Homework 3: Past Performance, Future Results

36-402, Spring 2025

Due at 6 pm on Thursday, 6 February 2025

AGENDA: More practice with cross-validation and with smoothing; first 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".

ADVICE: There are many sub-parts, but most of them are short. The two computation-intensive problems are 6 and 9. (Q 7 and Q8 can be done before Q 6.) The computations in Q 9b, in particular, may be time-consuming to run. Start early, cache your results, and do not wait to the last minute to ask for help (or make a first submission for partial credit).

A corporation's **earnings** in a given year is its income minus its expenses[1]. The **rate of return** or just **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 (`http://www.stat.cmu.edu/~cshalizi/uADA/25/hw/03/stock_history.csv`) 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 assets wearing out ("depreciation") and past investments ("amortization"). 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.

"Returns" will refer to `Return_10_fwd` throughout. If you need an algebraic symbol for this, call it R_t.

A basic but widely-used model of financial assets says that the expected value of the returns should be *exactly* equal to one over the historical average ratio of price to earnings[3], in symbols $\mathbb{E}[R_t|M_t = m] = \frac{1}{m}$, where M_t is the average ratio of price to earnings. This basic model does not say what the distribution of the noise around this relationship should be.

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  11.708  15.947  16.554  19.959  44.196     120
   ```

   Why are there exactly 120 NAs?

   (b) (1) Linearly regress the returns on `MAPE`. What is the coefficient and its standard error?

   (c) (2) What is the MSE of this model, under five-fold CV?

2. *The basic model* Remember that the basic model says that the expected returns over the next ten years should be *exactly* equal to 1/`MAPE`.

   (a) (1) Write the basic model out algebraically, in the form of a linear regression model, and explain why the theory fixes both the slope and the intercept.

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

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

   (d) (2) Make a $Q-Q$ plot for the residuals of this model. *Hint:* try subtraction, rather than `residuals`, and be careful about NAs.

   (e) (2) Do the residuals look Gaussian?

3. *The basic model generalized*

   (a) (2) Linearly regress the returns on 1/`MAPE` (and nothing else). What is the slope coefficient and its standard error? (For full credit, do not add a new column to the data frame, or create a new vector.) For future reference, call this slope $\hat{\beta}_1$. *Hint:* `lm(y ~ 1/x)` doesn't linearly regress Y on $1/X$, but `lm(y ~ I(1/x))` does.

[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, on average, to those over the last ten years. Now solve for the expected returns.

(b) (2) What is the five-fold CV MSE of this model? How does it compare to the model in Q1, and to the basic model?

4. *More fun with star-gazing*

   (a) (1) Linearly regress the returns on `MAPE` (as in Q1). Is the coefficient on `MAPE` statistically significant?

   (b) (1) Linearly regress the returns on `1/MAPE` (as in Q3). Is the coefficient on `1/MAPE` statistically significant?

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

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

   (e) (4) You should have found that the some predictor variables are statistically significant in some regressions but not others. Explain what is going on. Can you decide which variables matter, and which do not, based on these results?

5. *Testing a model (I)*

   (a) (4) Refer back to the linear regression model from Q3. Suppose we want to test, in this model, whether the slope $\beta_1 = 1$. Explain how we would do so. Also explain how this relates to testing whether the basic model is right.

   (b) (4) What you found in Q 2d and Q 2e means that you shouldn't rely on R's usual calculations of statistical significance here. Explain why.

   (c) (4) Explain how your results so far suggest that the basic model should be $R_t = \frac{1}{M_t} + \epsilon_t$, where ϵ_t follows a non-Gaussian distribution.

6. *Testing a model (II)* You can show code here in 6a, 6b and 6c.

   (a) (3) Explain what the following function will do when given a vector of values:

```
resample <- function(x) { sample(x, size=length(x), replace=TRUE) }
```

   (b) (5) Write a function which simulates the basic model in the form given in Q5c. The function should take as inputs (i) a vector of `MAPE` values, and (ii) the vector of residuals from Q2d. It should return a two-column data frame, with one column being `MAPE` and the other being `1/MAPE` plus noise coming from resampling the residuals. 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).

(c) (4) Write a function which takes as input a data frame, estimates the same linear model as in Q 3 on 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 6b.

(d) (4) Running the function from Q 6c on the simulation from Q 6b gives a random slope $\tilde{\beta}_1$. By repeated simulation, find the probability, under the basic model, of the coefficient on `1/MAPE` being at least as far from 1.0 (in either direction) as what you found in the data, i.e., $\mathbb{P}(|\tilde{\beta}_1 - 1| \geq |\hat{\beta}_1 - 1|)$.

(e) (5) 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 the test statistic, and give your p-value.

7. (5) Use `npreg` to estimate a kernel regression of the returns on `MAPE`. What is the bandwidth? The cross-validated MSE? How does it compare, in predictive accuracy, to the models you have already considered?

8. *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`.

   (b) (1) Add a curve showing the predictions of the basic model from Q 2.

   (c) (2) Add two lines, showing the predictions from the models you fit in Q 1 and Q 3.

   (d) (3) Add a line of the predictions of the kernel regression to the plot from Q 8. 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?

9. *Testing a model (III)*

   (a) (5) Write a function which takes as input a data frame, estimates the same kernel regression as in Q 7, 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. (You can show code here.)

   (b) (5) Create a plot of predicted returns versus `MAPE` for the basic model, as in Q 8b. Add 100 kernel regression curves, fit to 100 simulations of the model. Finally, add the kernel regression curve from the data, as in Q 8d. (You'll want to manipulate graphics settings.)

Hint/warning: Estimating the kernel regressions might well take a few seconds per simulation. Write and debug your code here with a small number of curves, then increase it for the final version. (You will get partial credit if you use less than 100 simulations here.)

(c) (5) Compare the kernel regression curve you got from the data to the kernel regression curves you fit to simulations of the basic model. Does the true-data curve look like the simulation curves, or does it differ from them in some way? If it does, does this indicate a problem for the basic model?

10. (5) Do your results support the general idea that expected returns are *roughly* inversely proportional to `MAPE`? Do they support the basic model that the expected returns are *exactly* inversely proportional to `MAPE`? Justify your answers by referring to Q 9c and 6e (and perhaps other problems, as you think appropriate).

11. (1) Roughly how long did you spend on this assignment?

PRESENTATION 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. 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 plots and tables are generated by code included in the R Markdown file. 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. All parts of all problems are answered with actual coherent sentences, and raw computer code or output are only shown when explicitly asked for. Text from the homework assignment, including this rubric, is included only when relevant, not blindly copied.

(In Gradescope, assign *all* pages to this rubric.)