Question 1

0 Points

Question#1 - Multiple Linear Regression

The human resources manager of BigCom, Inc., wants to predict the annual salaries of given employees using the potential explanatory variables in the file **salary_data.csv**.

salary: Current annual salary (in dollars)
experience: Number of years of relevant work experience prior to coming to BigCom
years _employed: Number of years employed at BigCom
education: Number years of education beyond high school
gender: 0=Female, 1=Male
department: 1=Sales, 2=Purchasing, 3=Advertising, 4=Engineering
supervised: Number of employees supervised by this employee

a) Develop a model that allows to predict **the annual salary** of a given BigCom employee using all predictors given in the data. Copy and paste your R code here. . Hint: Use **lm()** and **summary()**.

```
df_1 = read.csv("q1_salary_data.csv")
str(movies_data)

multiple_linear_regression_model = lm(salary ~ experience + years_employed + education
+ gender + department + supervised, data = df_1)
summary(multiple_linear_regression_model)
```

salary = 19589.47 + 621.06 * experience + -106.55 * years_employed + 1631.83 * education + -1654.07 * gender + 2134.29 * department + supervised * 134.01

b) Based on your model in part (a), predict the annual salary of a female employee who served in a similar department at another company for 10 years prior to coming to work at BigCom. This woman, a graduate of a four-year collegiate business program, has been supervising 12 subordinates in the purchasing department.
    a. $37,671.34

c) Are there any “useless” (not significant) independent variables in the regression model? If yes, list them, and for each variable explain why is it “useless” (i.e. write down the hypotheses being tested, and p-value for each of them). Use 10% significance level.
    a. Experience
        i. Hypothesis

1. H0: bi = 0
2. H1: bi != 0

ii. P-value

1. P = 1.46e-05

iii. Statistically Significant? Yes

b. Years Employed

i. Hypothesis

1. H0: bi = 0
2. H1: bi != 0

ii. P-value

1. P = 0.6199

iii. Statistically Significant? No

1. Given the p-value is larger than the 0.05 significance level, we fail to reject the null hypothesis, which means the variable is equal to 0. Thus, this predictor is considered insignificant, or useless, because it has no impact in

c. Education

i. Hypothesis

1. H0: bi = 0
2. H1: bi != 0

ii. P-value

1. P = 6.01e-05

d. Gender

i. Hypothesis

1. H0: bi = 0
2. H1: bi != 0

ii. P-value

1. P = 0.295

e. Department

i. Hypothesis

1. H0: bi = 0
2. H1: bi != 0

ii. P-value

1. P = 0.0015

f. Supervised

i. Hypothesis

1. H0: bi = 0
2. H1: bi != 0

ii. P-value

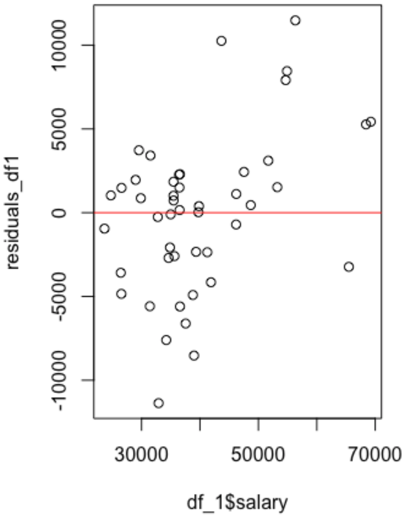
1. P = 0.1365

d) Based on your model in part (b), report the adjusted R-square value, and provide its interpretation. Characterize the model accuracy. Is this an excellent, good, or fair model?

a. The adjusted R squared is 0.7915,

e) Calculate the residuals (errors) for the model in part (b). Create a residuals plot vs salary. Copy-paste a residual plot, and report if residual plot indicates any problems with the regression model. Copy and paste your R code here. Use: the following R functions: **resid()** , **plot()**, and **abline(h=0, col="red")**

a.



b. residuals_df1 = resid(multiple_linear_regression_model)
plot(x= df_1$salary, y=residuals_df1)
abline(h=0, col="red")

f) Test residuals for normality as well. Comment on whether or not the residuals are normally distributed. Copy and paste your R code here. Use: **shaipro.test()**

a. shapiro.test(residuals_df1)

b. The p-value for the Shapiro test is p = 0.5851, which means the data is normally distribuated

g) Check the model for multicollinearity. Create a correlations matrix between " years_employed "," experience "," education ", "gender","department","supervised". Copy and paste the correlations matrix. Comment on whether or not the model suffers from multicollinearity. Copy and paste your R code here. First, install and load the "**GGally**" package in R as shown below. Then use the function **ggpairs()** function.

a.

