An Introduction to Statistical Learning 3.7 Exercise 8

Soodabeh

September 7

This question involves the use of a simple linear regression on the Auto data set.

(a) Use the lm() function to perform a simple linear regression with mpg as the response and horsepower as the predictor.Use the summary() function to print the results.
Answer:

```
#install.packages("ISLR")
library(ISLR)
y_Auto=Auto$mpg
x_Auto=Auto$horsepower
Model_auto=lm(y_Auto~x_Auto)
summary(Model_auto)

##
## Call:
## lm(formula = y_Auto ~ x_Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.5710  -3.2592  -0.3435   2.7630  16.9240
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 39.935861   0.717499   55.66   <2e-16 ***
## x_Auto      -0.157845   0.006446  -24.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.906 on 390 degrees of freedom
## Multiple R-squared:  0.6059, Adjusted R-squared:  0.6049
## F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16

cat(" the mean value for the response =",mean(y_Auto))

##  the mean value for the response = 23.44592

#(4.906)/mean(y_Auto)
```

i. Is there a relationship between the predictor and the response?

Answer:

The p-value (< 2.2e-16) corresponding to the F-statistic is very low, so we reject the hypothesis. We conclude that there is a relationship between mpg and horsepower.

ii. How strong is the relationship between the predictor and the response?

Answer:

The quality of a linear regression fit is defined by two criteria: (a) the residual standard error (RSE) and (b) the regression R-squared. For Auto data, the RSE is 4.906 while the mean value for the response is 23.44592, indicating a percentage error of approximately 21%. The coefficient of determination is 0.6059, so the regression model explains approximately 61% of the variation in the values of mpg. These indicate that there is a moderate to strong relationship between mpg and horsepower.

iii. Is the relationship between the predictor and the response positive or negative?

Answer:

The relationship between the predictor and the response is negative as the slope of the linear regression is negative.

iv. What is the predicted mpg associated with a horsepower of 98? What are the associated 95% confidence and prediction intervals?

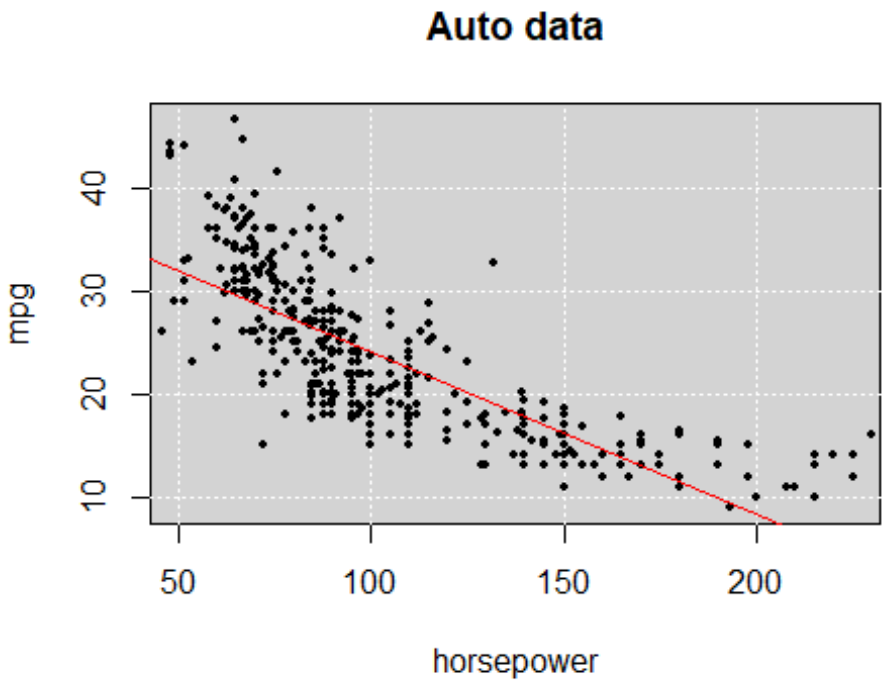
Answer: 24.46708

```
#iv.
#Confidence interval
predict(Model_auto,data.frame(x_Auto =98), interval = "confidence",conf.int =
0.95)

##        fit      lwr      upr
## 1 24.46708 23.97308 24.96108

#Prediction interval
predict(Model_auto,data.frame(x_Auto =98), interval = "prediction",conf.int =
0.95)

##        fit     lwr      upr
## 1 24.46708 14.8094 34.12476
```

(b) Plot the response and the predictor. Use the abline() function to display the least squares regression line.

