Empirical Problem Set
(219B, Spring 2015)

Stefano DellaVigna

May 5, 2015
Due by midnight on May 15, email PDF file

1 Introduction

The focus of the problem set is two-fold: (i) to induce you to work with a data set, prepare the necessary variable, and test hypotheses; (ii) to examine three anomalies that we discussed in class:

- The *post-earnings-announcement draft* (Chan, Jegadeesh, and Lakonishok, 1996; Bernard and Thomas, 1989). Announcements of good news in earnings are followed by higher returns over next 2-3 quarters, against the prediction that arbitrage would eliminate predictability in returns
- *Less Immediate Response and more Drift in the presence of more distractions* (DellaVigna and Pollet, 2009; Hirshleifer, Lim, and Teoh, 2009) Drift is stronger for announcements, and the immediate response is lower for announcements made on high-distraction days (Friday or day with more competing announcements). This is consistent with higher investor inattention with more distractions
- *CEOs adjust the earnings so as to meet analyst expectations* (Degeorge, Patel, and Zeckhauser, 1999)

The first part of the problem set asks you to go through a series of basic steps to analyze the response of stock prices to earning surprises. The second part of the problem set offers a choice between a number of alternative topics.

Earning Surprises. The main focus on the literature on earnings announcement has been on the response of investors to new information. Three main measures have been proposed in the literature to quantify the new information. The first two measures compare the earning announcement $e_{t,k}$ for company k in quarter t with the corresponding analyst forecast $\hat{e}_{t,k}$. The last measure compares the earning announcement $e_{t,k}$ with the earning announcement four quarters before, $e_{t-4,k}$. The analyst forecasts is defined as the median forecast among all the analysts that make a forecast in the last 45 (trading) days before the earning announcement. If an analyst made multiple forecasts in this time horizon, we consider the most recent one. In most of this problem set we consider Measure 1, but an optional question asks you to consider Measures 2 and 3.

Measure 1. Earnings surprise 1 is

$$s^1_{t,k} = \frac{e_{t,k} - \hat{e}_{t,k}}{p_{t,k}}. \tag{1}$$

The difference between the earning announcement and the forecast is divided by the lagged price of a share, $p_{t,k}$. The price of a share works as a renormalization factor: the earnings e are measured as earnings in dollar per share. The division by p implies that s^1 is the earning surprise as fraction of the value of the company. To see this, multiply numerator and denominator of expression (1) by the number of share $n_{t,k}$:

$$s^1_{t,k} = \frac{e_{t,k}n_{t,k} - \hat{e}_{t,k}n_{t,k}}{p_{t,k}n_{t,k}}.$$

In the numerator, $e_{t,k}n_{t,k}$ is the total profit for quarter t, and $\hat{e}_{t,k}n_{t,k}$ is the total forecasted profit. At the denominator is the market capitalization of a company, $p_{t,k}n_{t,k}$. The earning surprise measure, therefore, captures the unexpected profits as a share of total market value of the company. If $s^1_{t,k} = .01$, it means that the company earned unexpected profits equal to 1 percent of the value of the company.

Measure 2. Earnings surprise 2 is

$$s^2_{t,k} = \frac{e_{t,k} - \hat{e}_{t,k}}{\hat{d}_{t,k}},$$

where $\hat{d}_{t,k}$ is defined as the standard deviation between the earning forecasts of the analysts. This measure is therefore missing for companies with only one analyst, and in general for cases in which all the analysts agree in their assessment of the company's profits. This measure captures the intuition that the surprise is larger for companies in which the analysts agreed in their forecasts.

Measure 3. Earnings surprise 3 is

$$s^3_{t,k} = \frac{e_{t,k} - e_{t-4,k}}{d_{t,k}}.$$

The numerator is the difference between the earning announcement and the earning announcement 4 quarters before (the argument here is that there are seasonalities). The denominator $d_{t,k}$ is the standard deviation of the numerator over the previous 16 quarters. (Note: d is very different from $\hat{d}$ above)

Returns. We consider the response of stock returns to earnings surprises at different horizons. To capture the immediate response, one could look at $r^{(0,0)}$, that is, the stock return the same day as the announcement (measure as price at the close on day t minus the price at the close on day $t-1$). However, since announcements are often made after the markets are closed, one should look at $r^{(0,1)}$, that is the return for the same day and the next day. If one wants to look at the delayed response to the earning announcement, a typical measure is $r^{(3,75)}$, that is, the stock returns for the period $(t+3,\ t+75)$, where days are always meant as trading days. (this is finance!)

