Veera mukesh aripaka - U97302307

2022-10-04

#veera mukesh aripaka  
  
#Preprocessing:  
  
library(readxl)  
masterdata <- read\_excel("C:/Users/mukes/Downloads/6304 Module 4 Assignment Data.xlsx")  
  
colnames(masterdata)=tolower(make.names(colnames((masterdata))))  
  
attach(masterdata)  
  
set.seed(97302307)  
  
#mydata  
m4 = data.frame(masterdata[sample(1:nrow(masterdata),250),])  
  
#Analysis:  
  
#⦁ Conduct a simple regression analysis using your sample data with the dependent variable being "price" and the independent variable being "odometer".  
  
#veera mukesh aripaka  
  
odo.out = lm(price~odometer,data=m4)  
  
summary(odo.out)

##   
## Call:  
## lm(formula = price ~ odometer, data = m4)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15537 -5202 -2071 4051 32807   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.158e+04 1.127e+03 19.145 <2e-16 \*\*\*  
## odometer -7.922e-02 8.768e-03 -9.035 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8078 on 248 degrees of freedom  
## Multiple R-squared: 0.2477, Adjusted R-squared: 0.2446   
## F-statistic: 81.64 on 1 and 248 DF, p-value: < 2.2e-16

#from the regression conducted we can say that the equation of line is   
#price = (2.158e+04)-(7.922e-02)\*odometer  
  
# and different parameters can be seen in summary   
  
  
#⦁ Give verbal interpretations of all beta coefficients in your regression model. Make certain the language you use is understandable to a reasonably competent lay person shopping for a car on Craig's List.   
  
#veera mukesh aripaka  
  
summary(odo.out)

##   
## Call:  
## lm(formula = price ~ odometer, data = m4)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15537 -5202 -2071 4051 32807   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.158e+04 1.127e+03 19.145 <2e-16 \*\*\*  
## odometer -7.922e-02 8.768e-03 -9.035 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8078 on 248 degrees of freedom  
## Multiple R-squared: 0.2477, Adjusted R-squared: 0.2446   
## F-statistic: 81.64 on 1 and 248 DF, p-value: < 2.2e-16

#with the results plotted above we can say that the equation is   
#price = (2.158e+04)-(7.922e-02)\*odometer  
  
confint(odo.out)

## 2.5 % 97.5 %  
## (Intercept) 1.935650e+04 2.379598e+04  
## odometer -9.648758e-02 -6.195067e-02

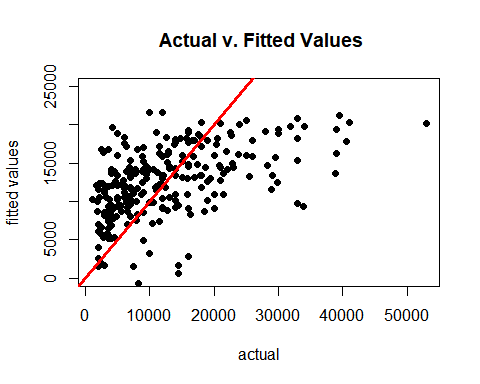
# plotted the confidence intervals of beta coefficients  
  
# here we have 2 beta coefficients one is the odometer which is the slope coefficient with an estimate of (-7.922e-02) and it lies in the confidence interval of (-9.648758e-02,-6.195067e-02) which means that the odometer might vary from -9.648758e-02 to -6.195067e-02 according to the regression model we achieved  
  
#similarly the other beta coefficient is the intercept with an estimate of 2.158e+04 and it lies in the confidence interval of (1.935650e+04,2.379598e+04) which means that the odometer might vary from 1.935650e+04 to 2.379598e+04 according to the regression model we achieved  
  
  
#⦁ Evaluate and interpret both the p value and the confidence interval on the "odometer" coefficient in your regression model.  
  
#veera mukesh aripaka  
  
summary(odo.out)

##   
## Call:  
## lm(formula = price ~ odometer, data = m4)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15537 -5202 -2071 4051 32807   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.158e+04 1.127e+03 19.145 <2e-16 \*\*\*  
## odometer -7.922e-02 8.768e-03 -9.035 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8078 on 248 degrees of freedom  
## Multiple R-squared: 0.2477, Adjusted R-squared: 0.2446   
## F-statistic: 81.64 on 1 and 248 DF, p-value: < 2.2e-16

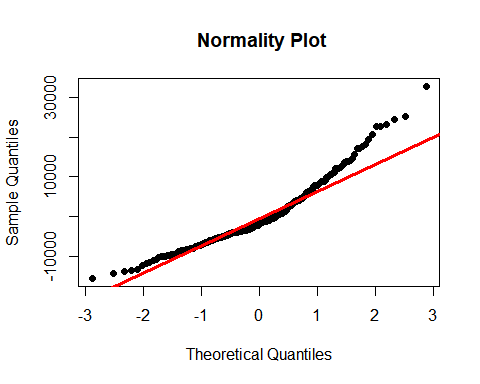
confint(odo.out)