Answer:

```
plot(x_Auto,y_Auto,ylab ="mpg" ,type = "n" , xlim=c(50,225), xlab
="horsepower",main="Auto data")
rect(par("usr")[1],par("usr")[3],par("usr")[2],par("usr")[4],col = "light
gray")
grid(col = "white", lty = "dotted",lwd = par("lwd"))
points(x_Auto,y_Auto,ylab ="mpg",pch = 20, cex = 0.8 , xlab ="horsepower")
abline(Model_auto, cex = 25, col = "red")
```

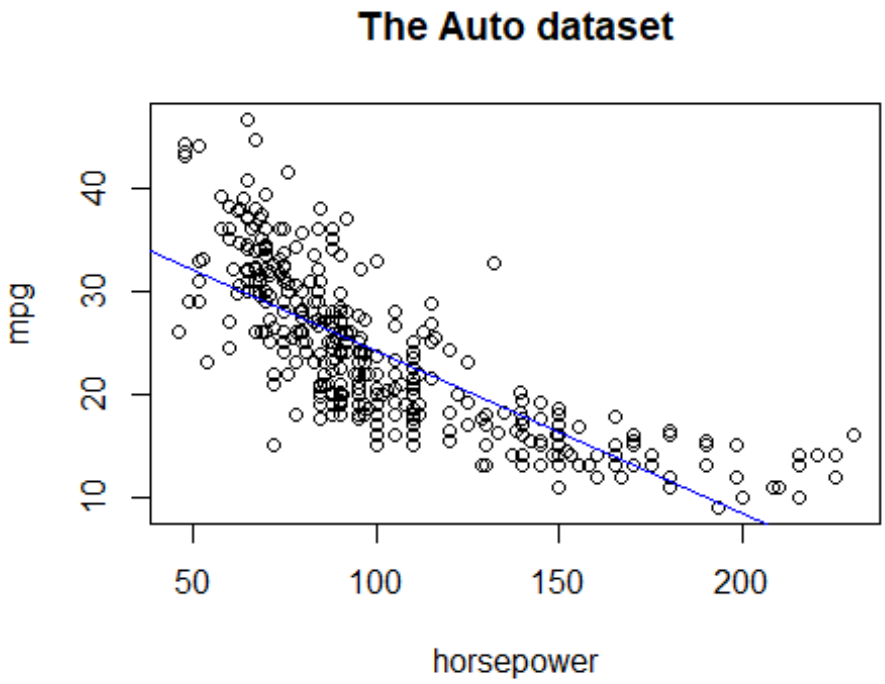


(c) Use the plot() function to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit.

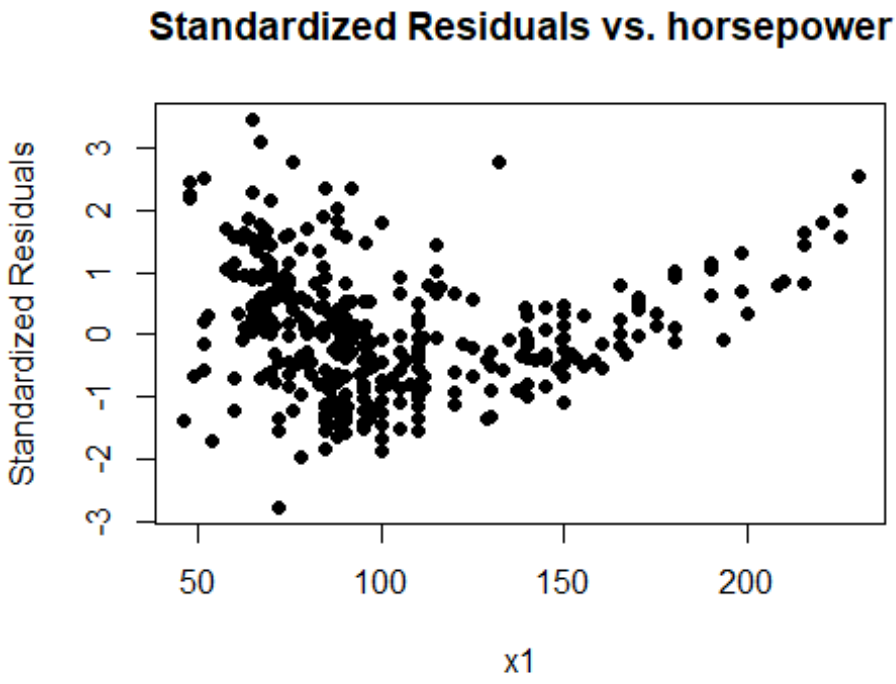
Answer:

In the scatter plot of standardized residual, a quadratic pattern is evident, so this indicates that the relationship between mpg and horsepower is in fact non-linear. We can also identify non-constant variances in the errors, from the response of a funnel shape in the residual plot. The normal Q-Q plot is close to a straight line, so the assumption that the errors are normally distributed is true. The large number of bad leverage points implies that an incorrect model has been fit to the data.

