Empirical Methods for Finance
First Graded Assignment
Analysts Forecasts and Decision Fatigue

Prof. Virginia Gianinazzi
Due: November 16, 2021 (midnight)

You recently learned about a phenomenon called *decision fatigue*, which refers to the idea that quality of decisions declines with the number of decisions that a person has to make during a day. You are interested in whether decision fatigue also affects analysts' forecasts of company earnings. In particular, you want to investigate empirically if, on any given day, the number of forecasts that an analyst has already made affects the accuracy of the next forecast.

For this purpose, you collect data on analyst earnings forecasts and actual earnings announcements in 2016-2020. Each data point corresponds to the forecast of a given analyst about the earnings of a given company issued on a given day. The data are saved in *ibes_ps1.dta*. The variables are described at the end of this document.

1. Sample selection and variable definition [25%]

   Open the data and perform the following tasks.

   For this section, you just have to generate variables and clean the data in your code. You do not have to present any explanation of what you are doing in the PDF. The only exception is question 1e) where you are asked to interpret the variable.

   (a) drop if either forecasted EPS or actual EPS are missing.

   (b) compute a variable *coverage* defined as the number of analysts covering a given earnings announcement. (*Hint:* in Stata check the command `egen...nvals` with the option `by`.)

   (c) generate a variable *fatigue* defined as the number of forecasts that an analyst has already made during the day plus 1. So, for a given analyst on a given day, *fatigue*=1 for the first forecast of the day, 2 for the second forecast, etc.

(d) define forecast error *fe* as the absolute difference between the actual EPS and the forecasted value.

(e) calculate the relative forecast accuracy as

$$relacc = \frac{median(fe) - fe}{std(fe)}$$

where *median(fe)* is the median forecast error across all forecasts made by analysts about a given company earnings announcement. *std(fe)* is the standard deviation of these forecast errors.

How do you interpret a positive vs. negative value of *relecc*?

(f) Generate a variable *followed* equal to the total number of firms that an analyst is covering in a given year.

(g) Generate a variable experience (*exper*) equal to the number of years the analyst has been following a given firm (the sample starts in 2016, therefore in 2020 an analyst can have a maximum experience of 4 years with any given company).

(h) Keep only forecasts made in year 2020.

(i) Finally, winsorize *relacc* at the 1% level. (*Hint:* in Stata check the command `winsor2` with the option `replace`.)

2. Summary Statistics and Plots [30%]

(a) How many earnings announcements are left in the final sample?

(b) How many unique analysts are in the sample?

(c) On average, how many firms does an analyst follow in 2020? Provide also minimum number, maximum number, median and standard deviation

(d) On average, how many analysts follow a given earnings event? Provide also minimum number, maximum number, median and standard deviation

(e) On average, how many forecasts does an analyst make during one day? Provide also minimum number, maximum number, median and standard deviation

3. Analysis [40%]

(a) You first consider the following univariate model:

$$relacc = \beta_0 + \beta_1 fatigue + u \qquad (1)$$

what sign do you expect for β_1 if decision fatigue affects financial analysts earnings forecasts?

(b) Which of the following statements is/are true? Briefly explain.

[A] Only under the zero conditional mean assumption, β_1 measures the causal effect of fatigue on forecast accuracy (*relacc*).

[B] β_1 is the ceteris paribus effect of fatigue on forecast accuracy (*relacc*).

[C] Increasing fatigue by 1 leads to an increase of forecast accuracy by β_1.

[D] Only if the zero conditional mean assumption holds, β_1 measures the change in average forecast accuracy (*relacc*) caused by a 1 unit increase in fatigue.

(c) Estimate the univariate model. Interpret the estimated coefficient on fatigue.

(d) Is the effect statistically significant? Explain.

(e) Re-run the previous regression, but this time use *log(fatigue)* as your independent variable. Interpret the estimated slope coefficient in the current specification.

(f) Now, estimate the following multivariate model, where you control for experience of the analyst with the firm and the total number of firms followed.

$$relacc = \beta_0 + \beta_1 logfatigue + \beta_2 exper + \beta_3 follow + u \qquad (2)$$

How do you explain the change in $\widehat{\beta}_1$? Pick one answer from below and justify your choice.

[A] The estimated coefficient from the univariate model appears to be biased downward. This indicates that *fatigue* is either negatively related to the number of firms followed by the analyst or positively related to analyst experience in our sample.

[B] The estimated coefficient from the univariate model appears to be biased upward. This indicates that *fatigue* is either positively related to the number of firms followed by the analyst or negatively related to analyst experience in our sample.

[C] The estimated coefficient from the univariate model appears to be biased downward. This indicates that *fatigue* is positively related with either the number of firms followed by the analyst or the analyst's experience in our sample.

[D] The estimated coefficient from the univariate model appears to be biased upward. This indicates that *fatigue* is negatively related with either the number of firms followed by the analyst or the analyst's experience in our sample.

(g) You are worried that analysts may decide to tackle firms for which the forecasting problem is more complex early in the day, when they are not yet fatigued, leaving less challenging tasks for later in the day. If this was in fact the case, in which direction would our estimated $\widehat{\beta}_1$ be biased? Explain

(h) You run a regression of coverage on fatigue. What do you find? What does this regression suggest about your concern expressed in the previous point?

(i) Now estimate the following model with dummy variables

$$relacc = \gamma_0 + \gamma_1 D_2 + \gamma_2 D_3 + \gamma_3 D_4 + \gamma_4 D_{5+} + u \qquad (3)$$

where $D_2 = 1$ for the second forecast of the day (zero otherwise), $D_3 = 1$ for the third forecast of the day (zero otherwise), $D_4 = 1$ for the fourth forecast of the day (zero otherwise) and $D_{5+} = 1$ for the fifth or later forecast of the day.

i. Interpret the estimated $\widehat{\beta}_0$

ii. Interpret the estimated $\widehat{\beta}_1$

iii. Comment on the statistical significance of the coefficients.

iv. Test the hypothesis that the first forecast is twice as accurate forecast made fifth or later.

(j) Finally, choose three variables that you think it is important to include in the model. At least one of them must be something that you can observe in the dataset. Briefly explain your choices.

4. OLS MECHANICS [5%]

You are still worried that you are omitting important factors from your model. Compute the residuals and predicted values from the last estimated model. Compute the correlation between these two variables. How does the result address your concerns? Comment.

Variables Description

The dataset *ibes_ps1.dta* contains the following variables:

- cusip: company identifier
- analystcode: analyst identifier
- forecast: forecasted earnings per share
- eps: actual company earnings per share
- date: date on which the forecast was made
- time: time at which the forecast was made
- epsdate: date of the actual earning announcement made by the company