Homework 3: Past Performance, Future Results

36-402, Spring 2017

Due at 11:59 pm on Wednesday, 8 February 2017

> AGENDA: More practice with cross-validation and with smoothing; baby steps in using simulation to see how a model behaves and to do hypothesis testing; reinforcement that "the variable matters" $\neq$ "the coefficient on the variable is statistically significant".
>
> WARNING: Some parts of this assignment are very computation-intensive. Start soon, cache your results, and do not wait to the last minute to make a first submission.

A corporation's **earnings** in a given year is its income minus its expenses[1]. The **return** on an investment over a year is the fractional change in its value, $(v_{t+1} - v_t)/v_t$, and the average rate of return over k years is $[(v_{t+k} - v_t)/v_t]^{1/k}$. Our data set this week looks at the relationship between US stock prices, the earnings of the corporations, and the returns on investment in stocks, with returns counting both changes in stock price and dividends paid to stock holders.[2]

Specifically, our data contains the following variables:

- Date, with fractions of a year indicating months
- Price of an index of US stocks (inflation-adjusted)
- Earnings per share (also inflation-adjusted);
- `Earnings_10MA_back`, a ten-year moving average of earnings, looking backwards from the current date;
- `Return_cumul`, cumulative return of investing in the stock index, from the beginning;
- `Return_10_fwd`, the average rate of return over the next 10 years from the current date.

"Returns" will refer to `Return_10_fwd` throughout.

[1] Accountants get into subtle issues about whether to include in expenses taxes, interest paid on loans, and charges for depreciation of assets and amortization of investments. Those of you who get jobs with certain kinds of tech company will grow only too familiar with these wrinkles. In our data set, earnings are very definitely after all these expenses.

[2] Nothing in this assignment, or the solutions, should be taken as financial advice.

1. *Inventing a variable*

   (a) (1) Add a new column, `MAPE`, to the data frame, which is the ratio of `Price` to `Earnings_10MA_back`. It should have the following summary statistics:

```
     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
 4.785  11.710  15.950  16.550  19.960  44.200     120
```

   Why are there exactly 120 NAs?

   (b) (1) Linearly regress the returns on `MAPE` (and nothing else). What is the coefficient and its standard error? Is it significant?

   (c) (1) What is the MSE of this model, under five-fold CV?

2. *Inverting a variable*

   (a) (3) Linearly regress the returns on 1/`MAPE` (and nothing else). What is the coefficient and its standard error? Is it significant? (For full credit, do not add a new column to the data frame, or create a new vector.)

   (b) (1) What is the five-fold CV MSE of this model? How does it compare to the previous one?

3. *Employing a variable* A simple-minded model[3] says that expected returns over the next ten years should be exactly equal to 1/`MAPE`.

   (a) (1) Find the in-sample MSE of this model.

   (b) (2) Explain why the in-sample MSE is an unbiased estimate of the generalization error for this particular model.

   (c) (2) Make a $Q - Q$ plot for the residuals of this model. *Hint:* try subtraction, rather than `residuals`.

   (d) (5) Estimate a t distribution from the residuals. Report the parameters and their standard errors. Plot a histogram of the residuals, and add the estimated t density. *Hint:* see the function `fitdistr` in the `MASS` package.

4. (5) Use `npreg` to estimate a kernel regression of the returns on `MAPE`. What is the bandwidth? The cross-validated MSE?

5. *One big happy plot* For this problem, you need to only include one plot, and one paragraph of writing, but make sure you clearly label, with comments, which parts of your code are answers to each question. (This does not mean showing your code in your report.) Also, in this problem, take "line" to mean "straight or curved line, as appropriate". Plotting disconnected points where a line is called for will get partial credit.

   (a) (1) Make a scatter-plot of the returns against `MAPE`.

[3] Assume that: future earnings get added to the value of an investment in the company's stock; that nothing else adds to the value of the investment; and that earnings over the next ten years will be equal to those over the last ten years. Solve for the returns.

(b) (6) Add two lines, showing the predictions from the models you fit in problem 1 and 2.

(c) (1) Add a curve showing the predictions of the simple-minded model from problem 3.

(d) (5) Add a line of the predictions of the kernel regression to the plot from problem 5. Which of the previous models does it most resemble? Is it just a slightly wiggly copy of that model, or does it do something qualitatively different?

6. *Simulating the simple-minded model*

(a) (10) Write a function which simulates the simple-minded model from problem 3. The function should take as inputs (i) a vector of `MAPE` values, and (ii) the three parameters of the t distribution. It should return a two-column data frame, with one column being `MAPE` and the other being `1/MAPE` plus t-distributed noise. The columns should have names which match the names used in the real data frame. Make sure that the output of your function has the right number of rows and columns, and that the `summary` statistics for the two columns are what they should be (at least approximately, in the case of the second column).

(b) (5) Write a function which takes as input a data frame, estimates the same linear model as in problem 2 to that data frame, and returns the coefficient on `1/MAPE`. Check that it works by running it on the original data. Check that it also works when the input comes from your simulation function from 6a.

(c) (7) By repeated simulation, find the probability, under the simple-minded model, of the coefficient on `1/MAPE` being as far from 1.0 (in either direction) as what you found in the data.

(d) (8) You can now report a p-value for testing the hypothesis that this slope is exactly 1.0. Carefully state the null and alternative hypotheses, and give your p-value.

(e) (7) Write a function which takes as input a data frame, estimates the same kernel regression as in problem 4, and returns the vector of fitted values from that regression. Check that it works by running it on your original data. Check that it also works when the input comes from your simulation function.

(f) (8) Create a plot of predicted returns versus `MAPE` for the simple-minded model, as in problem 5c. Add 200 kernel regression curves, fit to 200 simulations of the model. Finally, add the kernel regression curve from the true data, as in problem 5d. (You'll want to manipulate graphics settings.) How plausible is the simple-minded model? Explain your answer by referring to your plot.

Hint/warning: Estimating all the kernel regressions might well take a few seconds per simulation. Write and debug your code here with a smaller number of curves, then increase it for the final version.

7. *More fun with star-gazing*

   (a) (1) Linearly regress the returns on both `MAPE` and `1/MAPE` (without interaction). What are the coefficients? Which ones are significant?

   (b) (1) Linearly regress the returns on `MAPE`, `1/MAPE`, and the square of `MAPE`. What are the coefficients? Which ones are significant?

   (c) (8) Explain what is going on.

RUBRIC (10): The text is laid out cleanly, with clear divisions between problems and sub-problems. The writing itself is well-organized, free of grammatical and other mechanical errors, and easy to follow. Questions which ask for a plot or table are answered with both the figure itself and the command (or commands) use to make the plot. Plots are carefully labeled, with informative and legible titles, axis labels, and (if called for) sub-titles and legends; they are placed near the text of the corresponding problem. All quantitative and mathematical claims are supported by appropriate derivations, included in the text, or calculations in code. Numerical results are reported to appropriate precision. Code is properly integrated with a tool like R Markdown or knitr, and both the knitted file and the source file are submitted. The code is indented, commented, and uses meaningful names. All code is relevant to the text; there are no dangling or useless commands. All parts of all problems are answered with actual coherent sentences, and never with raw computer code or its output.