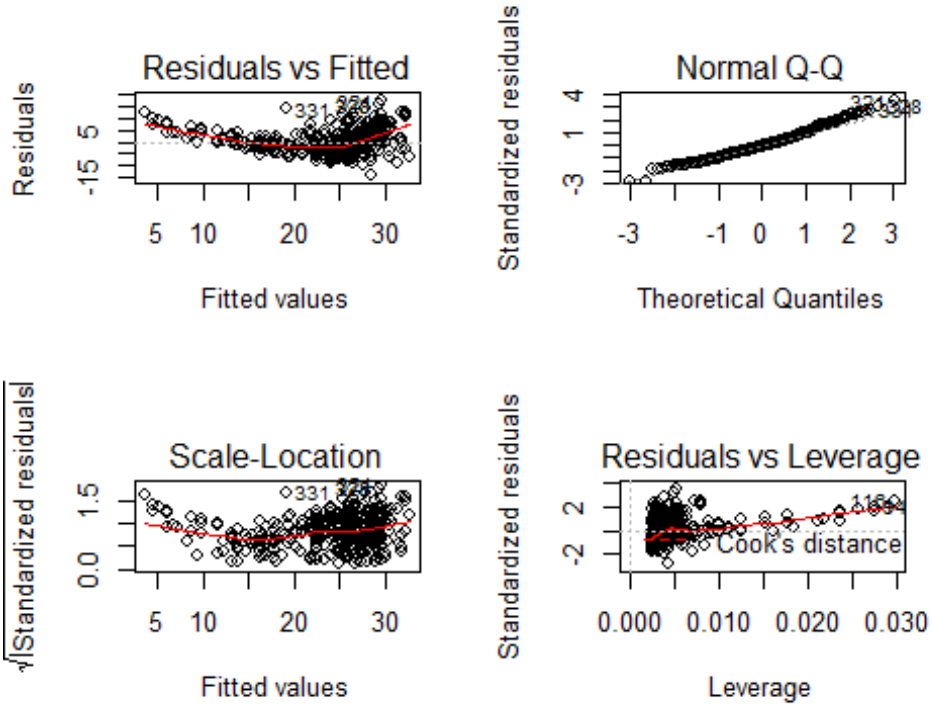
```
auto=data("Auto", package = "ISLR")
library(ISLR)
y1=Auto$mpg
x1=Auto$horsepower
Model_auto=lm(y1~x1)

plot(x1,y1,xlab = "horsepower", ylab = "mpg", main = "The Auto dataset")
abline(Model_auto, col="blue")
```



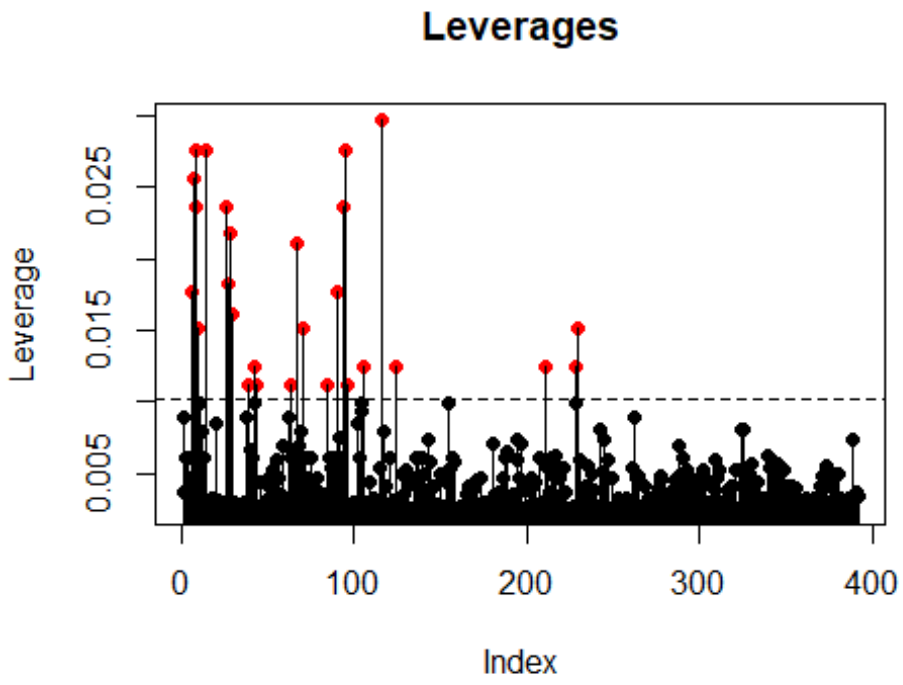

```r
#look at the standardized residuals
sdres1 <- rstandard(Model_auto)
plot(sdres1 ~ x1, ylab = "Standardized Residuals", main = "Standardized
Residuals vs. horsepower",pch = 16)
```



