DSI-06 Homework 1: Chapter 3, pg 123

Julia Gallucci

2023-02-20

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

```
install.packages("ISLR") #install package containing Auto dataset
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
library(ISLR)
attach(Auto) #attach Auto dataset to make the variables associated with Auto available.
head(Auto) #return the column names and first few rows of the dataset
```

```
##   mpg cylinders displacement horsepower weight acceleration year origin
## 1  18         8          307        130   3504         12.0   70      1
## 2  15         8          350        165   3693         11.5   70      1
## 3  18         8          318        150   3436         11.0   70      1
## 4  16         8          304        150   3433         12.0   70      1
## 5  17         8          302        140   3449         10.5   70      1
## 6  15         8          429        198   4341         10.0   70      1
##                        name
## 1 chevrolet chevelle malibu
## 2         buick skylark 320
## 3        plymouth satellite
## 4             amc rebel sst
## 5               ford torino
## 6          ford galaxie 500
```

(a) Use the lm() function to perform a simple linear regression with mpg as the response and horsepower as the predictor. Use thesummary() 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?

```
# Response variable Y : mpg,
# Predictor variable X: horsepower
```

```r
# The lm() function from the stats package performs the fitting of our linear model using the general s
# lm(y ~ x)
Auto_Model <- lm(mpg ~ horsepower)
# use summary() function to look at the results of our fit.
Auto_Model_summary <- summary(Auto_Model)
Auto_Model_summary
```

```
##
## 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
```

We can extract individual elements from the summary such as a information about the linear regression coefficients or R^2.

```r
Auto_Model_summary$coefficients
```

```
##               Estimate  Std. Error   t value      Pr(>|t|)
## (Intercept) 39.9358610 0.717498656  55.65984 1.220362e-187
## horsepower  -0.1578447 0.006445501 -24.48914  7.031989e-81
```

```r
Auto_Model_summary$r.squared
```

```
## [1] 0.6059483
```

i) So we have the coefficient estimates, their associated standard errors, and the t-statistic and p-value associated with the hypothesis test H0 : 1 = 0. The p-value for this test is significant so we can conclude there is a relationship between mpg and horsepower.
ii) According to the summary table, the R^2 value indicates that about 60.6% of the explained variance in mpg is due to horsepower.
iii) The t-statistic is negative, so we can conclude there is a negative relationship between mpg and horsepower.

the predict function predicts values based on a linear model

```r
# uses the syntax predict(object, newdata, interval)
predict(Auto_Model,data.frame(horsepower=c(98)),interval="prediction")
```

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

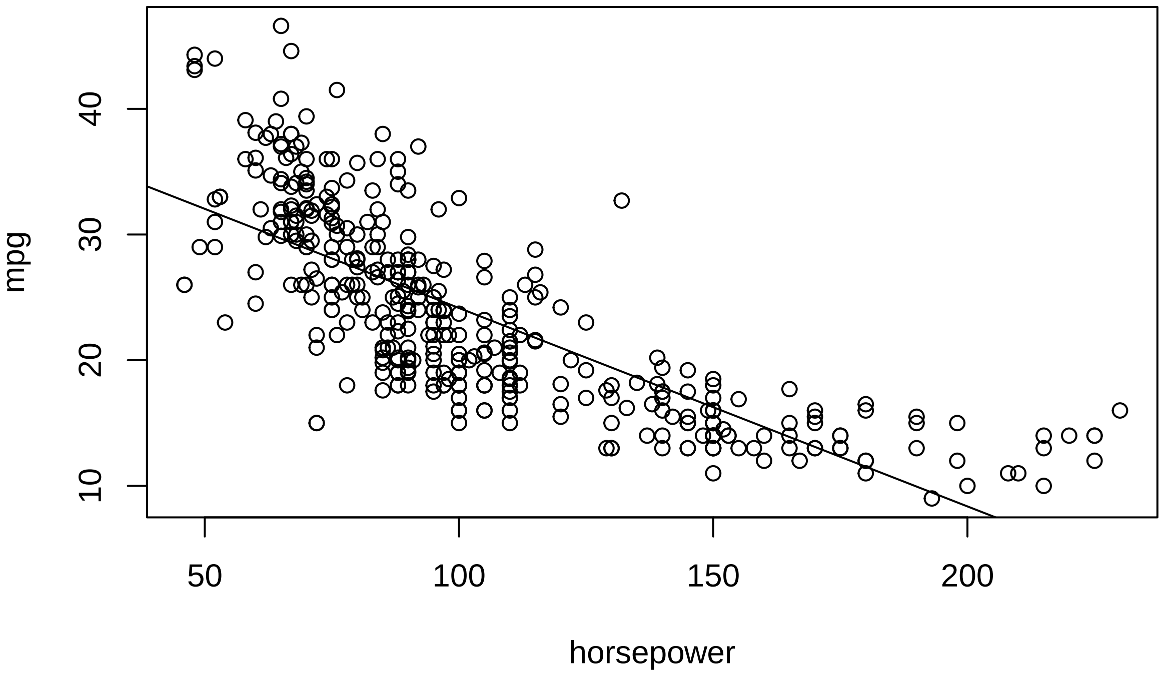
The confidence interval in the predict function will help us to gauge the uncertainty in the predictions.

