Q14

Rahul Atre

2023-10-12

Q14 [Regression Modeling]

This question involves the *Auto* dataset.

1. Use the *lm*() function to perform a simple linear regression with *mpg* as the response Y and *horsepower* as the predictor X. Use the *summary*() function to print the results. Comment on the output. For example:

i. Is there a relationship between the predictor and the response?
ii. How strong is the relationship between the predictor and the response?
iii. Is the relationship between the predictor and the response positive or negative?
iv. What is the predicted mpg associated with a horsepower of 98? What are the associated 95% confidence and prediction intervals?

First, we must import and load the data from Auto.csv. Let's also omit any missing values from our data.

```
autoData = na.omit(read.csv("Auto.csv"))

horsepower = autoData$horsepower
mpg = autoData$mpg

lin_model = lm(mpg ~ horsepower)
summary(lin_model)
```

```
##
## Call:
## lm(formula = mpg ~ horsepower)
##
## 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 ***
## horsepower  -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
```

The simple linear model that we have obtained is $\hat{Y} = 39.9359 - 0.1578X_1$.

- (i) For β_1, we can see that the corresponding p-value is very small. This indicates that β_1 is statistically significant, and that there is a strong relationship between the predictor and response variable.
- (ii) Looking at the Adjusted R^2 value, we can see that it is $0.6049 = 60.49\%$. This means that roughly 60.49% of the variation in from the response is due to the predictor. So, we can say that the relationship between mpg and horsepower is quite strong, at about 60.49% variability.
- (iii) The relationship between horsepower and mpg is negative, since $\beta_1 = -0.1578$. This implies that if the car's horspower were to increase, the overall mpg would decrease.
- (iv) We can use the following function to check the prediction and confidence interval for a horsepower of 98:

```
predict(lin_model, data.frame(horsepower=98), interval="prediction")
```

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

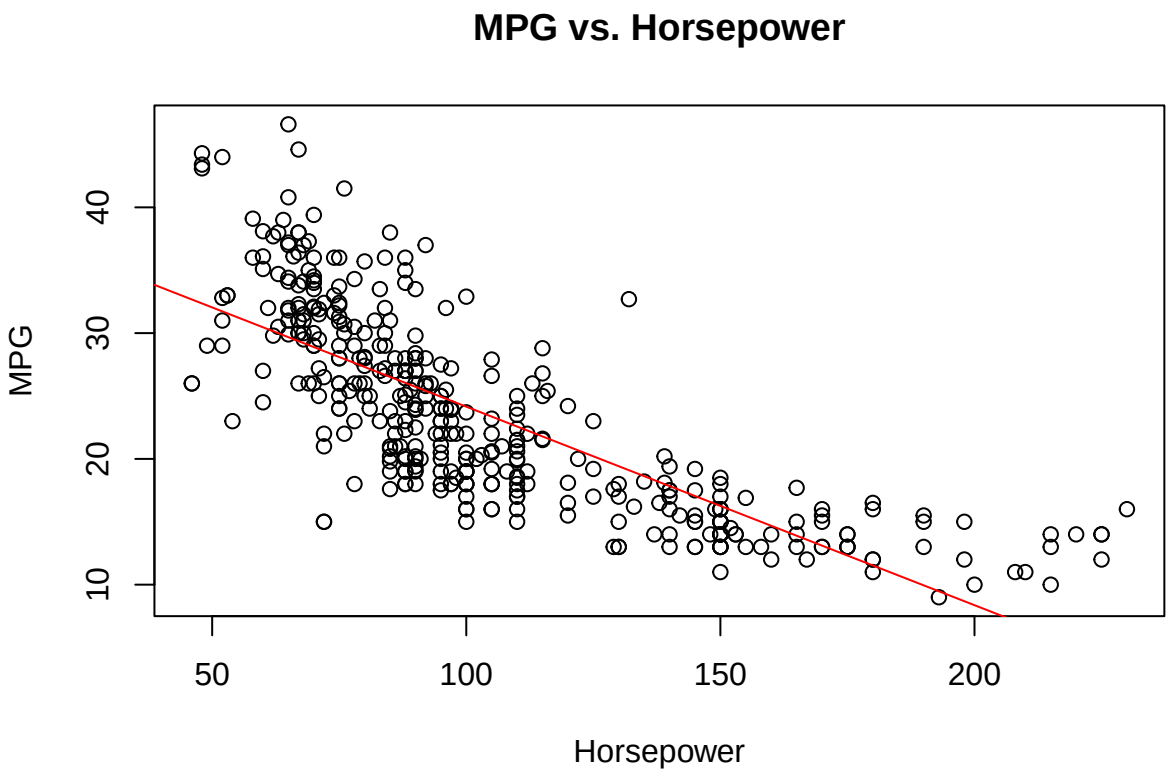
```
predict(lin_model, data.frame(horsepower=98), interval="confidence")
```

```
##        fit      lwr      upr
## 1 24.46708 23.97308 24.96108
```

Above are the 95% confidence and prediction intervals. Also, for a given 98 horsepower value, the prediction is 24.46708.