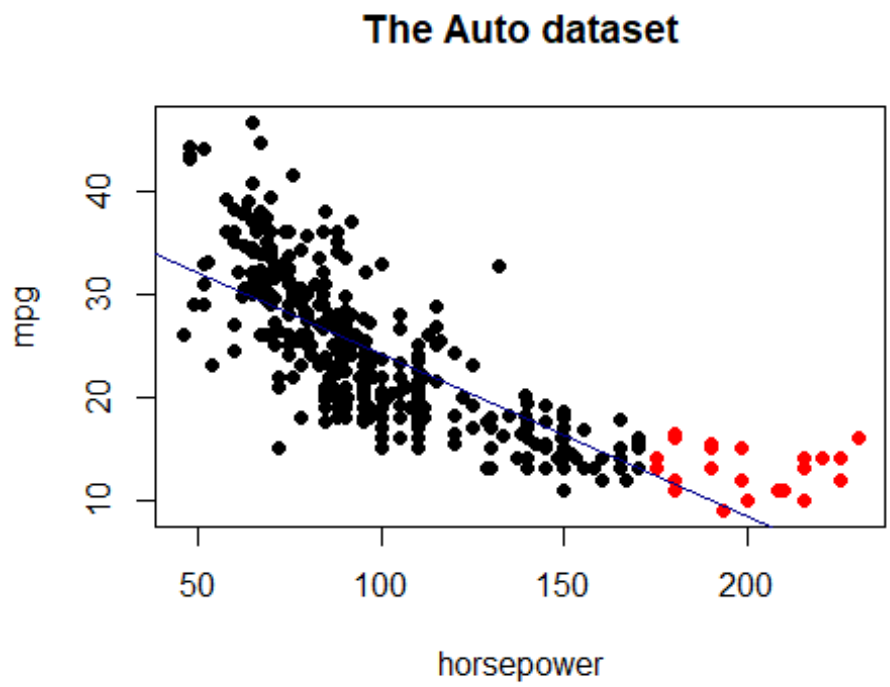

```r
#The 4 diagnosis plots in R
par(mfrow=c(2,2))
plot(Model_auto)
```



