Lab 4-Write Up and R Code

**Section A:**

**Store Data:**

patID<-c("0001","0002","0003","0004")

age<-c(25,53,37,59)

diabtype<-c("type1","type2","type1","type1")

status<-c("poor","improved","excellent")

patientdata<-data.frame(patID, age, diabtype, status) #assigns data to a data frame

diabtype<-factor(diabtype) #converts to a factor variable

status<-factor(status, ordered=TRUE, levels=c("poor","improved","excellent")) #divides the status variable categorically

patientdata<-data.frame(patID, age, diabtype, status) #updates data frame

summary(patientdata) #produces output below

Text

Description automatically generated

* Data is now structured as desired.

**Section B:**

**Read In and Store:**

library(xlsx)

data<-read.xlsx('Lab\_4\_Dataset.xlsx',1)

head(data) #gives data preview

y<-data$Median.Home.Price

x<-data$Interest.Rate

plot(x,y,xlab="Interest Rate",ylab="Median Home Price",main="Median Home Price($) vs. Interest Rate(%)")

Chart, scatter chart

Description automatically generated

lmod1<-lm(y~x) #produces the linear model

Chart, scatter chart

Description automatically generated

* Fitted least squares model layered onto scatterplot.

summary(lmod1) #produces the result below

Text, letter

Description automatically generated

* The p-values indicate that both the intercept “α hat” and the slope estimate “β hat” are statistically significant at a 95% confidence level. The null hypothesis α hat=0 and β hat=0 must be rejected.

**Creating a Revised Model:**

lmod2<-lm(y~x+I(x^2)) #our second model, here I(x^2) represents x2 only without the possibility for interaction between any other independent variable

summary(lmod2)

x11()

Text, letter

Description automatically generated

* The highlighted p-values indicate that this is a better model with a slightly less significant intercept value. We observe that the slope values become more significant with the introduction of the squared interest variable.

lines(x,fitted(mod2)) #plots the new model against the data

Chart, line chart

Description automatically generated

* The polynomial model is better suited for modelling the data(this is based on the plot observed above).

library(ggplot2)

ggplot(data, aes(x, y)) + geom\_point() +stat\_smooth(method='lm', formula = y ~ poly(x,2), size = 1) + ggtitle('Median Home Price v. Interest Rate')+ xlab('Interest Rate') + ylab('Median Home Price') #better illustration using ggplot2

Chart

Description automatically generated

summary(residuals(lmod1)) #gives of summary of the residual values



* The residuals for model 1 are relatively evenly dispersed, this could imply normality. High max and min values may mean outliers have considerable influence.

summary(residuals(lmod2))



* This distribution appears to have less spread than the previous one. A much smaller IQR “interquartile range” and R “range” can be observed.

par(mfrow=c(2,2)) #modifies R output window to display in a 2x2 matrix format

plot(lmod1)

Chart, schematic

Description automatically generated

* No real pattern in the Residuals vs Fitted values graph, variance is reasonably constant.
* The Normal Q-Q plot is promising however it does show that some residuals deviate from normality.
* Leverage points are present however they aren’t influencing the model drastically. Point’s are within reasonable distance.

par(mfrow=c(2,2))

plot(lmod2)

Chart

Description automatically generated

* No recognisable pattern in the Residuals vs Fitted values graph, even more random than before. This indicates normality in the original relationship.
* The Normal Q-Q plot shows most values touching the line, this indicates constant variance of error terms to a high degree.
* Leverage point influence is minimal. Outlier leverage influence is minimal.

hist(residuals(lmod1),freq=FALSE)

lines(density(residuals(lmod1)),na.rm=TRUE) #imposes the residual’s density course over a histogram of its discrete values

Chart, histogram

Description automatically generated

* The distribution illustrated here isn’t normal. Non-normally distributed residuals can imply that the original relationship isn’t linear in nature. This implies that a linear model would not be a good fit for the variables in question. A linear model isn’t ideal.

hist(residuals(lmod2),freq=FALSE)

lines(density(residuals(lmod2)),na.rm=TRUE)

Chart, histogram

Description automatically generated

* The distribution illustrated here is normal with µ approximately 0. This model is a good fit. In this case the model is polynomial.

**Confidence Intervals:**

confint(mod2) #constructs a 95% around model least square estimates

**Text

Description automatically generated**

**Prediction Intervals:**

predict(lmod2,data.frame(x=7.0),interval="prediction",level=.95) #predicts a value (with 95% confidence) for the mean house price at 7% interest.

predict(lmod2,data.frame(x=6.5),interval="prediction",level=.95)

predict(lmod2,data.frame(x=5.8),interval="prediction",level=.95)