Question 2

0 Points

Question#2

Tom Frederick is the computer support manager for a large company whose employees have been complaining about "spam", which many of us know as unwanted e-mail to solicit our money or our attention. Tom asked a sample of 40 employees to keep track of the number of spam messages they received during a week. He then installed spam filter software into the e-mail system in attempt to block some of the spam by identifying key words that can appear in such messages. If the filtering software is effective, the number of spam messages that employees receive should decrease. During the week following the installation, he asked same employees to keep track of the number of spam messages they receive. The results are recorded in the file "**spam.csv**".

a) What data type is being used in the study?

1. Numerical

b) Test sample differences for normality. Comment on whether or not the sample differences are normally distributed. Copy and paste your R code here. To receive credit, you must write hypotheses being tested, p-value for these hypotheses, and your conclusion. Use: **shaipro.test().**

c) Which is the appropriate test for this problem? Please explain shortly why? (Hint: Please use **TestTypes.pdf**)

2. ComparingTwo Populations – Matched Pairs (Wilcoxon Signed Rank Sum Test)

The p-values for both samples are less than 0.05, so they are not normal. They are also matched for you are finding the difference between data from before and after

d) Formulate the appropriate hypotheses to be tested to see if the anti-spam software installation has resulted in fewer "spam" messages received by employees. Your hypotheses should be problem-specific and NOT general. Do NOT test the hypothesis formulated.

3. We want to see if rated R versions of movies are morepopular than PG-13 versions. So
    1. H1: the locations of population (R movies) is to the right (>) from thelocations of population (PG-13)
        1. H1: the locations of population (PG-13 movies) is to the left (<) from thelocations of population (R movies)
        2. H0: the locations of population "PG13" and population "R" are the same

2.

Question 3

0 Points

Question#3

The Internet search engines are used by people on the daily basis. In 2008 Google processed 42%, Yahoo- 21%, Bing - 19%, of all searches. The rest of the Internet searches market (18%) was shared among smaller companies, many of which are now defunct (Ask.com, Quora.com, etc.). Since then, the online search engines were expanding aggressively. In order to see if the distribution of market among search engines changed, the survey was conducted. One thousand people were asked what is their search engine of choice, and responses were recorded as “Bing”, “Google” “Yahoo” and “other”. The data are recorded in the file “**internet.csv**”.

a) What data type is being used in the study?

1. Categorical - Nominal

b) Which is the appropriate test for this problem? Please explain shortly why? (Hint: Please use **TestTypes.pdf**)

2. Morethan Two Outcomes with Proportions (Chi-square Goodness-of-fit Test)
    1. This takes in multiple forms of nominal data

c) Formulate null and alternative hypotheses to test if the search engines market was redistributed compared to 2008. Your hypotheses should include numerical values to be tested.

H0: P_Google = 0.42,P_Yahoo = 0.21, P_Bing = 0.19, P_Other = 0.18

H1: That distribution of the searches by engine have changed since 2008

Question 4

0 Points

Question#4

Newborn babies normally lose between 5% to 7% of the weight in the first week of life. For proper development, it is important to ensure that babies quickly gain weight after the initial weight loss. Two leading manufacturers of baby-food are Infamil and Similac. Similac recently came up with the new baby formula, and claims that it allows babies to gain weight faster than the Infamil’s formula. To test this claim, a pediatrician at Riverside Health Care facility decides to track weight gain between week 1 and week 4 of life for a sample of babies fed with Similac and for a different sample of babies fed with Infamil formulas. The weight gains (in ounces) are recorded in the file “**baby_food.csv**”.

a) What data type is being used in the study?

3. Numerical

b) Check normality for each sample separately. Comment on whether or not the samples are normally distributed. Copy and paste your R code here. To receive credit, you must write hypotheses being tested, p-value for these hypotheses, and your conclusion. Use: **shaipro.test().**

c) Which is the appropriate test for this problem? Please explain shortly why? (Hint: Please use **TestTypes.pdf**)

4. Comparing Two Populations – Independent Samples (t-Test)

The data for the babies are not paired, so they are independent, and they are both normally distributed

d) What are the point estimates of the average weight gains for babies fed with Infamil and for babies fed with Similac?

e) Formulate null and alternative hypotheses that the pediatrician needs to test in order to statistically confirm or disprove that the Similac's new formula leads to a bigger weight gain than Infamil.

H0: (Mean of Similac - Mean of Infamil) <= 0
HA: (Mean of Similac - Mean of Infamil) > 0

q4_baby_food.csv

3.