As for the measure of returns, three are typically used:

1. RAW is just the unadjusted stock return: $r_{t,k}$
2. NET is the stock return minus the market stock return, $r_{t,k} - r_{t,m}$
3. CAR is the abnormal return defined as $r_{t,k} - \hat{\beta} r_{t,m}$ where β is the correlation between stock k and the market. This beta is meant to correct for correlation with the market in a CAPM framework. They are unlikely to make much of a difference for a short-run event study like this one.

Data. In the dataset earn219b2007.dta, which you find zipped on the webpage of the class, I have already merged for you the information from Compustat, CRSP, and IBES. The data includes earnings from 1995 on in which the Compustat and IBES announcement dates differ by no more than 5 days. I have also generated the forecast of earnings $\hat{e}$. The data set that you see includes therefore information on earnings (MEDACT, multiply by the adjustment coefficient ADJ: ADJ*MEDACT), earnings forecast in the last 60 days (MEDEST60, multiply by the adjustment coefficient ADJ: ADJ*MEDEST60), stock returns (RAWWIN*–raw returns, NETWIN*–returns net of market returns, CARWIN*–returns adjusted for correlation with market), volume information (VOLU*), aggregate volume information (VOLUA*). [VOLU31 is volume same day of earning announcement, VOLU32 is volume next trading day, etc.] It also contains number of analysts (NEST60), standard deviation of earning forecast (STDEST60), SIC code of industry (SICCODE), company name (CONAME), price of shares (LAGPRICE), number of shares outstanding (LAGSHR). In order to make the data set small enough, it contains only companies with name up to "M".

2 Assignment – part 1

Answer the eight questions below, and then four at your choice in the next Section.

1. **Short-run response, OLS.** Construct the earnings surprise s^1 (use (ADJ*MEDACT-ADJ*MEDEST60)/LAGPRICE) and summarize its mean and distribution (use SUM *,D). Now estimate a simple linear specification to relate raw returns $r^{(0,1)}$ to s^1 as a measure of surprise:

$$r^{(0,1)}_{t,k} = \alpha + \phi s^1_{t,k} + \varepsilon_{t,k}. \tag{2}$$

How do you interpret the coefficient ϕ? Now run the same regression restricting the sample to $s^1_{t,k}$ in the range $[-.01, .01]$ and then in the range $[-.001, .001]$. Does the coefficient $\hat{\phi}$ change? What does that suggest about the specification of this regression?

2. **Short-run response, Non-linearity.** Now we explore more directly the possible non-linearity of the relationship. To provide non-parametric evidence, do a local polynomial regression of stock returns $r^{(0,1)}_{t,k}$ on the earnings surprise $s^1_{t,k}$, for simplicity restricting the plot in the range $s^1_{t,k}$ in $[-.01, .01]$. You are going to have to make some choices about bandwidth and type of kernel. In Stata, the command is LPOLY. The type of kernel typically does not matter very much, but the bandwidth certainly does. Show the graphs for both a relatively broad and a relatively narrow bandwidth. (The first will give you a smoother picture, the second will show you more the raw data) Is the relationship between the stock returns and the earnings surprise linear? Provide at least one interpretation for the observed non-linearity. That is, what features does the information contained in the earnings news have to have to justify this shape? No need to be behavioral here.

3. **Short-Run Response, Quantiles.** Now we use the quantile methodology. Sort the announcements into 11 quantiles as a function of $s^1_{t,k}$ as follows. Define quantile 6 as the group of announcements with no surprise ($e_{t,k} = \hat{e}_{t,k}$). Divide the announcements with negative surprises ($s^1 < 0$) in 5 equal-sized groups, with group 1 being the one with the most negative announcements and group 5 the one with least negative. Similarly, divide announcements with positive surprises ($s^1 > 0$) in 5 equal sized groups (groups 7 through 11). Group 11 will be the one with the most positive surprises. Finally, plot raw returns $r^{(0,1)}$ as a function of these 11 quantiles (See Figures 1a-1c in DellaVigna and Pollet (2009) for an example). Interpret the economic magnitudes in this plot. How can you reconcile the fact that this plot is approximately linear with the non-linearity in the previous questions? From now on, we are going to exploit the approximate linearity in the 11 quantiles. Denote by $q^1_{t,k}$ the quantile implied by the variable $s^1_{t,k}$, that is, $q^1_{t,k} \in \{1, 2, ..., 11\}$. Then run the OLS regression

$$r^{(0,1)}_{t,k} = \alpha + \phi q^1_{t,k} + \varepsilon_{t,k.} \tag{3}$$

How does the R^2 of this regression compare to the OLS regressions in (2)? Comment.

