Homework 3

Advanced Methods for Data Analysis (36-402/36-608)

Due Thurs February 13, 2014 at 11:59pm

Instructions: each homework will be submitted in two parts.

1. *Writeup.* An important part of the learning the data analysis trade is learning how to communicate. Prepare a writeup to the questions below and work hard to make your submission as readable as you possibly can—this means no raw R code as part of the writeup, unless you find it necessary for clarity; no raw R output with 12 significant digits, etc., unless again you deem it necessary; making sure figures are of proper (large) size with labeled, readable axes; and so forth. Your submitted writeup must be in PDF format. You can produce this PDF using LaTeX, Word, or whatever you want, but at the end of the day, it must be a PDF. Any other type of file will not be accepted.

2. *Code.* Also submit your R code for the homework as a plain text file. You should demarcate sections of the code that correspond to the different homework problems using clearly visible comments (e.g., as in `##### Problem 1 #####`).

1 Minimizing investment variance

(a) Given random variables X, Y, prove that the value of $\theta \in \mathbb{R}$ that minimizes

$$\mathrm{Var}\big(\theta X + (1-\theta)Y\big)$$

is

$$\theta = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}},$$

where $\sigma_X^2 = \mathrm{Var}(X)$, $\sigma_Y^2 = \mathrm{Var}(Y)$, and $\sigma_{XY} = \mathrm{Cov}(X, Y)$.

(b) Recall the investment interpretation from lecture, i.e., think of X as the random return of one asset, and Y as the random return of another. If we were to put θ of our money in the first asset, and $1-\theta$ in the second, then $\theta X + (1-\theta)Y$ is the return that we would see; the optimal value of θ is the one minimizing the variance of this return, as given in part (a).

Now explain, in the context of investments: what happens to the optimal value of θ when σ_Y^2 is much larger than σ_X^2? When σ_X^2 is much larger than σ_Y^2? What is the role of the covariance σ_{XY}?

2 Saddle up your bootstraps, yeeeeeeeeeeeeeeha

Consider the set of the first n positive integers $\{1, \ldots n\}$. Suppose that we draw a bootstrap sample of size B from this set, written as $\{i_1, \ldots i_B\}$. In other words, this is sampling B numbers from $\{1, \ldots n\}$ with replacement.

(a) What is the expected number of numbers in $\{1, \ldots n\}$ that do not appear in the random sample $\{i_1, \ldots i_B\}$? (Hint: first determine the probability that a given number, say i, does not appear in $\{i_1, \ldots i_B\}$.) Hence, what is the expected proportion of numbers in $\{1, \ldots n\}$ that do not appear in the random sample $\{i_1, \ldots i_B\}$?

(b) Suppose that $B = n$, i.e., we draw n numbers from $\{1, \ldots n\}$ with replacement. What is now your formula from (a) for the expected proportion of numbers that do not appear in the random sample? What does this approach as $n \to \infty$? (Hint: yes, the title is silly, but it's also a clue...)

3 Abalone for dinner, or abalone for cash?

This problem uses real data on abalone fishing in Australia (taken from `http://archive.ics.uci.edu/ml/machine-learning-databases/abalone`). But the story is made-up (or is it?)...

Professor Tibshirani's friend, call him Bob, is an abalone diving enthusiast. Abalones are large edible sea snails, found in the coastal waters of nearly every continent. Diving for abalones is often done for sport, and has a rich history in some parts of the world like Australia and South Africa. Here is Bob's quandry: Bob has caught an abalone, and he either wants to eat it, or sell its shell for cash. (He cannot do both because preparing its shell for selling requires the use of some kind of cleaning agent that will ruin its meat.) Bob reasons that it will only be worth selling the shell if the abalone is older than 8.5 years old; otherwise, the shell won't be worth much at all, and he'd rather eat the meat.

The precise age of an abalone can be determined by counting the number of rings on its shell and adding 1.5. Bob doesn't quite have the eye for this, but he knows that its age should be related to its diameter. He has measured the diameter of his particular catch: $d_0 = 215$ mm. Bob is looking for some statistical help to determine the relationship between the diameter and age of abalones, and ultimately determine with he should try to sell or eat his big snail.

To help make the judgement call, Bob has tracked down a data set on abalones from the Marine Resources Division in Tasmania, Australia. It is available as `abalone.csv` on the course website, in the format of a data frame with 4177 and 9 columns. Each row corresponds to a particular caught abalone, and the columns correspond to the following attributes:

```
Name Data Type Meas. Description
----------------------------
Sex nominal M, F, and I (infant)
Length continuous mm Longest shell measurement
Diameter continuous mm perpendicular to length
Height continuous mm with meat in shell
Whole weight continuous grams whole abalone
Shucked weight continuous grams weight of meat
Viscera weight continuous grams gut weight (after bleeding)
Shell weight continuous grams after being dried
Rings integer +1.5 gives the age in years
```

(a) Looking at male abalones only (Bob's is a male), perform a simple linear regression of the age of abalones on the diameter. What are the coefficients (intercept and slope)? Plot this regression line on top of the data points. Does it look like a good fit?

(b) Use the linear model from (a) to predict the age of Bob's abalone, given that it has diameter $d_0 = 215$ mm. Is it older than 8.5 years? Looking at the plot from part (a), do you have any doubts about the accuracy of this prediction?

(c) Using standard linear modeling methodology, construct a 95% confidence interval for the expected age of the abalone at $d_0 = 215$ mm, using the linear model from (a). Does this interval contain 8.5? Again, do you believe it? Why or why not?

(d) Use the `smooth.spline` function to fit a smoothing spline of the abalone age on the abalone diameter (again, just for male abalones). E.g., the call

```
spmod = smooth.spline(x,y)
```

will fit a smoothing spline of y on x. (Don't worry about choosing a tuning parameter value for the smoothing spline; the above call will do it internally using something called generalized cross-validation.) Plot the predicted regression line on top of the data points.

(e) Does this look like a better or worse fit than the linear regression fit from part (a)? Think of a way to provably demonstrate this one way or another, and implement it. (Hint: sounds like "vross-calibration...")

(f) Use the smoothing spline model from (d) to predict the age of Bob's abalone, using the fact that its diameter is $d_0 = 215$ mm. Is it older than 8.5 years?

(g) Now use the smoothing spline model and the bootstrap to construct a 95% confidence interval for the abalone's age at $d_0 = 215$ mm. Does this contain 8.5?

(h) Do you trust more the confidence interval constructed using linear regression, or that using smoothing splines and the bootstrap? Why? Ultimately, what should Bob do—eat or sell?