Forecasting and Predictive Analytics

Homework 1

October 2017

One group will have to present their results in front of the class. So be prepared to present your results simply (using a PowerPoint presentation or directly from the Excel File)

Please work in groups of 3 students and hand in a set of a few slides that present your results. What I am waiting for here is not the actual result but that you make your own hypotheses to solve the problem at hand. You will have to make simplifying choices but you need to understand what you do. If you cannot do everything for lack of time, just do a part of it but do it well.

1 Forecasting

The question we want to ask this week is the following **:**

Is there evidence of stock return predictability?

To complete this exercise, you will need to use in turn two different variables. A question that we will ask later in the course is whether to use the *level* of a variable (i.e. the price) or its *rate of growth* (the return). Here we focus on the return and also on the squared returns. We focus on weekly returns on the Google stock price.

1.1 Forecasting Models

We will generically denote by y_t the variable to forecast.

We will see next lecture three types of models that could be useful in forecasting. Each of them depends on a parameter (in parenthesis):

- **uncentered moving averages:** the forecast $\widehat{y}_{T+1|T}$ of y_{T+1} made at time T is the average over the last k periods:

$$\widehat{y}_{T+1|T} = \frac{y_T + y_{T-1} + ... + y_{T-k+1}}{k}$$

  k is the window of the moving average and has to be chosen or estimated.

- **exponential smoother:** the forecast is a weighted average of all values so far:

$$\widehat{y}_{T+1|T} = (1-\alpha)\sum_{j=0}^{\infty}\alpha^j y_{T-j}$$

where in practice the sum is truncated because we do not observe an infinite number of observations. α is called the smoothing coefficient and needs to be chosen or estimated. Notice that

$$\begin{aligned}\widehat{y}_{T+1|T} &= (1-\alpha)\,y_T + \alpha\,(1-\alpha)\sum_{j=0}^{\infty}\alpha^j y_{T-1-j}\\ &= (1-\alpha)\,y_T + \alpha\widehat{y}_{T|T-1}\end{aligned}$$

so it is easier to update the forecast, given the previous forecast (starting with $\widehat{y}_{1|0} = y_1$).

- **autoregressive model of order 1:** this model, denoted AR(1) is

$$y_t = \tau + \rho y_{t-1} + \epsilon_t$$

where ϵ_t denotes a white noise (i.e. here we'll consider it unpredictable). τ is the intercept and ρ the slope or autoregressive coefficient. The forecast is

$$\widehat{y}_{T+1|T} = \tau + \rho y_T$$

and (τ, ρ) needs to be chosen or estimated.

QUESTION: we discussed the difference between the "Data Generating Process" (DGP) for y_t and our proposed *model* for it. Can you think of the DGPs for which it makes sense to use each of the three models? Ideally we would like to know for which DGP each model produces the optimal forecasts, but it's a bit early in the course to find an answer.

1.2 Estimation, Forecasting, Evaluation

We need to find a way to choose or estimate the parameter(s). We refer to "choosing a parameter" or "calibrating" it when the parameter values are not chosen by a proper estimation method that finds the optimal values based on the sample available (where "optimal" needs to be defined according to some criterion). Each of the model considered depends on one parameter (if we fix $\tau = 0$). To simplify the analysis, we have modified the original data by substracting the "recursive mean", i.e. the variable of interest is $y_t = z_t - \frac{1}{T_0}\sum_{j=1}^{T_0} z_j$, where z_t denotes the original data. We only do this to reduce the number of parameters to estimate.

One of the commonest methods for assessing the forecasting power consists in dividing your sample into two subsamples, say of 75% and 25% of the observations: the first 75% will be used

to estimate the model; the last 25% will be used to assess the forecasting power (validation). In practice, one would like to choose several such sample splits (80/20....) to see how much the results depend on these splits. Of course, it also depends on the number of observations in the sample.

It is possible to estimate parameters so as to minimize the in-sample mean-square error (MSE) of the residuals. The latter is defined as follows: assume that the model depends on parameter θ and provides a forecast $\widehat{y}_{t|t-1}(\theta)$ of y_t for $t = 1, ..., T_0$, where $T_0 = 0.75 \times T$ (or the integer part thereof). The MSE is

$$\mathsf{MSE}(\theta) = \frac{1}{T_0} \sum_{t=1}^{T_0} \left(y_t - \widehat{y}_{t|t-1}(\theta) \right)^2$$

we can find $\widehat{\theta} = \arg\min_{\theta} \mathsf{MSE}(\theta)$. In Excel, the solver[1] can do it simply. Note that for the moving average, the parameter needs to be an integer, this is imposed in my excel file (attached) but you may otherwise simply try several values. (Also pay attention on how to forecast with the various models, do not necessarily trust the attached excel file).

Once you have a parameter estimate, you may forecast y_t over $t = T_0 + 1, ..., T$ and compute the resulting Mean-Square Forecast Error (MSFE). Then compare the models. You will also notice that the parameter estimates that you first obtained may be modified.

2 What to think about and do

1. Think that the forecast should be true ex-ante forecast; i.e. never use the data that you want to forecast. Then only will you be able to see whether the methodology is appropriate.

2. You need to understand what the Excel spreadsheet does, so make sure that you understand exactly the functions and the output. **You need to modify this spreadsheet to suit your assumptions.**

3. If the model is stationary (this means if the distribution is constant over time), the in-sample residuals should constitute a good gauge of the out-of-sample forecast error. So the MSE and the MSFE should be close. If not (either way, if one is larger than the other), then it means that the model is probably not so good at forecasting; or that at least we cannot predict if it will perform well or not.

    - This means for instance that the smoothed value of y_t should constitute the forecast of y_{t+1} not of y_t

[1] It is very easy to use. First you need to load it via Excel Options/ Add Ins/ Analysis Toolpak. You will see that it appears in Excel in data/Analyze.You simply need to find the target cell (here the MSE), choose what you want it to be (here min) and then specify the cell(s) that need be adjusted. You may also impose restrictions (such that a parameter be ≤ 1). Press F1 for help.

- You may want to try with a model like $y_t = \alpha y_{t-k} + \varepsilon_t$ with $k = 2$ or 3 instead of 1.

4. There is no absolute measure of forecast performance based on the MSFE, this can only be specific to the process at hand. Hence you must compare the measure in-sample (where you estimate parameters) and out-of-sample (the posterior subsample).

   - The analysis of the forecast performance should hence be done twice, on the the 'estimation' subsample and on the 'evaluation' one.
   - The ratio of the MSFE over the MSE (when computed in a similar manner) gives a measure of the stability of the forecasting performance. This ratio is a variant of the so-called Theil statistic.

5. Combining forecasts (such as averaging across models) often proves a good strategy. This is an empirical feature often found but people haven't come up with good explanations (why should averaging different methods be better than using the best one overall?).

6. Think about the horizon, you can either produce forecasts at

   (a) a fixed horizon, i.e., fixing h and obtaining $\widehat{y}_{T_0+j+h|T_0+j}$ for $j = 0, ..., T - T_0 - h$

   (b) a variable horizon, i.e., obtaining $\widehat{y}_{T_0+j|T_0}$ for $j = 1, ..., T - T_0$

   Either type of forecast is possible, but it does not require the same dataset when you compute the actual forecasts, and also means you need to be very careful when you assess the quality of the forecast.

7. At the end of the first set of Slides on the website, there is a slide about Mincer-Zarnowitz regressions and the Diebold-Mariano test. Try figuring out what they are about and use them here to assess the quality of your forecasts.