4. **Clustering I.** In running a regression, so far you have made the assumption that all the observations are i.i.d. draws from a Normal distribution. This is problematic, here as in most papers. The observations are likely to be heteroskedastic: larger surprises are likely to have higher return errors. In addition, you may be concerned about the correlation of errors across companies making an announcement on the same day. To allow for both heteroskedasticity and correlation of errors within a day, re-run regression (3) but cluster observations by day of announcement t. In Stata, you add to your regression specification ", ROBUST CLUSTER(T)". How do the point estimates change? How about the standard errors? Argue that the increase in the standard errors due to clustering means

that we were neglecting a positive correlation and 'overcounting' observations. From now on, except in the next point, maintain the clustering by t in your specifications.

5. **Clustering II.** Above I have suggested that you allow for correlation across announcements in one day by clustering by time t. You may also be concerned about the correlation of errors over time for the same company. You can check this by running specification (3) with ", ROBUST CLUSTER(PERMNO)", that is, you cluster by company identifier. What happens to standards errors? What does this suggest about the clustering that one should adopt in order to be conservative?

6. **Clustering III.** Now use the procedure of Cameron, Gelbach, and Miller (2006) that allows for double-clustering and provide standard errors for the coefficients. Evgeny Yakovlev provided some simple code that you can use for it, see the end of the problem set. Provide an intuitive explanation for what the double-clustering is doing by adding the variance-covariance matrices and subtracting then the common part.

7. **Post-earning announcement drift** (Chan, Jegadeesh, and Lakonishok, 1996; Bernard and Thomas, 1989). Use the quantile Methodology to plot raw returns $r^{(3,75)}$ as a function of the 11 quantiles in the earnings surprise variable s^1. What does the theory of efficient financial markets predict? What do you find? Measure the drift as the difference between the return for the highest quantile minus the return for the lowest quantile. Compute a standard error for this difference. Repeat this using the specification (3) using $r^{(3,75)}$ as dependent variable. Is $\hat{\phi}$ significant?

8. **Manipulation of earnings.** (DeGeorge, Patel, and Zeckhauser, 1999). Companies have some discretion in the accounting procedure, so they can manipulate the earnings release at the margin. Consider the numerator of the earnings surprise, $e_{t,k} - \hat{e}_{t,k}$. This is the earnings surprise per share. Plot the distribution of this variable for $-\$.1 \leq e_{t,k} - \hat{e}_{t,k} \leq \$.1$. (Excel histogram cent-by-cent would work, for example) Comment on whether the distribution has a discontinuous drop at $e_{t,k} - \hat{e}_{t,k} = 0$, and interpret it relating it to manipulation of earnings.

3 Assignment – part 2

In this second part we use the data set on earnings announcements to explore a dozen of different questions. Pick four of these questions and address them.

1. **Drift II.** We now explore further the finding that earning surprises forecast stock returns over the horizon (3,75). This is called the post-earnings announcement drift. We now analyze how much of the drift occurs at the next earnings announcement. Consider the specification

$$r_{t,k}^{(0,1)} = \alpha + \phi q_{t-1,k}^{1} + \varepsilon_{t,k}. \tag{4}$$

that is, you regress the stock response at time of an announcement on the earning surprise at the *previous* announcement (hence the notation $t-1$ in $q^1_{t-1,k}$). What is the estimate for $\hat{\phi}$? Argue that in efficient financial market ϕ should be zero. Now regress $s_{t,k}$ on $s_{t-1,k}$. What does this suggest about the role of analysts? Give two possible reasons for this. Can this analyst bias help explain the result in specification (4)? Replicate regression (4) using the earning surprise 2 announcements ago, 3 announcements ago, and 4 announcements ago. How are the patterns?

2. **Drift III.** Again on drift. We now go back and plot again raw returns $r^{(3,75)}$ as a function of quantiles in the earnings surprise variable. However, instead of sorting into quantile based on the earnings surprise $s^1_{t,k}$, we sort into ten deciles based on the raw return at announcement, $r^{(0,1)}_{t,k}$. The immediate stock response is an alternative measure of good/bad news at announcement. Comment on the difference between this graph and the graph in point 4 above. Which specification gives the largest earnings drift, as measured as in point 5?

