Assignment 4

N.B:- I have done the assignment’s question answer in this word file and the codes in assignment 4a Linear Data script of R, but you will find some of the codes of non-linear data too on that R script as well as doc file because I have answered the question in one doc file. I have also shown the analysis of non-linear data in a separate script file also please check that also. You can find it in Assignment Non Linear Data R script. I didn’t gave a doc file for non linear data analysis because the answer to the question is already in this file.

1. Show the difference between lung capacities of non-smokers vs. smokers using a plot?

Now Lung Capacities are numeric variables and Smoke variable is categorical where smoker and non smoker are category, so Stratified Boxplots or Separate Boxplots for Non-Smokers and Smokers based on LungCap are required to visualize the relationship between a categorical variable(Smoke) and numerical variable(LungCap).

Means1 <- aggregate(LungCap~Smoke,Smoke,mean) gets the means of each categorical variable.

boxplot(LungCap~Smoke,col=c(3,2),xlab="Smoke",ylab="LungCapacity",main="Boxplots of LungCapacity vs Smoke",las=1,names=c("NonSmoker","Smoker"),cex.lab=1.5,cex.axis=1.5,cex.main=2.5)

points(1:2, Means1$LungCap, col = "White",pch=".",cex=3) place the mean point in the boxplot.

Arguments

Col colours the individual boxplot green for non smoker and red for smoker.

xlab labels the X axis.

ylab labels the y axis.

main labels the main heading.

names labels the non-smoker and smoker groups.

cex.lab changes the label size.

cex.axis changes the axis size.

cex.main changes the main heading size.

Pch =”.” set the point as mean of each groups.

Analysis

For Non smoker the value of the median lies in 8 and for smoker the value is a little bit above 8.

The data for Non smoker is very spread out while for the smoker it consistent.

The

1. Perform a two sided t-test and explain the test in your own language?
2. One-Sample t-test are parametric methods (Not Distribution Free) to test single variables. What do I mean when I say parametric methods.

Parametric methods means that the sample are coming from a particular distribution.

t.test(LungCap, mu = 8, alt="two.sided", conf=0.95)

I have performed a two sided t-test on mean(mu) with confidence level 95% or level of significance = 0.05.

Before performing the test I have called the summary on LungData and the mean of LungCap was 7.863.

So my hypothesis was both sided alternative

H0 : µ = 8 vs H1 : µ≠8 with 95% confidence level.

Arguments

Mu = 8 sets the testing parameter value.

Alt specifies “left sided”, “right sided”, “Twosided” test to perform. By default the value of alt is “Twosided” but I have mentioned it explicitely.

Conf is to set the confidence level.

Analysis

Since the p value is greater than 0.05 alpha we accept the null hypothesis and conclude that the mean of LungCap is 8 or not significant.

1. Similarly as above we did the two sample test for smokers and non smoker independent data.

T2 <- t.test(LungCap~Smoke, mu = 0, alt="two.sided", conf=0.95, var.eq=F, paired=F)

Arguments

Mu = 0 because we are checking equality of two means

Alt specifies “left sided”, “right sided”, “Twosided” test to perform. By default the value of alt is “Twosided” but I have mentioned it explicitely.

Conf is to set the confidence level.

Var.eq = F is to set variance not equal.

Paired = F means not paired test.

Analysis

Since the p value is less than 0.05 alpha we reject the null hypothesis and conclude that the mean of smoker and non smoker are significantly different or not equal

To verify just check the above boxplot you will see the means are different.

1. What is ANOVA? Conduct a one way ANOVA for two suitable variables and explain every aspect of this test?

For This we take the WeightLossDiet Data in the Dataset Folder because for testing of one numerical and other categorical we use ANOVA.

What is Anova?

Anova is a method to compare multiple mean from two or more population.This method is parametric method. It is also called multiple hypothesis testing problem i.e. if the number of hypothesis is to be test is greater than 2.

One-Way Anova Test

Since over here we are dealing with one categorical variable(Diet) and another numerical variable(WeightLoss) we might be interested in a boxplot and ONE WAY Anova Testing because we have one categorical variable to test not numerical.

H0 : µ1 = µ2 = µ3 = µ4 vs H1 : not H0

Analysis

We call aov(WeightLoss~Diet) and store it in Anova and then call the summary(Anova) to get the p value which is 0.00113 < 0.05 (alpha)

So we reject the null hypothesis(accept the alternative) and conclude that the means are not equal or significant.

To verify that we did the test correctly check the boxplot you will see the mean are significantly deviated from one another.

1. Conduct Chi-square and Fisher’s exact test on two variables. Explain the test.

Chi Square test is a parametric approach, is a test of independence between two categorical variables.

