

Session 11 – Linear Models

Assignment - 1



Contents

[Introduction](#page3) [2](#page3)



[Objective](#page3) [2](#page3)



[Prerequisites](#page3) [2](#page3)



[Associated Data Files](#page3) [2](#page3)



[Problem Statement](#page3) [2](#page3)



[Expected Output](#page3) [2](#page3)



|  |  |
| --- | --- |
| Copyrights© 2017, AcadGild. All Rights Reserved | 1 |
|  |  |



**Introduction**



This assignment will help you to understand the key concepts learnt in this session.

**Objective**



This assignment will test your skills on the Basics of Regression Analysis and Modeling.

**Prerequisites**



Not Applicable

**Associated Data Files**



Not Applicable

**Problem Statement**



#1. Use the given link and locate the bank marketing dataset. Data Set Link

#Perform the below operations:

**# a. Create a visual for representing missing values in the dataset.**

# using dataset airquality to create a visual for representing missing values in the dataset

newair = airquality

dim(newair)

str(newair)

summary(newair)

#before imputing

hist(newair$Ozone ,xlab = "ozone", ylab = "frequency",main="histogram of ozone",col="red")

median(newair$Ozone)

median(newair$Ozone,na.rm = T)

#imputed my median

newair$Ozone[is.na(newair$Ozone)]<- median(newair$Ozone,na.rm = T)

#check summary after done with imputing

summary(newair)

newair$Ozone

#visualize after imputing the variable ozone with the median

#lets visualize through histogram

#after imputing

hist(newair$Ozone ,xlab = "ozone", ylab = "frequency",main="histogram of ozone",col="red")

#with mice package

#The mice package provides a nice function md.pattern() to get a better

#understanding of the pattern of missing data

#using dataset airquality

library(mice)

md.pattern(airquality)

#using the VIM package as follows

#we can visualize like this too

library(VIM)

aggr\_plot <- aggr(airquality, col=c('navyblue','red'), numbers=TRUE, sortVars=TRUE, labels=names(airquality), cex.axis=.7, gap=3, ylab=c("Histogram of missing data","Pattern"))

#The plot helps us understanding that almost 73% of the samples are not missing any information,

#22% are missing the Ozone value, and the remaining ones show other missing patterns.

#Through this approach the situation looks a bit clearer in my opinion.

**#b. Show a distribution of clients based on a Job.**

#since in dataset I'm unable to find variable clients therefore i am using

#another variable say age for

#showing you distribution of a age based on job

#Set a different color for each group

library(ggplot2)

ggplot(bank.additional, aes(x=job, y=age, fill=job)) +

geom\_boxplot(alpha=0.3) +

theme(legend.position="none")+

ggtitle("Distribution of age based on a Job")

**#c. Check whether is there any relation between Job and Marital Status?**

#we are using Chi-Square Test for checking relation as both job and marital status are categorical variables

#so first defining the null hypothesis

#Ho: There is no relation between job and marital status

#Ha: There is relation between job and marital status

chisq.test(bank.additional$job ,bank.additional$marital)

#now as we can see p value is nearly 0 or less which is henceforth less than 0.05

#p value<0.05 hence we will reject the null hypo and accept the alternative hypothesis

#which says that There is relation between job and marital status

#also another way to check relation and more example

#by

#Pearson correlation test

#i'm using variable job and education in this

#Correlation test between job and education variables:

newbank = bank.additional

newbank$job <-as.numeric(newbank$job)

newbank$education <-as.numeric(newbank$education)

result <- cor.test(newbank$job, newbank$education, method = "pearson")

result

#In the result above :

#t is the t-test statistic value (t = 8.7235),

#df is the degrees of freedom (df= 4117),

#p-value is the significance level of the t-test (p-value nearly 0).

#conf.int is the confidence interval of the correlation coefficient at 95% (conf.int = [0.1046068, 0.1645802]);

#sample estimates is the correlation coefficient (Cor.coeff = 0.1347169).

#The p-value of the test is nearly 0, which is less than the significance level alpha = 0.05.

#We can conclude that job and education are significantly correlated with a correlation coefficient of 0.1347169 and p-value of 2.210^{-16}

**# d. Check whether is there any association between Job and Education?**

#we are using Chi-Square Test for checking association as both job and education are categorical variables

#hence Chi-Square Test for checking association

#so first defining the null hypothesis

#Ho: There is no association between job and education

#Ha: There is association between job and education

chisq.test(bank.additional$job ,bank.additional$education)

#now as we can see p value is nearly 0 or less which is henceforth less than 0.05

#p value<0.05 hence we will reject the null hypo and accept the alternative hypothesis

#which says that There is association between job and education

**Expected Output**

Not Applicable

|  |  |
| --- | --- |
| Copyrights© 2017, AcadGild. All Rights Reserved | 2 |
|  |  |