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

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

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

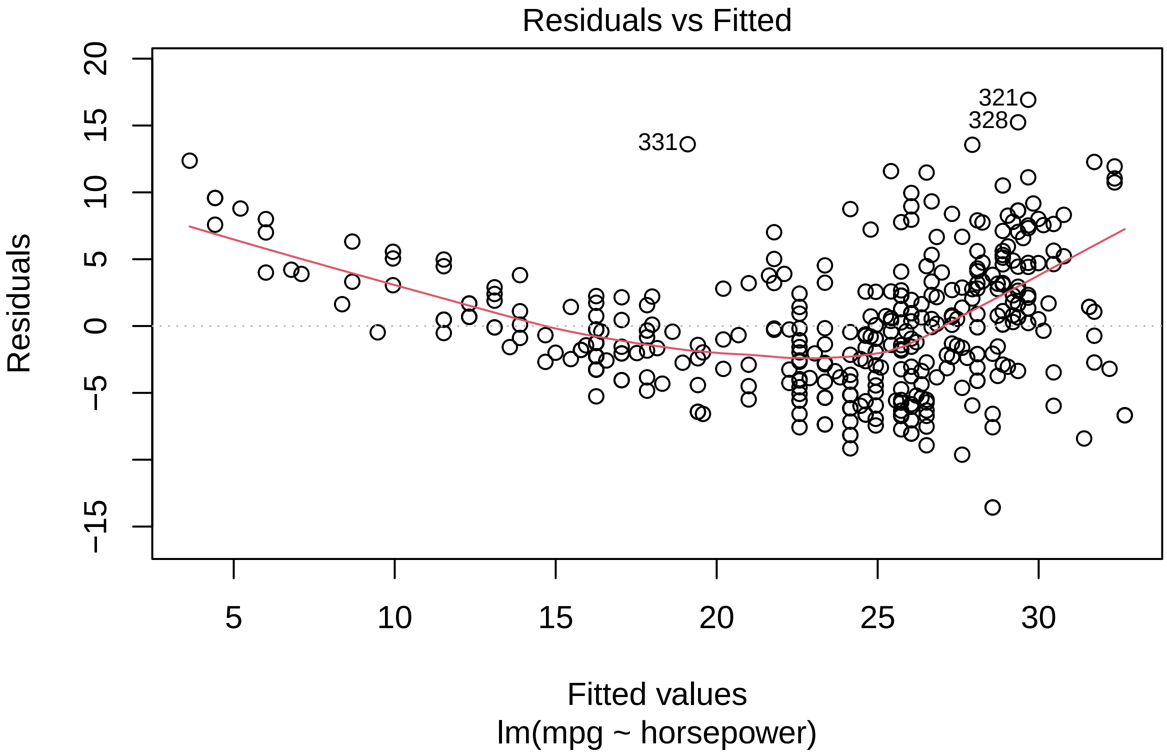
```
plot(mpg ~ horsepower) #we can plot our data
abline(Auto_Model) #and the linear regression model we fit.
```