## 2.5 % 97.5 %  
## (Intercept) 1.935650e+04 2.379598e+04  
## odometer -9.648758e-02 -6.195067e-02

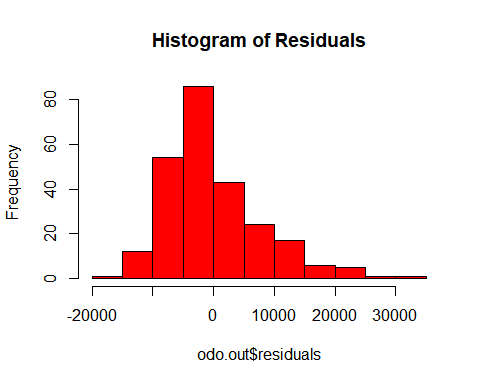
#from the summary plot we can say that the p value is less than 0.05 which means that we can reject the null hypothesis, which is slope beta coefficient is equal to zero. But we are not certain to accept the estimate since the p value is very low.  
#confidence interval shows that the slope of the regression model we plotted can range from -9.648758e-02 to -6.195067e-02 which is very low and near to zero from this we can say that the correlation between price and odometer is very less.  
  
# ⦁ Run appropriate diagnostics on your regression model to determine if it is in conformity with the LINE assumptions of regression.  
  
#Linearity  
  
plot(m4$price,odo.out$fitted.values,pch=19,ylim=c(0,25000),xlab=c("actual"),ylab=c("fitted values"),main=" Actual v. Fitted Values")  
  
abline(0,1,col="red",lwd=3)



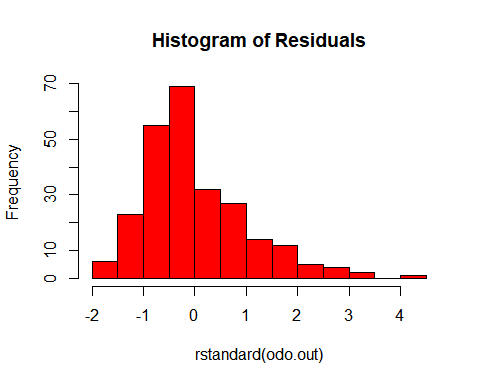
#looks like linear  
  
##Normality  
  
qqnorm(odo.out$residuals,pch=19,main=" Normality Plot")  
  
qqline(odo.out$residuals,col="red",lwd=3)



hist(odo.out$residuals,col="red",main="Histogram of Residuals")



hist(rstandard(odo.out),col="red",main="Histogram of Residuals")



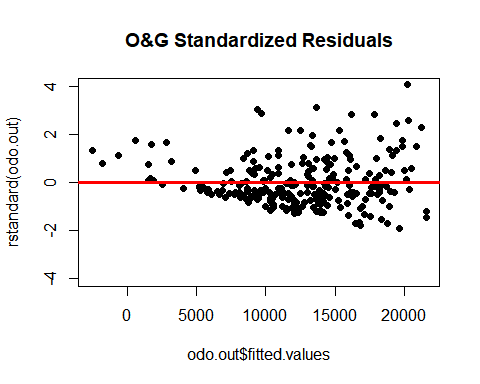
#from the QQPlot we can say that the majority of points does not lie on the qqline which shows that the data does not follow normal distribution. But the histograms plot show it as normal distribution.  
  
moments::skewness(odo.out$residuals)

## [1] 1.071147

moments::kurtosis(odo.out$residuals)

## [1] 4.348404

#Equality of Variances  
  
plot(odo.out$fitted.values,rstandard(odo.out),pch=19,ylim=c(-4,4),main="O&G Standardized Residuals")  
  
abline(0,0,col="red",lwd=3)



#equality of variances assumption does not satisfy because it have low standard deviations in the beginning and there is a spread as it goes on increasing. This is heteroscedasticity  
  
#independence:  
  
#here we assume there independence data since it does not show time variation.  
  
  
  
#⦁ Ms. Trayla Parks is considering offering her Toyota for sale on Craig's List. The Toyota currently has an odometer reading of 98,218 miles. Use your regression model to predict the price of the vehicle on Craig's List. Determine and verbally interpret the appropriate confidence interval on this prediction. If Trayla kept her vehicle one more year until the odometer showed 118,000 by how much would your model predict the price of her car would change?  
  
newdata = data.frame(odometer=98,218)  
  
predict(odo.out,newdata,interval = "predict")

## fit lwr upr  
## 1 21568.48 5504.373 37632.58

predict(odo.out,newdata,interval = "confidence")

## fit lwr upr  
## 1 21568.48 19350.24 23786.71

# so our model predicted that the price is 21568.48. But with the predictive intervals we can say that the it can range from value 5504.373 to 37632.58 and from confidence intervals we can say it ranges from 19350.24 to 23786.71, which is over long term . Lets compare it over a year.  
  
newdata1 = data.frame(odometer=118,000)  
  
predict(odo.out,newdata1,interval = "predict")

## fit lwr upr  
## 1 21566.89 5502.831 37630.95

predict(odo.out,newdata1,interval = "confidence")

## fit lwr upr  
## 1 21566.89 19348.97 23784.82

#we got new prediction as 21566.89 wich lies in the confidence interval as predicted above in CFI of 98,218  
#Also the new confidence interval got widened which means that the car price gets decreased over time.