Lab 4

Statistical Inference, Spring 98

In this lab we explore bootstrapping, fitting linear regression models, performing multiple linear regression, and logistic regression in R.

a) Bootstrapping

"airquality" dataset contains daily air quality measurements in New York city, from May to September 1973. Using this sample, we would like to construct a bootstrap confidence interval for the median value of the Solar Radiation measured in Central Park.

1- Take 1,000 bootstrap samples from the data and calculate the median for each one. Record the results in a vector.

2- Construct a 90% confidence interval using the percentile method for the Solar Radiation bootstrap median distribution. Explain briefly how you constructed the interval, and interpret this interval in the context of the data.

3- Calculate the bootstrap standard error. Note that this is basically the standard deviation of the bootstrap median distribution. Using this value, calculate a 90% confidence interval for the same parameter of interest.

4- Are the two intervals in exercises 3 and 4 approximately equal? Which method do you prefer?

The "inference" function (inference.R) is a custom function that allows you to perform simulation-based statistical inference methods. You can include the function in your R environment by the following command:

```
> source("inference.R")
```

By default, the function takes 10,000 bootstrap samples (instead of the 1,000 you've taken above), creates a bootstrap distribution, and calculates the confidence interval using the percentile method. This could be gain by the following function call:

```
> inference(sample, type = "ci", method = "simulation", conflevel =
0.90, est = "median", boot_method = "perc")
```

5- Calculate bootstrap 90% CI using the "inference" function with both percentile and standard error methods. Compare the results with the intervals that you calculated in exercises 3 and 4. (Study the parameters descriptions of the "inference" function carefully.)

b) Simulation

"ships" dataset ("COUNT" library) contains values on the number of reported accidents for ships belonging to a company over a given time period.

6- We say that if a ship has more than 10 incidents, it counts as a "High" incident ship; Otherwise, it is a "Low" incident one. Create a new categorical variable called "Risk" and append it to the dataset.

7- Does this data provide convincing evidence that "Type B" ships count as "High Risk" ships? To answer this question, perform a simulation using "Inference" function.

8- Is it necessary to perform simulation to prove the above claim? Are the conditions of CLT satisfied in the given setting?

c) Linear Regression, Multiple Linear Regression

“Galton” dataset (“mosaicData” library) contains results of a survey on the height of children and their parents.

9- Plot a scattered dot scheme that visualizes the relation between the height of the children and their fathers. Could it be conducted from the scatter plot that there is a relation between these two variables? Is it possible to state the claim that there exists a linear relation?

10- Find the correlation between “height” and “father”. Is there convincing evidence that a strong linear relationship exists between these two variables?

11- Construct a linear model using “lm” function. Select “height” as the response variable and “father” as the explanatory variable. Find the coefficient of “father” variable. Interpret its meaning?

12- Specify the line equation that fits the data using the result of the constructed linear model in exercise 11. What does the value of “intercept” mean?

13- According to the linear model constructed above, what would be the predicted height of a child if his/her dad has a height of 70”, 75” or 80”? Obtain the results for all three, only with one call of the “predict” function.

14- Construct a multiple linear model using “lm” function that uses “father” and “mother” as explanatory variables and fits into “height” as the response variable. Interpret the coefficients.

15- Plot the histogram of the residuals for the fitted model. Which MLR conditions might be checked by inspecting this plot?

16- Draw QQ-plot of the distribution of the residuals in order to compare it with a standard normal distribution. Which MLR conditions might be checked by inspecting this plot?

17- Plot the residuals in respect to the explanatory variables. Which MLR conditions might be checked by inspecting this plot?

18- Plot the residuals of the fitted model in respect to their indices. Which MLR conditions might be checked by inspecting this plot?

19- Is “height” of a child linearly related to his/her sexuality? Check the relation by setting up a multiple linear model with explanatory variables “father”, “mother” and “sex”. What is the p-value for the coefficient of this variable? Is it statistically significant?

20- Interpret the coefficient of “sex” variable in the model of exercise 19.

21- Perform a model selection using p-value-based backward elimination method considering all the variables of the Galton dataset with the response variable “height”. Point out the details of each step. Express the exact regression formula at the end. Check the MLR conditions for the final linear model.

d) Logistic Regression

“titanic” dataset (“COUNT” library) contains an observation-based version of the 1912 Titanic passenger survival log.

22- Form a logistic regression model on the whole titanic dataset. The response variable is “survived” which is a categorical variable that shows the survival with “yes” or “no”. There is no need to perform a model selection routine, simply choose the variables with significant p-values.

23- Based on the constructed model in exercise 22, what is the probability of survival for an adult man who sat in the 2nd class? Use the “predict” function.

24- Find the probability of survival for all possible cases in the titanic incident. Sort them by the probability of survival. Automate the process as much as you can.

25- Predict the probability of survival for the whole titanic dataset based on the logistic regression model that you have made. We decide that someone have been survived from the accident if the probability of survival for him/her is more than 50%. Now, match the predicted results with the “survived” variable of the dataset. How many cases did your model predicted correctly? State the result in the form of percentage.

26- 47 years later, 1959, the “Hans Hedtoft” ship, known as “Little Titanic”, sunk after striking an iceberg. Assume that the log of the survivals is not available. Can you use your logistic regression model that you have made for Titanic incident to predict Hans Hedtoft survivors? Discuss the conditions and terms that may affect the accuracy of the prediction for the imported model.