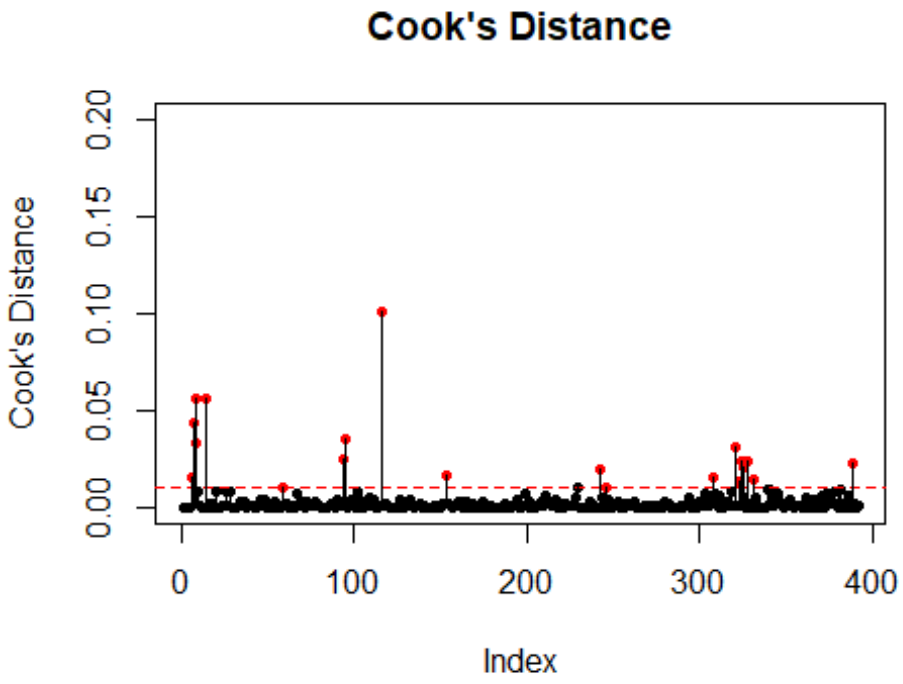
```r
#plot the leverage for the points
par(mfrow=c(1,1))
lev <- lm.influence(Model_auto)$hat
high <- 4/length(y1)
co= ifelse(lev <= high ,1,2)
plot(1:length(y1),lev, pch = 16, xlab = "Index", ylab = "Leverage", main =
"Leverages", col=co)
lines(1:length(y1), lev, type = "h")
abline(h = high, lty = 2) ## adding a dashed line for our rule of thumb
```



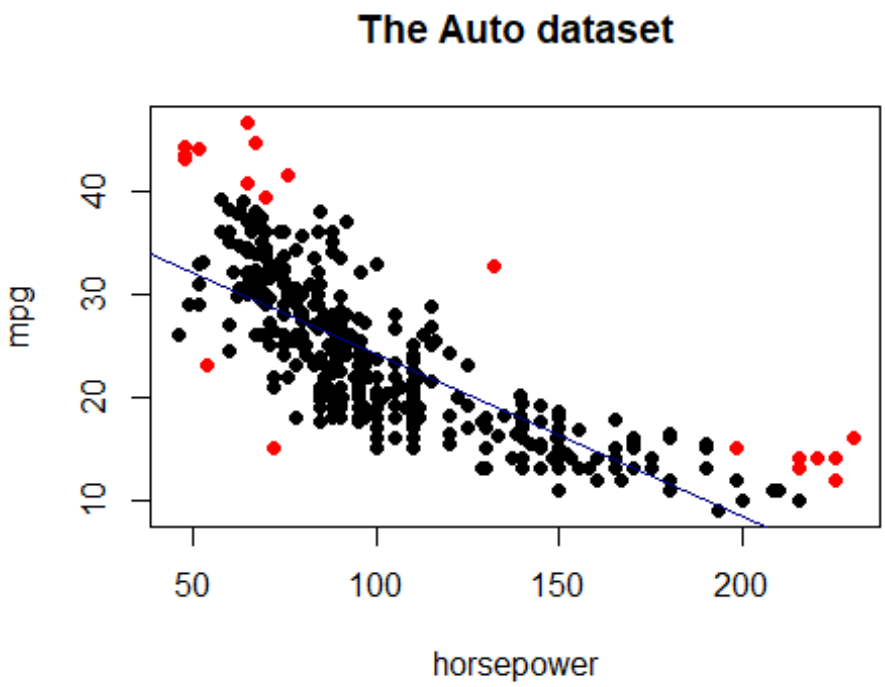

```r
#text(1:392, lev, labels=1:392, cex= 0.7)


## color the high leverage point red
plot(x1 ,y1,xlab = "horsepower", ylab = "mpg", main = "The Auto dataset", col
= co, pch = 16)
abline(Model_auto, col="blue4")
```




```r
#plot the Cook's distance for the points
sd_res <- rstandard(Model_auto)
cooks <- sd_res^2/2 * lev/(1-lev)
co_2= ifelse(cooks <= (4/(length(y1)-2)),1,2)
plot(x = 1:length(y1), y = cooks, pch = 20, main = "Cook's Distance", xlab =
"Index", ylab = "Cook's Distance", ylim = c(0,.2), col=co_2)
abline(h = 4/(length(y1)-2), lty = 2, col = 2)
lines(1:length(y1),cooks, type = "h")
```




```
## color the Cook's distance point red
plot(x1 ,y1, xlab = "horsepower", ylab = "mpg", main = "The Auto dataset",
col = co_2, pch = 16)
abline(Model_auto, col="blue4")
```


The Auto dataset
mpg
10
20
30
40
50
100
150
200
horsepower