wk7\_TechSupportDiscussion7\_Galton

SumbarajuAditya

1/31/2021

Use the Galton.txt data and answer the following two questions ( question#1 and question #2) You don’t need to provide a separate R file. Please copy paste the code and the output in the word file, add your interpretation and upload a single word file.

**Question1 : Use the data ‘Galton.txt’ and perform the followings:**

setwd("C:/BU/DSC520/assignment\_repo/dsc520/")  
getwd()

## [1] "C:/BU/DSC520/assignment\_repo/dsc520"

Galton\_df <- read.delim("data/Galton.txt")  
head (Galton\_df)

## Family Father Mother Gender Height Kids  
## 1 1 78.5 67.0 M 73.2 4  
## 2 1 78.5 67.0 F 69.2 4  
## 3 1 78.5 67.0 F 69.0 4  
## 4 1 78.5 67.0 F 69.0 4  
## 5 2 75.5 66.5 M 73.5 4  
## 6 2 75.5 66.5 M 72.5 4

1. Replace column names; y <- dataFather, X2<- Mother, X3 <- data$Gender ( convert to numeric 0 and 1)

Galton\_df1 <- Galton\_df  
y <- Galton\_df1$Height  
X1 <- Galton\_df1$Father  
X2<- Galton\_df1$Mother   
Galton\_df1$Gender[Galton\_df1$Gender == "M"] <- 0  
Galton\_df1$Gender[Galton\_df1$Gender == "F"] <- 1  
X3 <- as.numeric(Galton\_df1$Gender)  
head(Galton\_df1)

## Family Father Mother Gender Height Kids  
## 1 1 78.5 67.0 0 73.2 4  
## 2 1 78.5 67.0 1 69.2 4  
## 3 1 78.5 67.0 1 69.0 4  
## 4 1 78.5 67.0 1 69.0 4  
## 5 2 75.5 66.5 0 73.5 4  
## 6 2 75.5 66.5 0 72.5 4

class(X3)

## [1] "numeric"

1. Perform multiple regression using lm() function. Use y as dependent variable and all other three variable as independent variables. Hint: mod<- lm(y ~ x1+x2+x3) summary(mod)

mod<- lm(y ~ X1+X2+X3, data=Galton\_df1)  
summary(mod)

##   
## Call:  
## lm(formula = y ~ X1 + X2 + X3, data = Galton\_df1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.523 -1.440 0.117 1.473 9.114   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 20.57071 2.74067 7.506 1.48e-13 \*\*\*  
## X1 0.40598 0.02921 13.900 < 2e-16 \*\*\*  
## X2 0.32150 0.03128 10.277 < 2e-16 \*\*\*  
## X3 -5.22595 0.14401 -36.289 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.154 on 894 degrees of freedom  
## Multiple R-squared: 0.6397, Adjusted R-squared: 0.6385   
## F-statistic: 529 on 3 and 894 DF, p-value: < 2.2e-16

1. Obtain the confidence intervals ( 95%)for model parameters

confint(mod)

## 2.5 % 97.5 %  
## (Intercept) 15.1918205 25.9496022  
## X1 0.3486558 0.4633002  
## X2 0.2601008 0.3828894  
## X3 -5.5085843 -4.9433183

1. Check model goodness of fit.

From the output of the model summary information for the Residual standard error,Multiple R-squared, p values, and F-statistics; indicates that model is goodness-of-fit. Key Pointers: if we observe the values of Multiple R-squared and Adjusted R-squared the difference between the values is insignificant. P-Value: The p-values is 0.00000000000000022 for independent variables which is almost “0”, This proves a significant relationship to the dependent variable. In my assumption this model is goodness-of-fit.

\*\* Using General linear model ( glm() function) \*\* We can use glm() as well- this is convenient when we have categorical variables in our data et. Instead of creating dummy variables by ourselves, R can directly work with the categorical variables.