3. **Manipulation of earnings II.** (DeGeorge, Patel, and Zeckhauser, 1999). The earning surprise per share that we consider at the previous point is not the only obvious target of attention for investors. Two other obvious variables are the earning per share itself, $e_{t,k}$, and the difference from the previous year, same quarter, $e_{t,k} - e_{t-4,k}$. Again, do a plot for each of these two variables. Is there a discontinuity at zero? Where does the discontinuity appear to be larger? What does this suggest about what investors pay most attention to?

4. **Inattention and Distractions.** Hirshleifer, Lim, and Teoh (2009) analyze, similarly to DellaVigna and Pollet (2009), the impact of distractions on the speed with which returns incorporate the earnings information. Instead of using Friday as a proxy for distractions, though, they use the number of other announcements occurring on the same day as a proxy. The more announcements occur, the more an investor is likely to be distracted at any given announcement. Generate a measure of the number of announcements occurring on day t and generate a dummy variable for days with above-median number of announcements. Test whether on these days there is less immediate response of stock returns $r^{(0,1)}_{t,k}$ and more drift $r^{(3,75)}_{t,k}$ (Use the quantile methodology)

5. **Trading volume I.** I have provided you with data on a trading volume measure, that is, the value of the shares exchanged in a day for a company. It is interesting to examine what happens to volume of trading in response to earning surprises. What do you expect to find? (This is a little unfair, since there is no good theory of trading in financial markets) Denote by $v^{(s,s)}_{t,k}$ the value of the shares of company k traded s days after the day of announcement in quarter t. You will run a specification like:

$$\log\left(v^{(s,s)}_{t,k}\right) - \sum_{u=-20}^{-11} \log\left(V^{u}_{t,k}\right)/10 = \alpha + \varepsilon_{t,k}$$

Notice that the dependent variable is the difference between log volume around the announcement date and log volume the week before the announcement. Why is it important to control for baseline volume? Run the regression for $s = 0$. How do you interpret the estimated $\hat{\alpha}$? Now run the same regression for $s = -2, -1, 1, 2, 3, 4, 5$. How are trading patterns around announcement date? What does this suggest about the diffusion of information after the announcement? Why is this pattern different from the pattern for returns?

6. **Trading volume II.** We now look at the increase in abnormal volume as a function of the earning surprise. Run a specification controlling for the 11 quantiles of the earnings surprise (omit quantile 6, it's easiest to interpret the coefficients):

$$\log\left(v_{t,k}^{(s,s)}\right) - \sum_{u=-20}^{-11} \log\left(V_{t,k}^{u}\right)/10 = \alpha + \sum_{d=1}^{11} \phi_d s_{t,k}^{d} + \varepsilon_{t,k}$$

   What are the results for $s = 0$ (same day increase in volume)? How does the volume response vary depending on the earnings surprise, that is, what are you finding on the ϕ_ds? What are the interpretations of this result in terms of attention and information content? Do the results on the ϕ_ds vary for $s = 2$ or $s = 5$ (two or five days later)?

7. **Response over time to earnings announcement.** Consider specification (3) with the usual sample restriction and surprise measure 1 and net returns $r_{t,k} - r_{t,m}$ as the dependent variable. Now we focus on when stock prices react to the news contained in the earning announcement. Repeat the regression with returns at (0,0), (1,1), (2,2), (3,75). Is the coefficient ϕ significantly positive for the (2,2) horizon? How about for the (3,75) horizon? How do you interpret the results? Now do the regression with returns at (-1,-1), (-2,-2), and (-30,-3). Do you find any positive coefficients? What does this suggest about the possibility that the part of the information contained in the earning surprise was leaked to the market in the days before the announcement?

8. **Different surprise measures I.** Construct measures s^1 , s^2, and s^3 in the dataset. What is the average for each measure? (use SUM) How high is their correlation? (use PWCORR). Now consider the distribution of these measures. (Use SUM VARNAME,D) Does it seem that the variables have extreme outliers? Construct variables obtained from s^1, s^2 and s^3 by trimming (dropping) 2 percent on either tail of the distribution. What is the correlation between the trimmed measures?