Question 5

0 Points

Question#5

A taxi company manager is trying to decide whether the use of radial tires instead of regular belted tires <u>improves</u> fuel economy. Twelve cars were equipped with radial tires and driven over a prescribed test course. <u>Without changing drivers, the same cars</u> were then equipped with regular belted tires and driven once again over the test course. The gasoline consumption, in km per liter, was recorded in the file **tires.csv**

a) What data type is being used in the study?

1. Numerical

b) Test sample differences for normality. Comment on whether or not the sample differences are normally distributed. Copy and paste your R code here. To receive credit, you must write hypotheses being tested, p-value for these hypotheses, and your conclusion. Use: **shaipro.test().**

c) Which is the appropriate test for this problem? Please explain shortly why? (Hint: Please use **TestTypes.pdf**)

2. ComparingTwo Populations – Matched Pairs (t-Test)

d) What are the point estimates of the average gasoline consumption for radial tires driver and regular belted tires?

> mean_radial = 5.75
> mean_belted = 5.608333

e) Formulate the appropriate hypotheses to test in order to determine if the two patterns differ in level of interest towards them from potential consumers.

4.

Question 6

0 Points

Question#6

Can you become addicted to exercise? The chain of franchise workout facilities performed a study to answer that question. A random sample of dedicated exercisers who usually work out every day was drawn. Each completed a questionnaire that gauged their mood on a 5-point scale, where
5=very relaxed and happy,
4=somewhat relaxed and happy,
3=neutral feeling,
2=tense and anxious, and,
1=very tense and anxious.
The selected group was then instructed to abstain from all workouts for the next three days. Moreover, they were told to be as physically inactive as possible. After three days the same group completed similar questionnaire that gauged their mood on a 5-point scale. From the data collected, can we conclude that after 3 days of inactivity the mood of the group of dedicated exercisers started to deteriorate? The data are in the file "**exercise.csv**".

a) What data type is being used in the study?

1. Categorical- Ordinal

b) Which is the appropriate test for this problem? Please explain shortly why? (Hint: Please use **TestTypes.pdf**)

2. ComparingTwo Populations – Matched Pairs (Sign Test)

c) Formulate null and alternative hypotheses that need to be tested to see after 3 days of inactivity the mood of the group of dedicated exercisers started to deteriorate. Your hypotheses should include numerical values to be tested. Do NOT test the hypotheses formulated.

H0: the moods before and after are the same
H1: the moods before the experience are different(NOT=) from the moods after the experience

q6_exercise.csv

1.

Question 7

0 Points

Question#7

A construction material company that manufactures concrete wants to investigate how do the properties of the concrete change over time. One of the concrete properties is the ability to withstand the compression. The company poured 120 concrete samples. Two days after pouring the company measured the compressive strength (in thousands of pounds per square inch) of 40 out of 120 randomly selected samples. Seven days after pouring the compressive strength was measured for another randomly selected 40 samples, and the strength of last 40 samples was measured 28 days after they were poured. Is there a difference in the compressive strength of the concrete after 2, 7, and 28 days? Use the file "**concrete_strength.csv**".

a) What data type is being used in the study?

1. Numerical

b) Check normality for each sample separately. Comment on whether or not the samples are normally distributed. Copy and paste your R code here. To receive credit, you must write hypotheses being tested, p-value for these hypotheses, and your conclusion. Use: **shaipro.test().**

2. H0: The locations of all k populations are the same.
3. H1: At least two population locations differ.

c) Which is the appropriate test for this problem? Please explain shortly why? (Hint: Please use **TestTypes.pdf**)

4. ComparingThree or More Populations (Kruskal-Wallis)

q7_concrete_strength.csv

2.

Question 8

0 Points

Question#8

Lack of Sleep is a serious medical problem. It has been linked (among other things) to heart attacks and car accidents. A Statistics Canada study asked a random sample of Canadian adults to report the amount of sleep (hours) they normally get. Can we conclude from these data that the average amount of sleep is different for men and women? The data are recorded in the file “**sleep.csv**”.

a) What data type is being used in the study?

1. Numerical

b) Check normality for each sample separately. Comment on whether or not the samples are normally distributed. Copy and paste your R code here. To receive credit, you must write hypotheses being tested, p-value for these hypotheses, and your conclusion. Use: **shaipro.test().**

H$_1$: The location of “new drug” is to the right of the location of“aspirin”

H$_0$: The two population locations are the same

c) (Which is the appropriate test for this problem? (Hint: Please use **TestTypes.pdf**)

2. ComparingTwo Populations – Independent Samples (Wilcoxon Rank Sum Test)

d) Formulate the appropriate hypotheses to be tested to determine whether the average amount of sleep is different for men and women. Do NOT test the hypotheses formulated.