Assignment14

John Bute

2024-10-14

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.

We perform simple linear regression, where mpg = B0 + B1 * horsepower

```
auto <- read.csv("C:/Users/johnb/Desktop/Machine Learning/data/Auto.csv",
stringsAsFactors = TRUE)

auto <- na.omit(auto)

attach(auto)

model <- lm(mpg ~ horsepower)
summary(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
```

i. Is there a relationship between the predictor and the response? We have an F statistic of 599.7, and a p-value very close to 0. Our F-statistic measures whether our model is a good fit by testing the null hypothesis that all regression coefficients are equal to 0. Thus, a high F-statistic means that the predictor (horsepower) has a very strong effect on mpg. Meanwhile, a very small p-value tells us that the probability of

obtaining the observed results given that the null hypothesis is true is very small, thus the null hypothesis can be rejected with very high confidence, indicating that horsepower is a significant predictor of mpg.

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

A R-squared of 0.6059 means that 60% of the variability of mpg can be explained by horsepower. It is a moderately strong relationship as a result, since 40% of the variance remains unexplained by horsepower alone.

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

The coefficient for horsepower (-0.157) tells us that the relationship between horsepower, our predictor, and mpg is negative. Thus, as horsepower increases, mpg decreases.

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

Y = 39.935861 + -0.157845 * X, where X is horsepower. Plugging in, 98, we get Y = 24.467 mpg

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

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

The 95% C.I.(mpg(98)) = (23.97, 24.96)

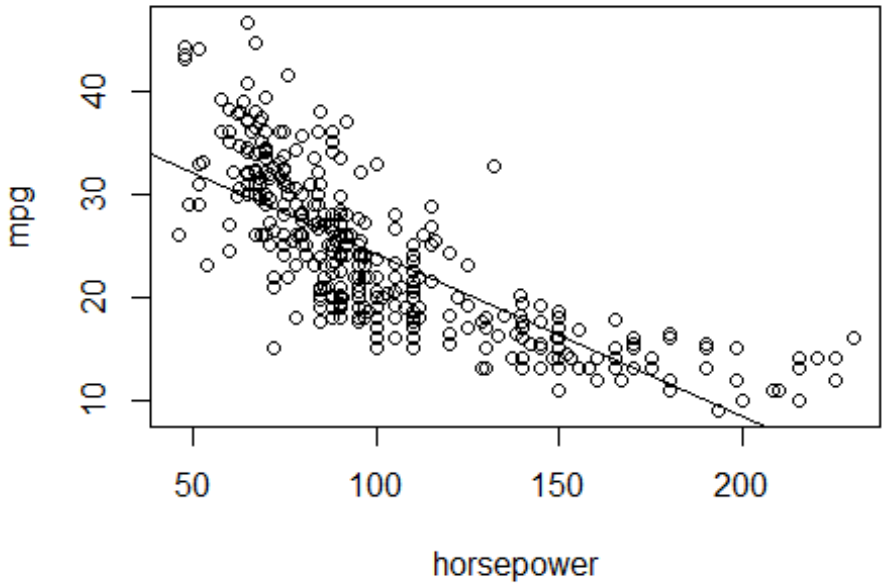
```
predict(model, data.frame(horsepower= 98),interval="prediction")
```

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

The 95% P.I(mpg(98)) = (14.80, 34.12)

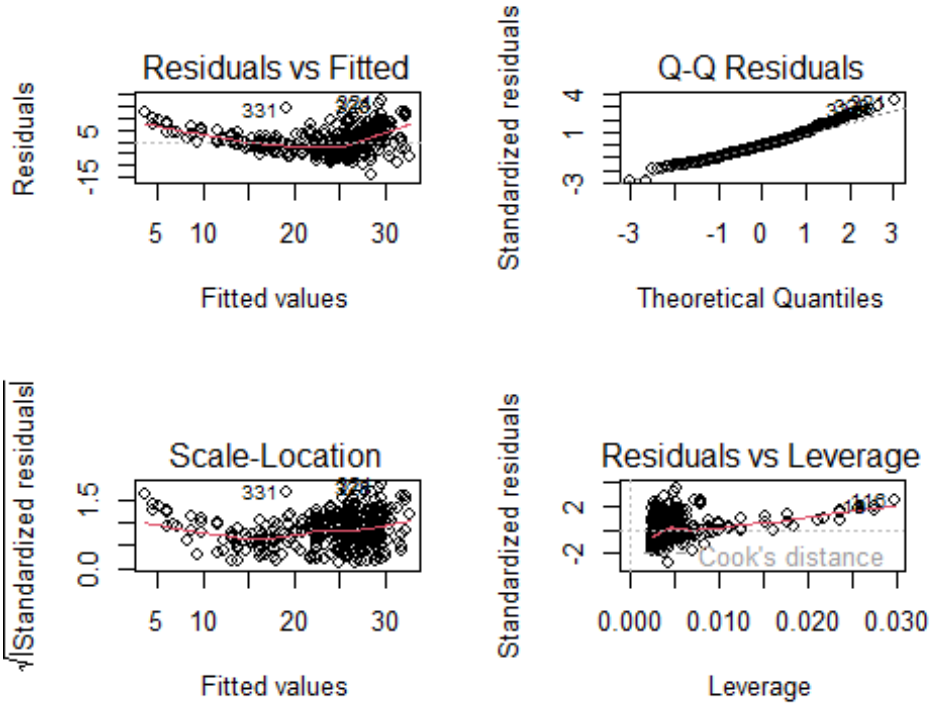
2. Plot the response and the predictor. Use the abline() function to display the least squares regression line.

```
plot(horsepower, mpg)
abline(model)
```



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(model)
```



The residuals vs Fitted plot is used to determine whether the assumption of linearity is a good one or a bad one. In this scenario, we see a u-shaped curve (non-random distribution), suggesting that the pattern of our data (horsepower with mpg) is not entirely linear or homoscedasticity.

The Q-Q residual plot is normally distributed, which is good, as the normality of residuals is one of the assumptions and that we picked an important predictor, with very little outliers.

The scale-location plot checks the assumption of homoscedasticity aka, whether the error variance is constant for all predictor values. It seems that the spread is pretty good across all fited values, meaning that the model meets the assumption of homoscedasticity.

Finally, the Residuals vs Leverage plot helps plot influential data points that had a significant impact on the regression model. There are certain points that have high leverage and influential observations (top right). Thus, it might be worth removing those points from the dataset.