So basically we go back to the LungCapdata and lets test the independence between two categorical variables say Gender and Smoke

Analysis

We accept the null hypothesis and conclude there is no independence between the variables.

Fisher test is a similar but non parametric test for independence

So basically we go back to the LungCapdata and lets test the independence between two categorical variables say Gender and Smoke.

Analysis

We accept the null hypothesis and conclude there is no independence between the variables.

The code is given on the script.

1. How to determine correlation and covariance in R? Give two examples with help of above data?

Correlation is a measure to check linear association between two numerical variables.

We have Pearson Correlation, Spearman Rank Correlation and Kendall’s Ranks Correlation.

Pearson is a parametric test and it measure linear relationship.

Spearman is non parametric test and it measure monotonic (Rank) relationship.

Kendall is non parametric test and it measure Ordering Consistency or Concordance Discordance relationship.

To determine correlation in R we use the cor(object x, object y).

e.g. are give in R script with comments.

Covariance is the deviation of two variables from their respective means.

To determine correlation in R we use the cov(object x, object y).

e.g. are give in R script with comments.

1. Compare the linear model (single + multiple) for every set of data. How to check non-linear relationship?

Given on the R script

To check non-linear relationship

From the graph given by

Plot(NONMultiple) we can make inference about the dataset is that it is non linear data.

The graphs that are given is residual vs fitted, normal q-q, scale location, residual vs leverage and from these graphs we can see the point does not lies on the curve and hence say it is non linear.

1. How to change a numeric variable to categorical variable? How to include categorical variables or factors in a linear regression model?

To convert the numeric(i.e. Height) variable into categorical variable, we make use of the cut() command.

L <- LETTERS[1:5]

CatH <- cut(Height, breaks = 5, labels =L, right = F)

In the above code I have broken the numerical height into 5 categories with labels A to E as levels.

To include categorical variable in linear regression line just use the lm() command with argument LungCap~Age+Gender+Smoke+Height.

I have already done this on other questions.

1. How to include categorical or qualitative variables(also known as factors) in a regression model using dummy or indicator variables? Explain with an example (use dataset).

Using Dummy or indicator variable we include variables in linear regression

Categorical variables with k levels or categories requires k-1 dummy or indicator variables.

e.g. XSmoke=1 if Smoke

0 if Non-Smoke

These are also called indicator variable.

I am doing the same concept for height in the code script.

Check the regression line and the means of each category of height

The means of each group are

2.148611 3.654640 5.396983 7.170552 8.692063 10.802778

And the coefficient of regression line is

2.148611 + 1.506029β1 + 3.248372 β2 + 5.021941 β3 + 6.543452 β4 + 8.654167 β5

Now put

1. β = (1,0,0,0,0,0) to get mean 2.148611
2. β = (0,1,0,0,0,0) to get mean 2.148611+1.506029 = 3.654640
3. β = (0,0,1,0,0,0) to get mean 2.148611+3.248372 = 5.396983

and so on …..

the intercept remain constant because it does not have any variable associated with it.

1. What do you mean by residual sum of squares (RSS)? What is the significance of its (RSS) value with respect to selecting a variable in a linear regression model? Explain with an example.

RSS(RESIDUAL SUM OF SQUARE) measures the overall difference between actual data and the values predicted by the model. The differences are squared and then added. It represents unexplained variation - a smaller RSS means that the model fits the data well. Also called the **Sum of Squared** Errors of prediction (SSE).

The test is called partial F-test which helps us decide whether to keep the variable or not in the model which compares the SSE or RSS to see if there’s any significant change in SSE due to removal of term.

The hypothesis we are testing is:-

H0 : no significant difference in SSE vs H1 : significant difference in SSE

If we accept H1 we accept that model which contains all the variables.

The test Statistics is

F=(SSE(model2)-SSE(model1)/df)/MSE

Where MSE is mean square error term and df is degree of freedom ()

Let us explain what is the significance of the RSS measure with respect to selecting a variable in a linear regression model by an example. The code is given in the R script is done for both linear(Lung Cap Data) and nonlinear(LungCapData2):-

e.g.

Let say the model which contains all the variables called model1.

Let say the model which contains one or more variable removed model2.

Now suppose we are working with LungCapData dataset which has linear rerlationship

Model1 is created from LungCapData1 which has linear relationship And suppose we want to estimate Lung Capacity using Age, Gender, Smoke and Height

Then Model1 will contains

**Lungcap = β0 + β1Age+β2Gender+β3Smoke+β4Height**

and Model 2 contains

**Lungcap = β0 + β1Age+β2Gender+β3Smoke**

We do anova testing on both the model to find which model has less RSS in this case for LungCapData dataset the first model has less RSS.

It is shown separately in the R script for both linear and non linear respective dataset.