Homework Four

Statistics 151a (Linear Models)

Due by 11:59 PM on October 23, 2018

1. a) Suppose $X_1, \dots X_n$ are i.i.d observations from a distribution with known variance σ^2. Describe a bootstrap-based algorithm to compute a 95% confidence interval for σ. (**0.5 points**)

   b) Take $M = 1000$. For each $i = 1, \dots, M$, simulate $n = 100$ observations from a normal distribution with $\sigma = 1$. Construct your confidence interval in the previous part and check if the interval contains the true value $\sigma = 1$. For how many $i = 1, \dots, M$, does your interval contain the true value? (**0.5 points**)

2. Consider the dataset "twoyear.Rdata" available in bcourses. We want to fit a linear model for log(wage) based on the variables jc (number of years in junior college), univ (number of years in university) and exper (number of years in the workforce). I want to test the null hypothesis $H_0 : \beta_1 = \beta_2$ (that the effects of number of years in junior college and number of years in university are the same) against the alternative $H_1 : \beta_1 \neq \beta_2$ at the 95% significance level.

   a) Find the value of the t-statistic for this test. Does the t-test reject the null hypothesis at the 95 % level? (**0.5 points**).

   b) Find the value of the F-statistic for this test. Does the F-test reject the null hypothesis at the 95 % level? (**0.5 points**).

   c) Design a permutation test for testing this hypothesis. Does your test reject the null hypothesis at the 95% level? (**0.8 points**).

   d) Construct a 95 % confidence interval for $\beta_1 - \beta_2$ via bootstrap. Does this interval contain the value zero? (**0.8 points**).

3. Consider the *savings* dataset (from the R package *faraway*) that we used in class. Fit a linear model for the response variable *sr* based on the explanatory variables *pop*15, *pop*75, *dpi* and *ddpi*.

   a) Use R to report the usual normality based confidence intervals for each of $\beta_1, \dots, \beta_4$. (**0.4 points**)

   b) Compute confidence intervals for $\beta_1, \dots, \beta_4$ using residual bootstrap. How do these intervals compare with those in part (a) above? (**0.8 points**).

4. In the Bodyfat dataset, consider the linear model

$$\text{BODYFAT} = \beta_0 + \beta_1\text{AGE} + \beta_2\text{WEIGHT} + \beta_3\text{HEIGHT} + \beta_4\text{THIGH} + e$$

In R, plot the following graphs (**2.7 points = 0.3 for each graph**)

a) Residuals against fitted values.

b) Standardized Residuals against fitted values.

c) Residuals against Standardized Residuals.

d) Predicted residuals against fitted values.

e) Residuals against predicted residuals.

f) Residuals against leverage.

g) Predicted residuals against Standardized Predicted Residuals.

h) Standardized residuals against Standardized Predicted residuals.

i) Cooks Distance against the ID number of the subjects.

Comment on these plots. Based on these plots, assess whether there are any outliers in the dataset; are there any influential observations. (**0.5 points**)

For each subject, calculate the p-value for testing whether the ith subject is an outlier based on the standardized predicted residual. Plot these p-values against the ID number of the subjects. How may of these p-values are less than 0.05? Does it make sense to rule all such subjects as outliers? (**1 points**)

Based on the analysis, does it make sense to fit the linear model with any of the subjects removed? If not, why not? If so, which ones; and in this case, report the summary for the linear model with the subjects removed. (**1 points**)