Assignment2

Username: zhu51

Name: Ziling Huang

# Question1

a)

Text

Description automatically generated with medium confidence

Text

Description automatically generated

b)

Chart, scatter chart

Description automatically generated

Chart, histogram

Description automatically generated

Comment: from the histogram, we can see it looks approximate normal distributed but with a little bit skewed on the right as the left tails are much lower the right tails. And from the plot graph, there is no obvious relationship between deprivation and crime, because I did not observe any pattern from it.

c)

poverty.lm1<-lm(Crime ~ Dep, data = poverty)

d)

Text, letter

Description automatically generated

Crime = 0.4931 \* deprivation + 0.4062

As there is one unit increase in deprivation, there is 0.4931 increase in crime.

When the deprivations is 0, the crime is 0.4062

e)

Chart, scatter chart

Description automatically generated

Comment: From Residuals vs Fitted graph we can see the observations are approximately scattered around 0, and the red line is nearly fitted the dash line. so we can say there is not too much deviation from linearity. And for the constant variance, we can see all the observations are evenly distributed. So this graph is really solid.

Chart

Description automatically generated

Comment: We can observe although there are some deviation between residuals and the line at the tail, overall the observations fit the line, so we can say the data are generally normally distributed.

Chart

Description automatically generated with medium confidence

Comment:

From the Cook’s distance graph, there are three observations that are away from our dataset which means they influence our model. Three observations are 9, 19 and 32, but none of them has Cook’s distance value > 0.4. So, we can say they are not significantly changing our model.

f)

our equation:

# crime = 0.4931 \* Dep + 0.4062

# when a deprivation score of 10

# crime = 0.4931 \* 10 + 0.4062 = 5.3372

g)

Text

Description automatically generated

# Question2

a)

Text

Description automatically generated with medium confidence

Chart, scatter chart

Description automatically generated

Comment: from the graph we can observe that there is a negative relationship between time and pollution, which means as the time increases, the pollution is decreasing.

Chart, histogram

Description automatically generated

Comment: from the histogram we can see there is positive skewed as the distribution with tail on its right side, with the most frequency between pollution from 13-14.

b)

Text

Description automatically generated

# the relationship between pollution and time is

# pollution = -0.12267 \* time + 15.78676

# when the time is 0, there are 15.78676 pollution

# As one unit increase in time, there is a decrease in 0.12267 in pollution

c)

Chart, scatter chart

Description automatically generated

Comment: The graph looks solid. We can see the red line is almost fit the dash line. And all the observations are scattered around 0 constantly. So that we can say there are not too much deviation between linearity.

Chart

Description automatically generated

Comment: from this graph we can observe the observations are nearly fitted the line although there are some deviations at the tail. But overall, we can say our data meet normality.

A picture containing timeline

Description automatically generated

Comment: there are 3 significant observations that influence out model which are 9, 19 and 32. But as their Cook’s distance are below 0.4, we can say they don’t not significant change our model, there is no need to delete them.

d)

Text

Description automatically generated

e)

As the P-value is low, the model is good.

Although the slope for Time^2 is 0.006565 which is low, the P-value for quadratic term is 4.81 \* 10 ^(-6) which is really low. Our null hypothesis is there is not relationship between quadratic term and our model. As P-value < 0.05, we reject null hypothesis which means the quadratic term is important to our model. So, we cannot reduce the quadratic term to simpler. And the R-squared increases from 0.7501 to 0.9012 which means the new predictor(Time^2) enhance the model. Adjusted R-squared increases from 0.739 to 0.892 which shows the same thing: out model becomes better.

f)

Chart, scatter chart

Description automatically generated

Comment: from the graph we can observe the red line is still fit the dash line and all the observations are constantly scattered. We can say the model meets the linearity.

Chart

Description automatically generated

Comment: we can see although there are some deviations at the tail, overall most of the observations fit the dash line/ So we can say the model meets the normality.

A picture containing chart

Description automatically generated

Comment: from the graph we can still observe there are three significate influential observations which are away from our model. But their Cook’s distance is all below 0.4, same as the Cook’s distance of linear graph. So no need to remove those observations.

g)

Chart, line chart

Description automatically generated

Comment: from the graph here, we can see the quadratic model fit the observations better. As it is curved, it fits more observations that sit at the corners. From the summary we printed out before, we can observe that the R-squared increases from 0.7501 to 0.9012 which means the new predictor(Time^2) enhance the model. Adjusted R-squared increases from 0.739 to 0.892 which shows the same thing: out model explains the variance better. Also, the residual standard error decreases from 0.6605 on 24 degrees of freedom to 0.4243 on 23 degrees of freedom, which means the our model is better. From the P-value of variable Time^2, we can see it is significantly smaller than 0.05, which means this variable is related to our model. Overall, quadratic model is better.

Code :

# question1

# (a)

poverty <- read.csv("poverty.csv")

head(poverty)

tail(poverty)

# (b)

plot(poverty$Dep, poverty$Cime,

xlab = "deprivation score",

ylab = "crime",

main = "the relationship between deprivation and the crime")

hist(poverty$Crime, xlab = "Crime", ylab = "Percentage", main="Crime percentage")

# (c)

poverty.lm1<-lm(Crime ~ Dep, data = poverty)

# (d)

summary(poverty.lm1)

# (e)

plot(poverty.lm1, which = 1)

# Residuals vs Fitted detects the linearity

plot(poverty.lm1, which = 2)

# Normal Q-Q detects normality

plot(poverty.lm1, which = 4)

# Cook's distance to see if there is any outliers.

# (f)

# crime = 0.4931 \* Dep + 0.4062

# when a deprivation score of 10

# crime = 0.4931 \* 10 + 0.4062 = 5.3372

# (g)

pred<-data.frame(Dep=c(10,15,20))

predict(poverty.lm1, pred, interval = "prediction")

# question2

# (a)

airquality <- read.csv("airquality.csv")

head(airquality)

tail(airquality)

plot(airquality$Time, airquality$Pollution)

hist(airquality$Pollution)

# (b)

air.lm1 <- lm(Pollution~Time, data = airquality)

summary(air.lm1)

# (c)

plot(poverty.lm1, which = 1)

plot(poverty.lm1, which = 2)

plot(poverty.lm1, which = 4)

# (d)

air.lm2<-lm(Pollution~Time+I(Time^2), data = airquality)

summary(air.lm2)

# (e)

# Does the summary output suggest you can reduce the quadratic model to the simpler linear model?

# (f)

# residual plots

plot(poverty.lm1, which = 1)

plot(poverty.lm1, which = 2)

plot(poverty.lm1, which = 4)

# (g)

x<-airquality$Time

plot(Pollution~Time, data = airquality)

lines(airquality$Time[x], fitted(air.lm1)[x], lwd=2)

lines(airquality$Time[x], fitted(air.lm2)[x], lwd=2, col="blue")

# quadratic model performs better is because it matched the observations on the tails.