Homework 3: Past Performance, Future Results

36-402, Spring 2015

Due at 11:59 pm on Monday, 2 February 2015

AGENDA: Trying out bootstrapping for both coefficients and for curves; more practice with smoothing; more practice with cross-validation; more practice separating "the variable matters" from "the variable is statistically significant".

GRADING: The problems add up to 90 points. The remaining 10 points are reserved for style and clarity. (See rubric at the end of this assignment.) This will apply to all future homework until further notice.

WARNING: Some parts of this assignment are very computation-intensive. Do not wait until Monday night to start this.

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.

[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 go on to careers in finance or in certain kinds of start-up 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.

"Returns" will refer to `Return_10_fwd` throughout.

1. (5) *Doing what comes naturally*

   (a) (1) Run four linear regressions for the returns: on `Price`; on `Earnings`; on both `Price` and `Earnings`; and on both variables and their interaction. Report coefficients and standard errors.

   (b) (1) Find (in-sample) R^2 for these four models. Can their R^2's be meaningfully compared? If so, which model is preferred by R^2?

   (c) (3) Use five-fold cross-validation to estimate the generalization error of all four models. Can these be meaningfully compared? If so, which model is preferred by cross-validation?

2. (5) *Inventing a variable*

   (a) (2) 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) (2) What is the R^2 of this new model? What is its CV MSE? Are these better or worse than the models in the previous question?

3. (5) *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) (2) What are the R^2 and the CV MSE of this model? How do they compare to the previous ones?

4. (10) For this problem, you need to only include one plot, but make sure you clearly explain which parts of your code are answers to each question. Also, in this problem, read "line" as "straight or curved line, as appropriate". Disconnected points in place of a line will get partial credit.

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

   (b) (3) Add a line showing the predictions from the model you fit in problem 2.

   (c) (3) Add a line showing the predictions from the model you fit in problem 3. (Again, disconnected points will get partial credit.)

(d) (1) A simple-minded model[3] says that returns over the next ten years should be exactly equal to 1/MAPE. Add a curve showing the predictions of this model.

(e) (1) Explain why the in-sample MSE is an unbiased estimate of the generalization error for the "simple-minded" model of the previous part. That is, why do we not need to do cross-validation to estimate its generalization error?

(f) (1) Based on these plots and your previous results, which model fits best?

5. (10) *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) What do you think is going on?

6. (25) *Bootstrapping a parametric model* In this problem, use the model you fit in problem 3.

(a) (1) What are the conventional 90% confidence limits for the coefficient on 1/MAPE? *Hint:* `confint()`.

(b) (10) Use resampling of residuals to get 90% confidence limits for that coefficient. What are they?

(c) (10) Use resampling of cases to get 90% confidence limits for that coefficient. What are they?

(d) (4) Are these compatible with each other? If so, why? If not, explain which seems best.

7. (10) *More smoothing*

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

(b) (5) Add a line of the predictions of the kernel regression to the plot from problem 4. Which of the previous models does it most resemble? Is it just a slightly wiggly copy of that, or does it do something qualitatively different?

8. (20) *Bootstrapping the kernel regression*

(a) (10) Use resampling of residuals to get a 90% confidence band for the kernel regression. *Advice:* Re-estimating the model on a new bootstrap replicate may take several seconds. De-bug your code first, with a small number of replicates, and then go do something else while the main run happens.

[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) (10) Add the confidence band to the previous plot. Does the band include the line for the best-fitting parametric model you've plotted? What does that tell you about the parametric model?

9. (5, extra credit) *Feature discovery* Run a new kernel regression of returns on price and the 10-year moving average of earnings, but *not* on their ratio `MAPE`. Create a color or three-dimensional plot showing the predictions as a function of both variables. What should this look like if the relevant variable is really `MAPE`?

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. Figures and tables are easy to read, with informative captions, axis labels and legends, and are placed near the text of the corresponding problems. 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 either properly integrated with a tool like R Markdown or knitr, or included as a separate `.R` file. In the latter case, the code is clearly divided into sections referring to particular problems. In either case, 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.