MA308: Statistical Calculation and Software

Assignment 2 (Oct 11– Nov 08, 2021)

2.1 [18 points] For the *"kid.weights"* dataset from R package "UsingR",

(a) [5 points] First draw two boxplots for the heights of two genderss of children, i.e. male (M) and female (F), put two boxplots side by side in one figure. What will be the conclusion for testing the height of all the children at $\alpha = 0.05$ level of significance,

$$H_0 : \ \mu = 36, \ \text{v.s.} \ H_1 : \ \mu \neq 36, \tag{2.1}$$

with unknown variance? What if the variance is known to be the current sample variance?

(b) [5 points] Carry out the likelihood-ratio test in (2.1) for male with unknown variance and draw the conclusion at $\alpha = 0.05$ level of significance. Compare the result with that of using *t-test*.

(c) [4 points] Test whether the height of the male and female have the same mean value at $\alpha = 0.05$ level of significance. ~~What if there is a "pairing" between the control and treatment1 group?~~

(d) [4 points] Test whether the spread of height for the male and female are the same or not.

2.2 [59 points] This question should be answered using the **Carseats.csv** data set.

(a) [4 points] Test whether *Sales* follow normal distribution.

(b) [5 points] Fit a multiple regression model to predict *Sales* using *Price*, *Advertising*,*Age*, and *Urban*.

(c) [5 points] Provide an interpretation of each coefficient in the model. Be careful some of the variables in the model are qualitative!

(d) [4 points] Write out the model in equation form, being careful to handle the qualitative variables properly.

(e) [4 points] For which of the predictors can you reject the null hypothesis $H_0 : \beta_j = 0$?

(f) [4 points] On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.

(g) [5 points] How well do the models in (b) and (f) fit the data?

(h) [5 points] Using the model from (f), obtain 95% confidence intervals for the coefficient(s).

(i) [6 points] Is there evidence of outliers or high leverage observations in the model from (f)?

(j) [6 points] There is an indicator *"Urban"* in the *"Carseat"* data set, compare the mean *Sales* of the *"Urban"* area with that of the *"Non-Urban"* area, show the results of the likelihood ratio test and the Mann-Whitney test for testing the equality of these two mean values. Can we use the Wilcoxon's Signed-Rank test? Why?

(k) [5 points]Fit a multiple regression model to predict *Sales* using all the other variables, implement variable selection by *stepwise methods* and *all-subsets regression.*

(l) [6 points] Consider using all the other variables to predict *Sales*, find out the most important variable in predicting *Sales* via the concept of *Relative Importance*, compare with the results in **(k).**

2.3 [23 points] This question should be answered using the **weekly.csv** data set.

(a) [5 points] Produce some numerical and graphical summaries of the *Weekly* data. Do there appear to be any patterns?

(b) [6 points] Use the full data set to perform a logistic regression with *Direction* as the response and the five lag variables plus *Volume* as predictors. Use the

summary function to print the results. Do any of the predictors appear to be statistically significant? If so, which ones?

(c) [6 points] Compute the confusion matrix and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of mistakes made by logistic regression.

(d) [6 points] Now fit the logistic regression model using a training data period from 1990 to 2009, with *Lag2* as the only predictor. Compute the confusion matrix and the overall fraction of correct predictions for the held out data (that is, the data from 2010).