**STAT 2600 *Exam3 from 4pm to 5:30pm on 4/19* NAME: BRYAN GREENER**

**In order to get full credit, you must use R to answer the problems along with the R codes used. Download ‘final.csv’ file. Write your answer right after each subproblem.**

Data contains sales of child car seats at 400 different stores and consists of the following 11 variables:

|  |
| --- |
| ***Sales***: Unit sales (in thousands) at each location  ***CompPrice***: Price charged by competitor at each location  ***Advertising***: Local advertising budget for company at each location (in thousands of dollars)  ***Population***: Population size in region (in thousands)  ***Price***: Price company charges for car seats at each site  ***Age***: Average age of the local population  ***Education***: Education level at each location  ***ShelveLoc***: A factor with levels Bad, Good and Medium indicating the quality of the shelving location for the car seats at each site  ***Urban***: A factor with levels No and Yes to indicate whether the store is in an urban or rural location  ***US***: A factor with levels No and Yes to indicate whether the store is in the US or not  ***Income***: A factor with levels low, middle and high indicating community income level |

1. We will attempt to predict ***Sales*** based on ***Price*** by running a simple linear regression, i.e., we treat ***Price*** and ***Sales*** as an explanatory and dependent variable, respectively.

(a) [1point] What are the intercept and the slope of the least-squares regression line?

(Intercept) Price

13.64192 -0.05307

(b) Test to see whether ***Sales*** are influenced by ***Price***.

(b-1) [1 point] State null and alternative hypotheses.

H\_0: Price does not have any effect on sales.

H\_a: Price influences sales.

(b-2) [2 points] Report the test statistic and p-value and state your conclusion.

T: -9.912 P: <2e-16

Since the pvalue is far below 0.05, we reject our null hypothesis meaning that price does have a significant effect on sales.

(c) [1 point] Construct the 95% confidence interval for the slope.

105.9550 110.6423

(d) [1 point] Predict a value of Sales at ***Price***=150.

5.680962

(e) [1 point] Construct the 95% confidence interval for the mean of Sales at ***Price***=150.

5.243237 to 6.118688

2. We will now attempt to predict ***Sales*** based on ***Price, Advertising,*** and ***Population*** by carrying a multiple linear regression.

a) [1 point] Report the fitted regression equation.

Sales=13.164-0.055Price+0.127Advertising-0.00066Population

(Intercept) 13.1643149 0.6409944 20.537 < 2e-16 \*\*\*

Price -0.0547121 0.0050822 -10.765 < 2e-16 \*\*\*

Advertising 0.1270215 0.0187666 6.768 4.7e-11 \*\*\*

Population -0.0006623 0.0008461 -0.783 0.434

(b) [1 point] Which independent variables are associated with the dependent variable?

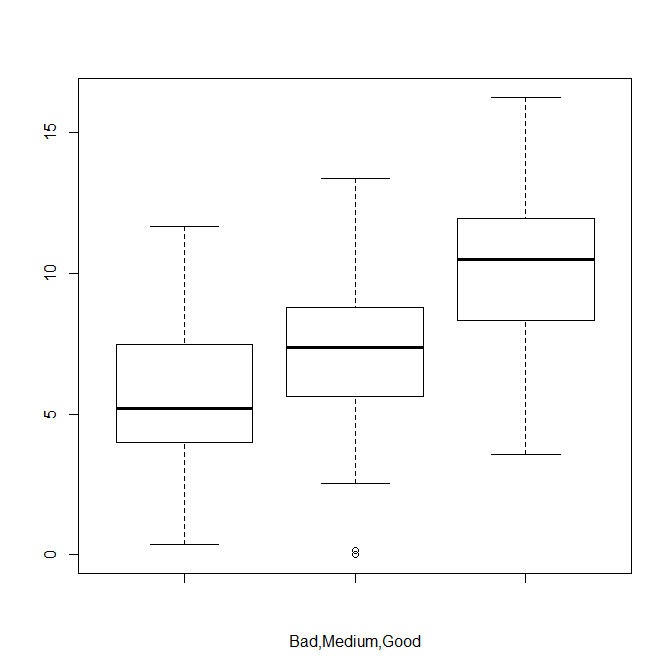
Price and Advertising are associated with Sales however Population is not associated with sales since its pvalue is greater than 0.05.

(c) [1 point] What is the predicted Sales at ***Price***=150, ***Advertising***=5, and ***Population***=270?

5.413801

3. Run ANOVA test to compare population means of Sales at different qualities of the shelving location for the car seats.

(a) [1 point] Make a side-by-side boxplot and describe the differences briefly.



You can easily see that the lower quality shelves cause the mean sales to be lower with the Good shelves having the highest range of sales.

(b) Test whether all means are equal.

(b-1) [1 point] State null and alternative hypotheses.

H\_0: All means are equal

H\_a: Means are different

(b-2) [2 points] Via ANOVA, report the test statistic and p-value. Do you reject H\_0?

Test Statistic: 92.23 Pvalue: 2.2e-16

Based on these two values, I am going to reject the null hypothesis and say that the means are different.

(c) [2 points] Use a multiple comparisons method to compare the three groups.

Bad Good

Good < 2e-16 -

Medium 3.6e-09 < 2e-16

Since all of these are below 0.05, none of the means are the same.

4. Test whether Income and US are associated or not.

(a) [1 point] Fill out the actual counts in the following table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | Income | | |
| Low | Middle | High |
| US | Yes | 60 | 131 | 67 |
| No | 41 | 70 | 31 |

(b) [1 point] State null and alternative hypotheses.

H\_0: Income and US status are associated

H\_a: Income and US status are not associated

(c) [2 points] Report the chi-squared test statistic, p-value. Do you reject H\_0?

X-squared = 1.8246, df = 2, p-value = 0.4016

Since this pvalue is greater than 0.05, we fail to reject the null hypothesis.

**CODE:**

data=read.csv("final.csv")

attach(data)

names(data)

fix(data)

# 1

fit1a=lm(Sales~Price)

summary(fit1a)

fit1a

predict(fit1a, se.fit=TRUE, interval="confidence")

newdata = data.frame(Price=150)

newdata

predict(fit1a, newdata, se.fit=TRUE, interval="confidence")

t.test(Price,Sales)

# 2

fit2a=lm(Sales~Price+Advertising+Population)

summary(fit2a)

newdata2 = data.frame(Price=150,Advertising=5,Population=270)

newdata2

predict(fit2a, newdata2, se.fit=TRUE, interval="confidence")

# 3

fit3a=lm(Sales~ShelveLoc)

anova(fit3a)

x1=Sales[ShelveLoc=="Bad"]

x2=Sales[ShelveLoc=="Medium"]

x3=Sales[ShelveLoc=="Good"]

com3=cbind(x1,x2,x3)

boxplot(x1,x2,x3,beside=TRUE,xlab="Bad,Medium,Good")

anova(fit3a)

pairwise.t.test(Sales,ShelveLoc,p.adjust="bonferroni")

#4

table(US,Income)

chisq.test(table(US,Income))

detach(data)