} = \delta_j + \gamma_t + \beta' X_{i,t} + \varepsilon_{i,t+1}$$

Interpretation

- Remove average response (market factor)
- For instance: “I am not interested in trying to predict the market-level movements in earnings, just out- or under-performance relative to market”

d. Fixed Effects

```
➢ roe_panel5 = felm(lead_lnROE ~ lnROE | year + ff_ind | 0 | year + FirmID, StockRetAcct_DT)
➢ stargazer(roe_panel5, type = 'text', report = 'vc*t') # Output regressions as text, report t-stats
```

Fixed effects, at the industry and year levels

```
==============================================
                 Dependent variable:
             ---------------------------
                     lead_lnROE
-----------------------------------------------
lnROE                 0.504***
                     t = 15.681


-----------------------------------------------
Observations            60,467
R2                      0.291
Adjusted R2             0.290
Residual Std. Error   0.199 (df = 60421)
==============================================
Note:          *p<0.1; **p<0.05; ***p<0.01
```

Now, the R2 again increased slightly

- But, not a large effect. Thus, variation in earnings is mainly cross-sectional

d. Predicting firm earnings, big model

```
===============================================
                 Dependent variable:
            ---------------------------
                      lead_lnROE
-----------------------------------------------
lnROE                 0.362***
                     t = 16.054

lnBM                 -0.053***
                     t = -9.305

lnProf                0.229***
                      t = 7.721

lnLever               -0.010**
                     t = -2.133

lnIssue              -0.062***
                     t = -7.773

lnInv                -0.056***
                     t = -3.935

-----------------------------------------------
Observations           55,040
R2                    0.351
Adjusted R2            0.351
Residual Std. Error   0.173 (df = 55022)
===============================================
Note:          *p<0.1; **p<0.05; ***p<0.01
```

```
➢ roe_panel6 = felm(lead_lnROE ~ lnROE + lnBM +lnProf +
  lnLever + lnIssue + lnInv         | ff_ind | 0 | year +
  FirmID, StockRetAcct_DT)

➢ stargazer(roe_panel6, type = 'text', report = 'vc*t') #
  Output regressions as text, report t-stats
```

Now, the R2 again increased

- Many variables are significant
- Note: marginal effect of higher investment is lower future earnings…!

d. Predicting earnings in 5 years, big model

```
==========================================================
                   Dependent variable:
              -----------------------------------
                 lead_lnROE      lead5_lnROE
                    (1)              (2)
----------------------------------------------------------
lnROE              0.362***         0.068***
                t = 16.054       t = 3.430

lnBM              -0.053***        -0.025***
                t = -9.305       t = -5.332

lnProf             0.229***         0.168***
                t = 7.721        t = 4.786

lnLever           -0.010**          0.002
                t = -2.133       t = 0.459

lnIssue           -0.062***        -0.080***
                t = -7.773       t = -7.922

lnInv             -0.056***        -0.053***
                t = -3.935       t = -5.381

----------------------------------------------------------
Observations          55,040          36,235
R2                 0.351           0.088
Adjusted R2            0.351           0.087
Residual Std. Error 0.173 (df = 55022) 0.192 (df = 36217)
==========================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```

```
➢ roe_panel7 = felm(lead5_lnROE ~ lnROE + lnBM +lnProf +
  lnLever + lnIssue + lnInv          | ff_ind | 0 | year +
  FirmID, StockRetAcct_DT)

➢ stargazer(roe_panel6, roe_panel7, type = 'text', report
  = 'vc*t') # Output regressions as text, report t-stats
```

Now, the R2 is quite a bit lower, as we would expect

- Still, many variables are significant
- Note: marginal effect of higher investment is ***still*** lower future earnings…!

d. Earnings prediction model

Used over 50,000 firm-year observations to create earnings forecasting model

- In-sample R2 was around 35%
- About 9% when predicting the annual earnings 5 years from now

Benefit a lot from cross-section

- Year fixed effects not that important
- Most variation is cross-sectional, that's how we get such high *t*-stats
- Can use model for a new firm even
  - Set all characteristics you do not have to their unconditional average
- Impossible to run individual model at the firm-level given only 10 year median firm survival in data

d. Firm variance prediction model

We can use the same model to predict firm variance.

In fact, that is what you will do in Problem Set 2

e. Panel regression postscript

Main idea:

- ***Get power*** from looking at both time-series and cross-section
- Assumes 'betas' are the same across time and firms
- Can remove time and firm (or industry) fixed effects
  - Time f.e.: Identification of beta entirely from cross-sectional variation in response (y) from feature (x)
  - Firm f.e.: Identification of beta entirely from time-series variation in response (y) from feature (x)
- Need to make sure standard errors are appropriate to account for potential correlation patterns (across T and N; clustering)
- Routines exist for both unbalanced and balanced panels.
- Big data: time is typically short, cross-section can be huge!

e. Fama-MacBeth and Panel: When to use

A cross-sectional (Fama-MacBeth) regression is a convenient way to do portfolio sorts!

- Especially powerful when we have multiple right-hand side variables
- Realized excess return on portfolio *k* is the regression coefficient, lambda_{t,k}
- Risk premium on each portfolio is then estimated using the portfolio's average sample return
  - This estimate will be quite noisy, unless you have a really long time-series sample

Panel regression estimates (constant) regression coefficients using both time-series and cross-sectional variation

- Big increase in power (regression coefficients have low standard error)
- However, requires assumption that regression coefficients are constant over time and in cross-section (potentially with exception of intercepts)

e. Fama-MacBeth and Panel: When to use

Year fixed effects are appropriate if all you are interested in are cross-sectional differences

- E.g., predict the difference of two stocks' returns or earnings

Industry-fixed effects are appropriate if you believe each industry has permanent differences in the y variable (e.g., one can argue this is the case for ROE).

- If there are many firms in each industry, these industry fixed effects can be estimated with relatively little noise

Firm fixed effects I do not advise in forecasting exercises. These typically lead to overfitting and small-sample biases that can be quite severe.

- A cross-validation exercise, which we discuss later (Topic 4), will typically show that this is indeed the case and lead you to drop firm fixed-effects in your prediction exercise