1. Create a model named mod1 <- glm(y ~ x1+x2+X3 ) Note: X3 is data $ Gender, and it should be converted to factor.

glm\_galton\_df <- Galton\_df  
y <- glm\_galton\_df$Height  
X1 <- glm\_galton\_df$Father  
X2<- glm\_galton\_df$Mother   
X3 <- as.factor(glm\_galton\_df$Gender)  
mod1 <- glm(y ~ X1+X2+X3, data=glm\_galton\_df)  
summary(mod1)

##   
## Call:  
## glm(formula = y ~ X1 + X2 + X3, data = glm\_galton\_df)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -9.523 -1.440 0.117 1.473 9.114   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 15.34476 2.74696 5.586 3.08e-08 \*\*\*  
## X1 0.40598 0.02921 13.900 < 2e-16 \*\*\*  
## X2 0.32150 0.03128 10.277 < 2e-16 \*\*\*  
## X3M 5.22595 0.14401 36.289 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 4.641121)  
##   
## Null deviance: 11515.1 on 897 degrees of freedom  
## Residual deviance: 4149.2 on 894 degrees of freedom  
## AIC: 3932.8  
##   
## Number of Fisher Scoring iterations: 2

head(glm\_galton\_df)

## Family Father Mother Gender Height Kids  
## 1 1 78.5 67.0 M 73.2 4  
## 2 1 78.5 67.0 F 69.2 4  
## 3 1 78.5 67.0 F 69.0 4  
## 4 1 78.5 67.0 F 69.0 4  
## 5 2 75.5 66.5 M 73.5 4  
## 6 2 75.5 66.5 M 72.5 4

1. R uses “F” as X3’s reference level because F comes before M in the alphabetical order.Now, change the order and observe the summary of the model1.

glm\_galton\_df$Gender[glm\_galton\_df$Gender == "M"] <- 1  
glm\_galton\_df$Gender[glm\_galton\_df$Gender == "F"] <- 0  
X3 <- glm\_galton\_df$Gender  
mod2 <- glm(y ~ X1+X2+X3, data=glm\_galton\_df)  
summary(mod2)

##   
## Call:  
## glm(formula = y ~ X1 + X2 + X3, data = glm\_galton\_df)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -9.523 -1.440 0.117 1.473 9.114   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 15.34476 2.74696 5.586 3.08e-08 \*\*\*  
## X1 0.40598 0.02921 13.900 < 2e-16 \*\*\*  
## X2 0.32150 0.03128 10.277 < 2e-16 \*\*\*  
## X31 5.22595 0.14401 36.289 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 4.641121)  
##   
## Null deviance: 11515.1 on 897 degrees of freedom  
## Residual deviance: 4149.2 on 894 degrees of freedom  
## AIC: 3932.8  
##   
## Number of Fisher Scoring iterations: 2

head(glm\_galton\_df)

## Family Father Mother Gender Height Kids  
## 1 1 78.5 67.0 1 73.2 4  
## 2 1 78.5 67.0 0 69.2 4  
## 3 1 78.5 67.0 0 69.0 4  
## 4 1 78.5 67.0 0 69.0 4  
## 5 2 75.5 66.5 1 73.5 4  
## 6 2 75.5 66.5 1 72.5 4

In contrast with model “mod” in Q2 vs “mod2” in Q6; I have interchanged the values between male and Female. Assigned 1 to male and ) to Female. After the Model execution the results depicts that almost all the values in comparision with previous model remains same except for the coefficient values for X3, which is now measuring males rather than females since we have now assigned the 1 to Male.

**Question 2: Multiple regression model can be developed to fit almost any data set if the level of measurement is adequate and enough data points are available. Once a model has been constructed, it is important to test the model to determine whether it fits the data well and whether the assumptions underlying regression analysis are met. Assessing the model can be done in several ways, including testing the overall significance of the model, studying the significance test of the regression coefficients, computing the residuals, examining the stansard error of the estimate, and observing the coefficient of determination. Discuss how can we test the overall significance of the model.**

The null hypothesis, is usually the hypothesis that sample observations result purely from chance.

Alternative hypothesis. The alternative hypothesis, is the hypothesis that sample observations are influenced by some non-random cause.

Below are the factors proves that our model is a goodness of fit and the alternative hypothesis will be accepted. 1.Multiple R-squared: 0.6397, Adjusted R-squared: 0.6385. Difference is insignificant. 2.The P Values is almost close to 0 3.The coefficient intervals all being close 4.T-Value is >0 and looks far from 0 and this is an indication lack of null hypothesis