Homework 4

Instructions

Due: 5:00pm on Wednesday, October 6th

1. Add your name between the quotation marks on the author line in the YAML above.
2. Compose your answer to each problem between the bars of red stars.
3. Commit your changes frequently.
4. Be sure to knit your .Rmd to a .pdf file.
5. Push both the .pdf and the .Rmd file to your repo on the class organization before the deadline.

Theory

Problem 1

Based on ISLR Exercise 5.2

We will now derive the probability that a given observation is part of a bootstrap sample. Suppose that we obtain a bootstrap from a set of n observations. Let j be an arbitrary integer between 1 and n.

a. What is the probability that the first bootstrap observation is *not* the jth observation from the original sample? Justify your answer.

b. What is the probability that the second bootstrap observation is *not* the jth observation from the original sample?

c. Argue that the probability that the jth observation is *not* in the bootstrap sample is $(1 - 1/n)^n$.

d. Approximate the probability that the jth observation is *not* in the sample in terms of the constant e.

e. For each of $n = 5, 10, 100$, what is the probability that the jth observation *is* in the bootstrap sample?

f. Create a plot that displays, for each integer n from 1 to 100, the probability that the jth observation is in the bootstrap sample. Comment on the results of your plot.

g. Use the `sample` function to create a bootstrap sample of the numbers 1 through 100. Then use either `dplyr` or base `R` code to assess whether the bootstrap sample contains the number 1. Repeat a total of 10000 times and compute the proportion of times the number 1 appears in your bootstrap samples. Compare to the results of the previous parts.

Problem 2

For each of the three scenarios listed below, determine which of the following techniques would be most appropriate to implement in order to achieve the desired result. Briefly justify your answer.

Techniques

(i) Use a single randomly selected validation set to estimate test MSE.
(ii) Use summary statistics like R^2 and RSE from the linear model summary table to assess model accuracy.
(iii) Use k-fold cross validation to estimate test MSE.
(iv) Using bootstrapping to estimate a statistic's bias and variance.
(v) None of these methods are appropriate.

Scenarios

a. A statistics professor wants to create a model exploring the relationship between midterm exam scores and final exam scores for a statistics class. The professor has data for 100 students over the past year, and is interested in finding the best of several polynomial models to predict final score as a function of midterm score.

b. Researchers are interested in building a model to predict house prices in the Woodstock neighborhood as a function of the house's square footage. Currently, they have data for 10 houses and want to assess whether the non-zero correlation they observed between price and size is likely to just have occurred by random chance.

c. A data set contains observations on 50,000 samples of red and white wines from the north of Portugal. Researchers are interested in building a model to predict wine quality (on scale 1 - 10) based on physicochemical data for each sample. The research want to compare the performance of two complicated models, each of which takes significant time to code and compute.

Applied

Probelm 3

Based on ISLR Exercise 5.8

We will now perform cross-validation on a simulated data set.

a. Use the following code chunk to generate a simulated data set. Then write out the explicit equation for the model used to generate the data.

```
set.seed(1010) #Change this to your favorite number
x <- rnorm(100, 0, 1) ## Generates 100 variables from N(0,1)
e <- rnorm(100,0, 1) ## Generates 100 errors from N(0,1)
y <- x- 2*x^2 + e ##Specifies Y as a function of X plus errors

sim_data<-data.frame(x,y) ## Creates a data frame of X and Y
```

b. Create a scatterplot of X against Y. Describe the relationship observed. Calculate mean and standard deviation for each of X and Y, as well as the correlation between X and Y.

c. Set a seed and compute LOOCV errors from fitting each of the following 4 models using least squares:

d. $Y = \beta_0 + \beta_1 X + \epsilon$

ii. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$

iii. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$

iv. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \epsilon$

v. Repeat part c using a different seed and report your results. How do they compare to your answer from part c? Explain why this occurred.

e. Set a seed and compute 5-fold CV from fitting each of models in part c.

f. Set a difference seed from the previous part and again compute 5-fold CV from fitting each of the models in part c. How does your answer compare to part e. Explain why this occurred.

g. Which of the models in c. had the smallest LOOCV? Explain why this makes sense.

h. Comment on the statistical significance of the coefficient estimates that results from fitting each of the models in c. using least squares. Do these results agree with the conclusions drawn based on the cross-validation results?

Problem 4

In this problem and the next, we look 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. A corporation's **earnings** in a given year is its income minus its expenses. 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

$$[(v_{t+k} - v_t)/v_t]^{1/k}.$$

Read this data from the csv in the accompanying data folder:

```
stocks <- read.csv("data/stocks.csv")
```

The dataset 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.

Inventing a variable

a. Add a new column, `MAPE` to the data frame, which is the ratio of `Price` to `Earnings_10MA_back`. MAPE stands for the *monetary-adjusted price-earnings* ratio, and represents the number of years it would take to recoup the cost of a share, assuming the earnings stayed constant at their current level.

Bring up the summary statistics for the new column using the `summary()` command. Why are there exactly 120 NAs? For ease of computing for the rest of the lab, you should remove all rows with any missing data.

b. Build a simple linear model to predict Returns as a function of `MAPE`. What is the slope of the model and its standard error? Is it statistically significant?

c. What is the MSE of this model under 5-fold CV?

Inverting a variable

d. Build a simple linear model to predict Returns as a function of `1/MAPE` . What is the slope of the model and its standard error? Is it statistically significant?

e. What is the CV MSE of this model? How does it compare to the previous one?

A simple model

A simple-minded model says that the expected returns over the next ten years should be exactly equal to `1/MAPE` (i.e earnings / price)

f. Find the *training* MSE for this model.

g. Explain why the training MSE is equivalent to the estimate of the test MSE that we would get through five-fold CV.

Problem 5

Is simple sufficient?

The model that we fit in part 4d is very similar to the simple-minded model. Lets compare the similarity in these models. We could go about this in two ways. We could *simulate* from the simple-minded model many times and fit a model of the same form as 4d. to each one to see if our observed slope in 4d. is probable under the simple-minded model. We could also *bootstrap* the data set many times, fitting this model each time, then see where the simple-minded model lays in that distribution. Since we already have practiced with simulation, lets do the bootstrap method.

a. Form the bootstrap distribution for the slope of `1/MAPE`. Plot this distribution with the parameter of interest (the slope corresponding to the simple-minded model) indicated by a vertical line.

b. What is the approximate 95% bootstrap confidence interval for the slope? How does this interval compare to the one returned by running `confint()` on your model object from 4d. of Problem 4? Explain any differences you observe.

One big happy plot

c. Make a scatterplot of the returns against `MAPE`. Add two curves showing the predictions from the models you fit in 4b. and 4d. Add a line showing the predictions from the simple-minded model.

The big picture

d. **Cross-validation for model selection**: using CV MSE, which model would you select to make predictions of returns? Looking at the plot in part c., does this seem like a good model? What are its strengths and weaknesses for prediction?

e. **Bootstrapping for uncertainty estimation**: based on your bootstrapping procedure for the slope of the linear model using `1/MAPE` as a predictor, is the simple-minded model a plausible model given our data?