9. **Different surprise measures II.** Reestimate specification (3) using the measures 2 and 3 of earning surprises (usual sample restrictions). In which specification is the R^2 higher? Compare the third measure with the other two. Notice that the third measure does not use at all the forecasts of analysts. Do the analyst forecasts help in increasing the explanatory power? What happens if you run a specification with all three surprise measures in it? Do they all remain significant predictors?

10. **Time-varying effects and measurement error.** We now explore a different aspect of the findings in point 1. Break down the sample in three time periods, 1984-1989, 1990-94, and 1995-2002 and re-run specification (3). Notice that the coefficient ϕ of returns (0,0) on earning surprises is quite a bit higher in the later than in the earlier period. How about returns at (-1,-1)? How would you explain this? Part of the explanation is measurement error in the date of announcement. A team of Berkeley undergrads used newswires to locate the exact time of the announcement for about 1,500 announcements. This information is recorded by the variable tn. Compare the variable tn to the (reported) date of announcement in IBES, as recorded by the variable t. How close are the two dates for the pre-1990 and the post-1990 period? Argue that measurement error in the date can explain part of the differences in the results of the return regressions in the three different periods.

11. **Open-ended.** Have you noticed any other interesting phenomenon in the data? Write about it. Is this related to a feature of the trading environment, to an informational story, to a behavioral story? Any general lessons?

4 Names of Variables

Brief explanation of variables. In square parentheses are the ones that you will not need for the problem set

T - Date of earning announcement

[TC and TI - Date of earning announcement according to Compustat and IBES respectively]

NEST-number of analysts following stock

STDEST-standard deviation of analist forecasts about earning announcement

MEDEST-Median earning forecast (IBES)

MEDACT-Earning announcement (IBES)

CONAME-Company name

[GAAP-Earning announcement (Compustat)]

SICCODE-SIC code of company making announcement

PERMNO-Identifier number of company making announcement (CRSP)

RAWWIN*-Raw return of stock k on Window * around earning announcement

NETWIN*-Return of stock k on Window * around earning announcement minus aggregate stock

CARWIN*-Return of stock k on Window * around earning announcement minus β * aggregate stock

Window Explanation: Type SUM CARWIN*,D. (0,1) for example means return between the announcement day and the next day.

LAGPRICE-Price of a share of company k right before announcement

LAGSHR-Number of shares outstanding of company k right before announcement

VOLU*-Volume of shares of company k traded (in \$) on Window * around announcement day. VOLU31 is volume traded on announcement day, VOLU32 is volume traded on the trading day follwoing the announcement day, etc.

VOLUA*-Total volume of shares traded (in \$) on Window * around announcement day

Time indicators

5 Code for double clustering (by Evgeny Yakovlev)

* a b c d - there are your variables in the regression

* permno and t are the two variables you want to cluster on. Compute permno_t as a variable that indicates unique combinations of the two

* For example, permno*100,000+t

local varlist "a b c d"

*obtaining b

reg ‘varlist’, cl(permno)

```
        matrix b = e(b)
        matrix list b
    *obtaining variance
        foreach X of varlist permno t permno_t {
    * Here substitute the 'nl' command
        reg `varlist', cl(`X')
            estat vce
        matrix VCE`X'=e(V)
        }
    matrix VCE=VCEid+VCEyear-VCEyear_id
    matrix list VCE
```

*This gives you the right variance-covariance matrix, you need to compute the square root of the relevant variances to find the s.e.s of the coefficients

```
    ereturn post b VCE
```

References

[1] Bernard, Victor L. and Jacob K. Thomas. "Post-Earnings-Announcement Drift: Delayed Price Response or Risk Premium?", *Journal of Accounting Research,* Vol. 27, 1-36, 1989.

[2] Chan, Louis, Narasimham Jegadeesh, and Josef Lakonishok. "Momentum Strategies", *Journal of Finance,* Vol. 51, 1681-1713, 1996.

[3] Degeorge, Francois, Patel, Jay, and Zeckhauser, Richard. "Earnings Management to Exceed Thresholds," *Journal of Business*, 1999.

[4] DellaVigna, Stefano and Joshua Pollet. 2009. "Investor Inattention and Friday Earnings Announcements" *Journal of Finance.*

[5] Hirshleifer, David A., Sonya S. Lim, and Siew Hong Teoh. 2009. "Driven to Distraction: Extraneous Events and Underreaction to Earnings News" *Journal of Finance.*