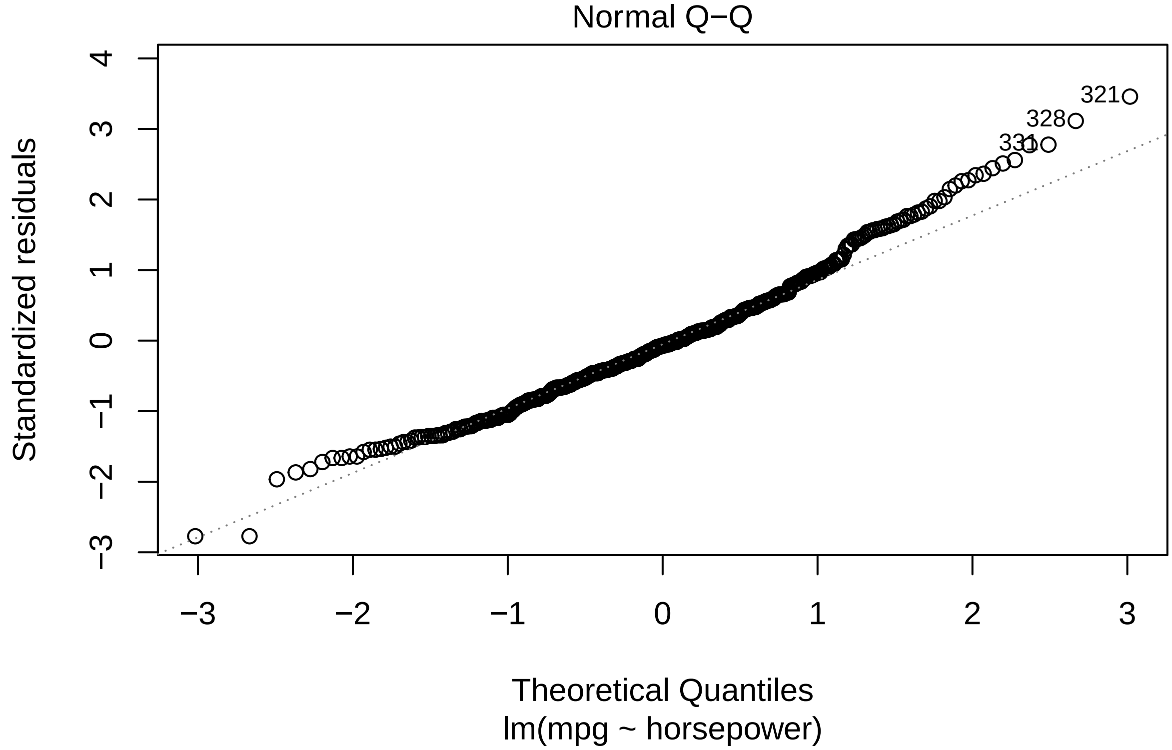


(c) Use the plot() function to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit.

There are a few plots that can help to identify problems with our data or with our fit. If we use the plot() function there are 4 plots that are automatically generated. We are particularly interested in the first 2 so we use the argument which = c(1, 2).

```
plot(Auto_Model, which = c(1, 2))
```





The first graph shows that there is a non-linear relationship between the response and the predictors We can also note the heteroskedasticity: as we move to the right on the x-axis, the spread of the residuals seems to be increasing. Finally, points 331, 321, and 328 may be outliers, with large residual values. The linear model assumes our residuals are normally distributed, we can use a Normal Q-Q plot to check that assumption. That appears to be a fairly safe assumption. The points seem to fall about a straight line.