

# Empirical Methods for Finance

## First Graded Assignment

### Analysts' Earnings Forecasts

Prof. Virginia Gianinazzi

Due: September 27, 2020 (midnight)

Companies release their earnings-per-share (EPS) on scheduled announcement dates, typically every quarter. Preceding earnings announcements, financial analysts' publish their forecasts of companies' earnings, based on their expectations about companies' growth and profitability.

In this assignment you empirically investigate the determinants of analysts' forecast errors, i.e the difference between forecasted earnings and announced earnings. For this purpose, you are given two datasets, *ibes\_2019.dta* and *crsp\_daily\_2019.dta*. The variables contained in both datasets are described at the end of this document.

#### 1. BASIC DATA MANIPULATION [25%]

- (a) In STATA, open *ibes\_2019.dta*. Briefly describe the structure of the data. What is the smallest unit of observation?
- (b) Rename *anndats\_act* as *date*, then merge (`m:1`) *ibes\_2019.dta* to (`using`) *crsp\_daily\_2019.dta*. Keep only the observations that are successfully merged. Compute companies' market capitalization (*mktcap*).
- (c) Compute the consensus forecast (*consensus*), defined as the median forecast across analysts, and the standard deviation of the forecasts (*dispersion*).

Then, compute the following two measures of forecast accuracy:

- $fe = \left| \frac{actual - consensus}{actual} \right|$
  - $fd = \frac{dispersion}{|actual|}$
- (d) Generate a variable *coverage* equal to the number of analysts providing a forecast for a given earnings announcement.

- (e) Drop the following industries from the dataset: international affairs and non-op. establishments (SIC 9000-9999), foreign governments (SIC 8888), utilities (SIC 4000-4999) and agriculture, fishing and hunting (SIC 0000-0999).  
Create a dummy variable *financials* for financial firms (SIC 6000-6999).

## 2. SUMMARY STATISTICS AND PLOTS [30%]

- (a) How many distinct earnings announcements events are in the data?
- (b) Collapse the data at earnings announcement level and keep the mean of *fe*, *fd*, *coverage*, *mktcap*, *financials*.
- (c) Produce a summary statistics table with the mean, standard deviation, min and max of all the variables in the dataset. Include also a correlation matrix between all the variables.
- (d) Do a scatter plot of *fe* against *coverage*, and of *fd* against *coverage*. Label the axes in a meaningful way. Briefly comment on the two charts.

## 3. OLS: ESTIMATION AND INTERPRETATION OF THE RESULTS [30%]

- (a) Run a regression of *fe* on *financials* and then of *fd* on *financials* and *coverage*. Interpret the coefficients and the  $R^2$  of both regressions.
- (b) Re-run the above regression, where you additionally control for *mktcap*. How does the interpretation of the coefficient on *coverage* change?
- (c) Units of measurement.
  - i. Scale the variable *mktcap* such that it gives the company market capitalization in billions of dollars. Re-run the last regression. How does the coefficient change? And the t-statistic? Explain.
  - ii. Suppose that you are allowed to report only two decimals in your tables. Do you see a problem? What would you do?
- (d) Under which assumptions can we interpret the coefficient on *coverage* as measuring the causal effect of analysts' coverage on forecast accuracy? What threats to the identification of the causal effect do you see in this case?

## 4. OLS: ASSUMPTIONS AND MECHANICS [15%]

- (a) After running the previous regression of *fe* on *financials*, *coverage* and *mktcap*, compute<sup>1</sup>

$$\sum_{i=1}^N \hat{u}_i(\hat{y}_i - \bar{y}).$$

---

<sup>1</sup>Hint: after the estimation use `predict` and `predict, residuals` to obtain the fitted values and the residuals.

What is the result of the summation? Explain.

(b) [THEORY QUESTION] Consider the following population model

$$y_i = \beta_0 + \beta_1 D_i + \beta_2 x_i + \varepsilon$$

where  $D_i$  is a dummy variable = 1 if  $i$  is a financial firm. What would be the consequence of estimating the model without  $D_i$ ?

### Datsets and Variables Description

The dataset *ibes\_2019.dta* contains the following variables:

- *permno*: CRSP unique stock level identifier
- *analys*: analyst identifier
- *estimator*: analyst's firm identifier
- *value*: earnings (EPS) forecast
- *actual*: realized earnings (EPS)
- *anndats*: date of forecast announcement
- *anndats\_act*: date of actual earnings announcement

The dataset *crsp\_daily\_2019.dta* contains the following variables:

- *permno*: CRSP unique stock level identifier
- *date*: date
- *siccd*: industry code
- *prc*: stock price
- *shrout*: number of shares outstanding (in thousands)