Homework Assignment 5: Bootstrapping Will Continue Until Morale Improves

36-402, Advanced Data Analysis

Due 15 February 2011

The goal of this homework is to practice using bootstrapping to quantify the uncertainty in regression models.

The data set `cats` in the library `MASS` contains measurements of the total body weight for 47 female and 97 male adult cats, as well as the weights of their hearts. The medical rationale for this experiment was that the dosage of many heart medicines needs to be calibrated to the mass of the heart, and so one wants to know how to predict that from the total body weight.

Load the data, and check the loading, as follows:

```
> library(MASS)
> data(cats)
> summary(cats)
 Sex         Bwt             Hwt       
 F:47   Min.   :2.000   Min.   : 6.30  
 M:97   1st Qu.:2.300   1st Qu.: 8.95  
        Median :2.700   Median :10.10  
        Mean   :2.724   Mean   :10.63  
        3rd Qu.:3.025   3rd Qu.:12.12  
        Max.   :3.900   Max.   :20.50  
```

Body weights (`Bwt`) are in kilograms, and heart weights (`Hwt`) are in grams — a cat's heart is not actually bigger than its entire body.

The goal here is to provide an accurate estimate of the weight of a cat's heart from its body weight, including some measure of uncertainty, and to assess whether the prediction should take account of the cat's sex.

As always, include the code for the computational parts of each problem as an appendix to your report, clearly labeling which block of code goes with which problem.

1. (5 points) Use `lm` to linearly regress heart weight on body weight, without using sex as a predictor variable, and ensuring that the regression line goes through the origin. (That is, forcing the intercept to be zero.) Report the estimated coefficient, the standard errors given by R, and the corresponding 95% confidence interval (using a t-distribution).

Extra Credit: (5 points) Why is it reasonable to force the intercept to be zero in this case?

2. (5 points) Plot the distribution of residuals from the model you fit in Problem 1. Does it look Gaussian? (Explain.)

   Extra credit (5 points): Suggest a formal test of the hypothesis that the residuals are Gaussian. Is the departure from a Gaussian distribution significant at the 5% level? At the 1% level?

3. (5 points) Using `lm`, fit a linear regression model for heart weight, in which body weight *interacts* with `Sex`. Again, ensure that the regression lines for both females and males go through the origin. How many coefficients should there be? Report the estimated coefficients, the standard errors calculated by `lm`, and the corresponding 95% confidence intervals.

4. *Testing the significance of a model expansion*

   (a) (5 points) Describe, briefly and in your own words, a method for formally testing the hypothesis that adding `Sex` as a predictor in the model, as in Problem 3, does not significantly improve over the model you fit in Problem 1. Clearly state the assumptions underlying the test. (Here a "formal test" means one where you calculate both a test statistic and a p-value for the test statistic under the null distribution, e.g., the "Partial F-test" you learned in 36-401.)

   (b) (5 points) Apply this test to the `cats` data and the models from Problems 1 and 3. What is the value of your test statistic? Is the difference significant at the 5% level?

5. We now compare the measures of uncertainty from problem 1, which are calculated assuming Gaussian and homoskedastic noise, with the measures obtained by bootstrapping the data points. *Hint:* Look at section 4.1 of the notes for lecture 8.

   (a) (5 points) Write a function `resample.cats`, to resample the data points in `cats`. It should take no arguments, but produce a new data frame with the same column names as `cats`. Check that it is working properly by running `summary(resample.cats())` and confirming that the result is close to that of `summary(cats)`.

   (b) (5 points) Write a function `fit.cats.1` to re-estimate the coefficient of the model from Problem 1 on a new data frame. It should take a data frame as an argument, estimate the same model as in Problem 1, and return the value of the regression coefficient (a single number, not the whole regression object). Check that it works by confirming that `fit.cats.1(cats)` gives the same number as the regression coefficient you got in Problem 1.

(c) (5 points) Using your functions `resample.cats` and `fit.cats.1` from Problems 3a and 3b, write a function `cats.1.se` to find the bootstrap standard error for the coefficient in this linear model. The function should take the number of bootstrap replicates, and return the estimated standard error. What standard error do you find with 100 replicates? With 1000?

(d) (5 points) Using your functions `resample.cats` and `fit.cats.1` from Problems 3a and 3b, write a function `cats.1.cis` to find confidence intervals for the coefficient in this model. The function should take as arguments the number of bootstrap replicates and one minus the confidence level. It should return the upper and lower confidence limits. What 95% confidence interval do you get with 100 replicates? With 1000 replicates?

(e) (5 points) How do your results in (5c) and (5d) compare to those from (1)? Based on your findings in (2), which set of error estimates seems more trustworthy?

6. *Cross-validation on a model expansion*

   (a) (5 points) Explain, briefly and in your own words, how to use cross-validation to tell whether including `Sex` in the model improves its ability to generalize.

   (b) (5 points) Check, using five-fold cross-validation, whether adding `Sex` to the model improves it.

   (c) (5 points) Can you say whether the cross-validation comparison is significant at the 5% level?

7. *Bootstrap testing of a model expansion.*

   (a) (5 points) Write a function to simulate new data sets from the model you fit in Problem 1 by re-sampling the residuals. The function should take no arguments, but return a data frame with columns `Sex`, `Bwt` and `Mwt`. (*Hint:* Look at section 4.3 of the notes for lecture 8.

   (b) (10 points) Write a function to calculate the test statistic from your hypothesis test in Problem 4. The input should be a data frame, which you can assume has the columns `Sex`, `Bwt` and `Hwt`, and the output should be the value of the statistic. Check your function by seeing that it gives the right value of the test statistic when applied to the original `cats` data frame, i.e., the one you calculated in Problem 4b. *Hint:* Look at Code Example 8 in the notes for lecture 8.

   (c) (10 points) Using the simulator from (7a) and the test statistic calculator from (7b), find a bootstrap p-value for the significance of adding `Sex` as a predictor. Use at least 500 replicates. Is it significant at the 5% level? *Hint:* Look at Code Example 8 in the notes for lecture 8.

8. (10 points) A veterinarian wants to know whether they should adjust for a cat's sex when calibrating how much heart medicine to administer. Based on your findings in problems 4, 6 and 7, what would you recommend, and why?