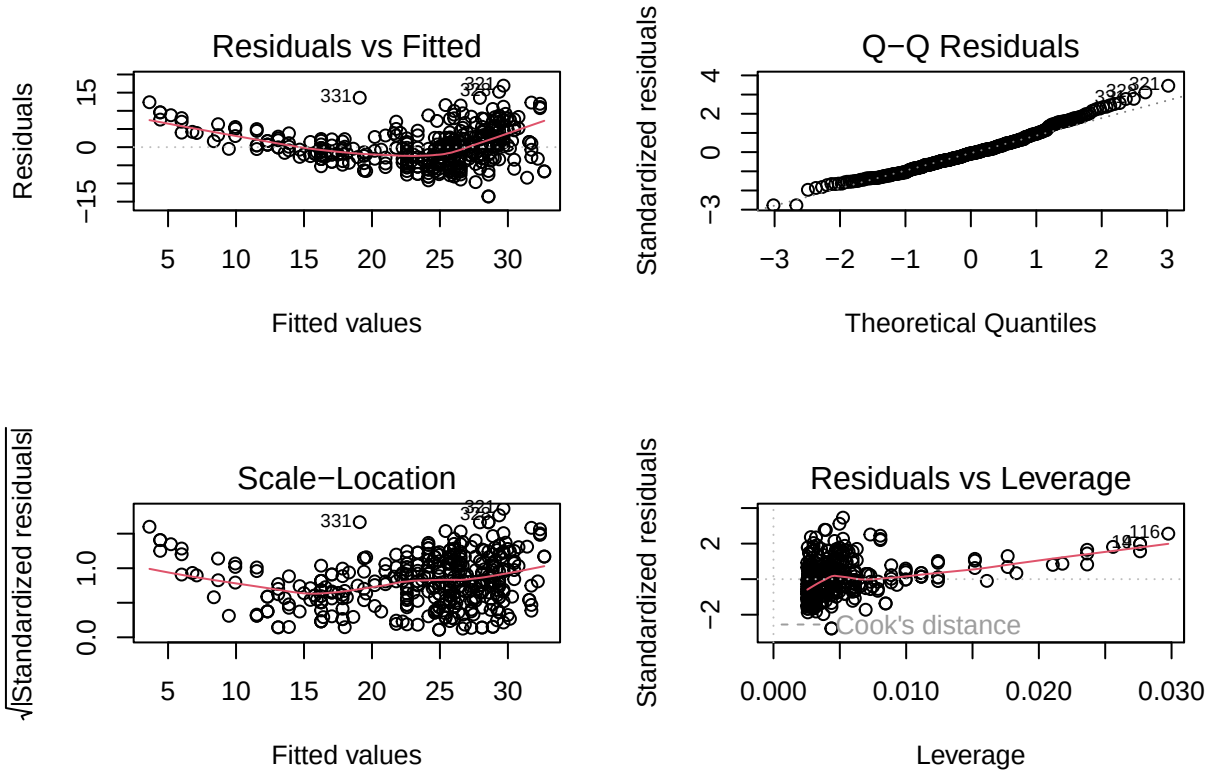
2. Plot the response and the predictor. Use the *abline*() function to display the least squares regression line.

```
plot(horsepower, mpg, main = "MPG vs. Horsepower", xlab = "Horsepower", ylab ="MPG")
abline(lin_model, col = 'red')
```



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

```
par(mfrow = c(2, 2))
plot(lin_model)
```



From the Residuals vs. Fitted plot, we can see that the residuals follow a parabola-like, which implies that the data is non-linear. Also, from the Q-Q plot, we can see that the residuals towards the upper end are not normally distributed, and do not follow the line of best fit. The Residuals vs. Leverage plot shows that there are some data points with high leverage, meaning that they can have a large impact on the results of the model. Using both plots together as analysis, this could strongly imply that there are a few outliers in our dataset.