Discussion - Week 12

Jai Jeffryes

4/15/2020

Contents


```r
library(MASS)
library(ISLR)
```

Introduction to Statistical Learning with Applications in R

Authors: Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

Between semesters, I started this classic of machine learning. I continued into part of this semester. Now, in DATA 605, our scope turns to regression analysis and I have an opportunity to return to *ISLR*. My work here is based on Exercise 8, p. 121.

Auto dataset

The dataset I examine comes from the `ISLR` package, which supports the textbook. The exercise as stated in the book is:

This question involves the use of 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. 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?

(b) Plot the response and the predictor. Use the `abline()` function to display the least squares regression line.
(c) Use the `plot()` function to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit.

```r
data(Auto)
lm.fit <- lm(mpg ~ horsepower, data = Auto)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = mpg ~ horsepower, data = 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 ***
## 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
```

Summary commentary

- There is a relationship between the predictor (`horsepower`) and the response variable (`mpg`).
- The relationship is strong and negatively correlated, indicated by a low p-value, near 0 for the `horsepower` variable, and the negative coeffecient, -0.157845.
- The R^2 statistic is 0.6059, indicating that 61% of the value for `mpg` is explained by the `horsepower` variable.

Confidence and prediction intervals

```r
# Confidence intervals for the coefficient estimates.
confint(lm.fit)
```

```
##                 2.5 %     97.5 %
## (Intercept) 38.525212 41.3465103
## horsepower  -0.170517 -0.1451725
```

```r
# Produce confidence intervals and prediction intervals for the prediction of mpg
# for a given value of horsepower.
# predict(lm.fit,data.frame(lstat=c(5,10,15)), interval = "confidence")

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

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

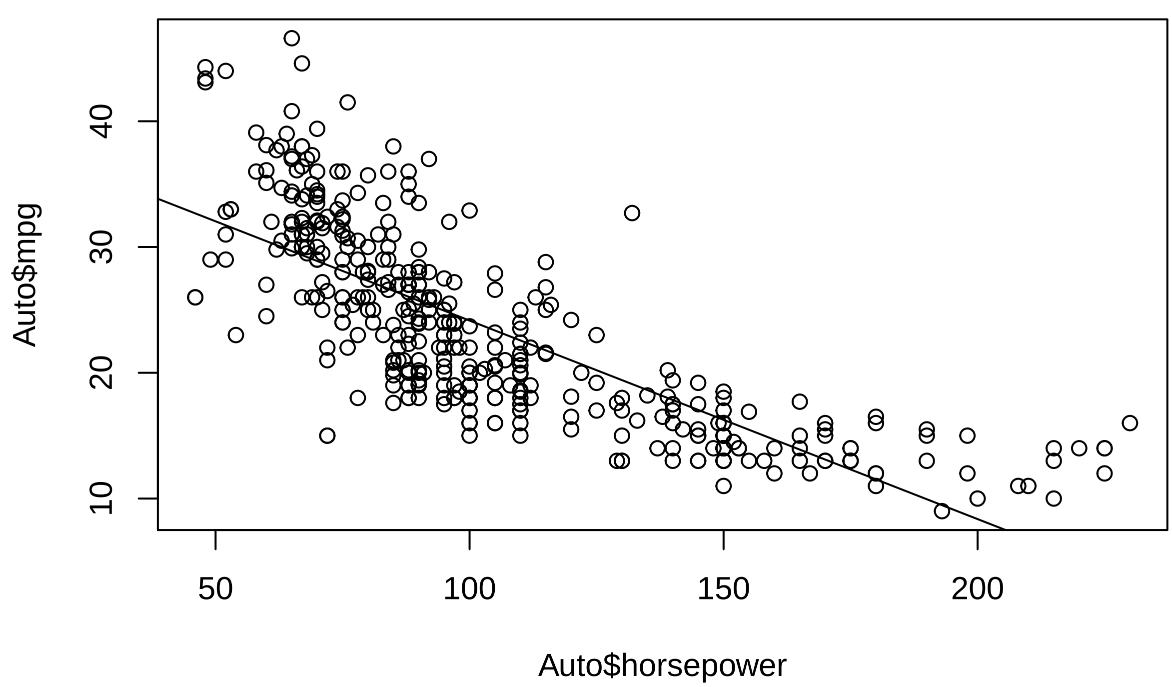
```r
predict(lm.fit, data.frame(horsepower = 98), interval = "prediction")
```

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

- Predicted `mpg` when `horsepower` = 98: 24.47.
- Confidence interval (23.97, 24.96). Prediction interval (14.81, 34.12)

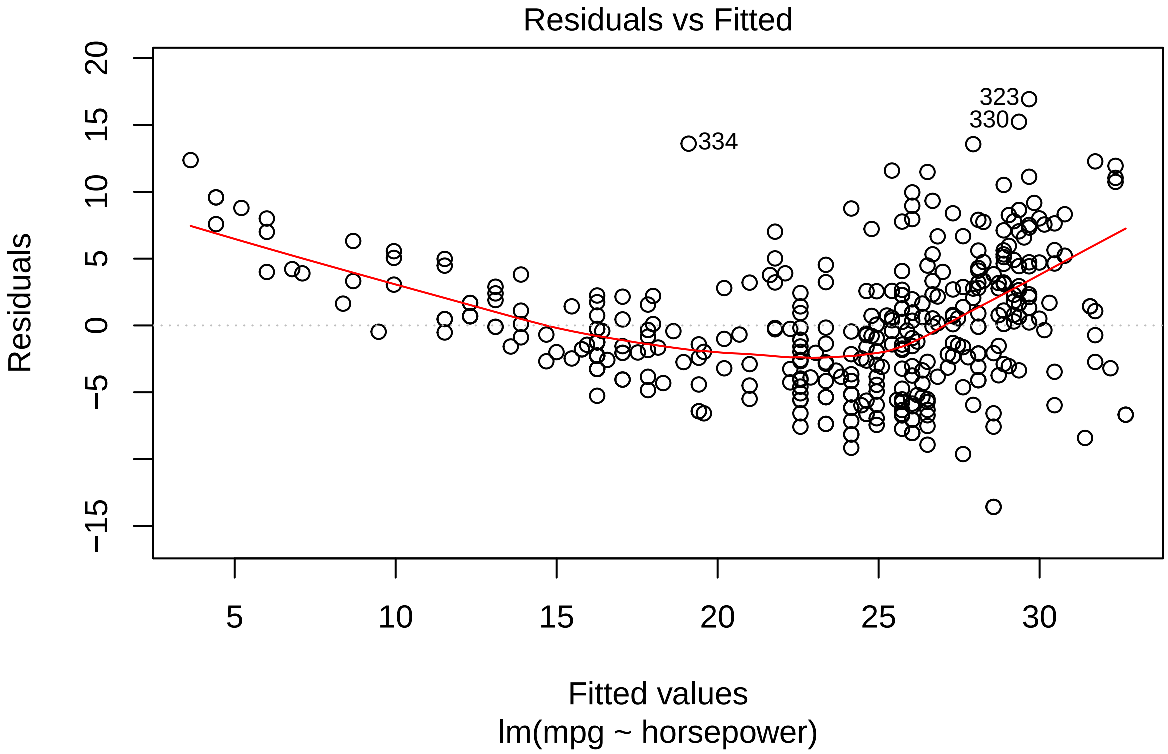
Regression plot

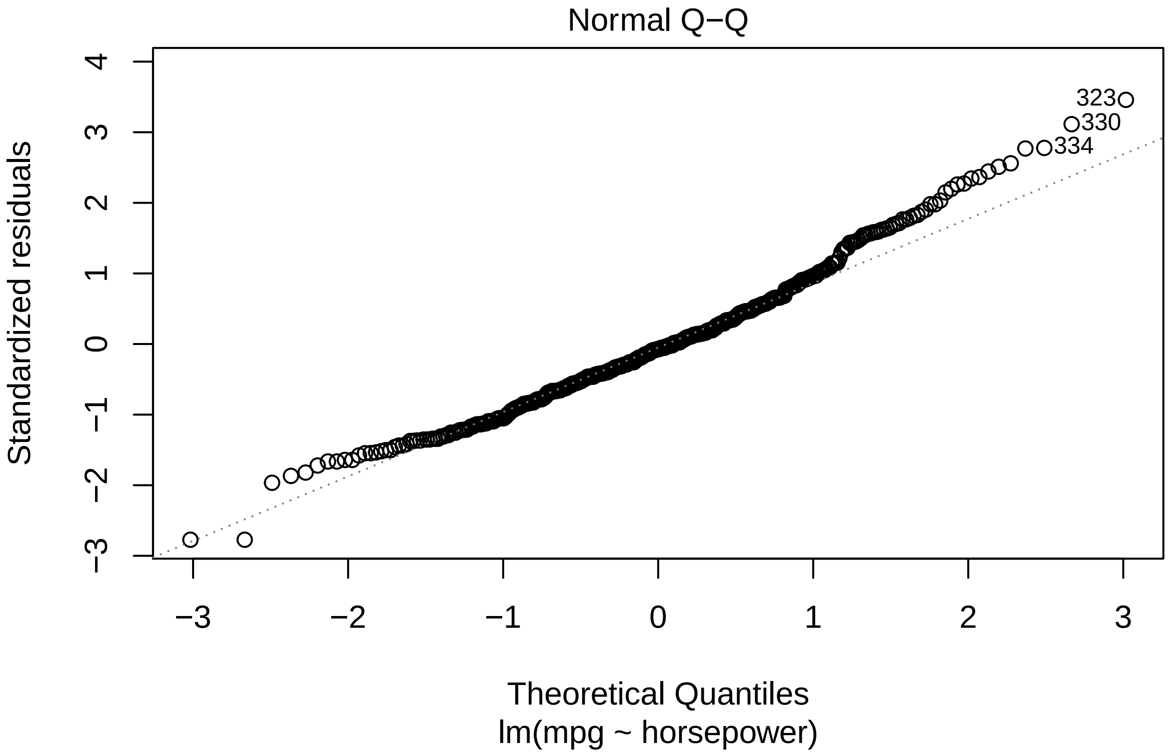
```r
plot(Auto$horsepower, Auto$mpg)
abline(lm.fit)
```

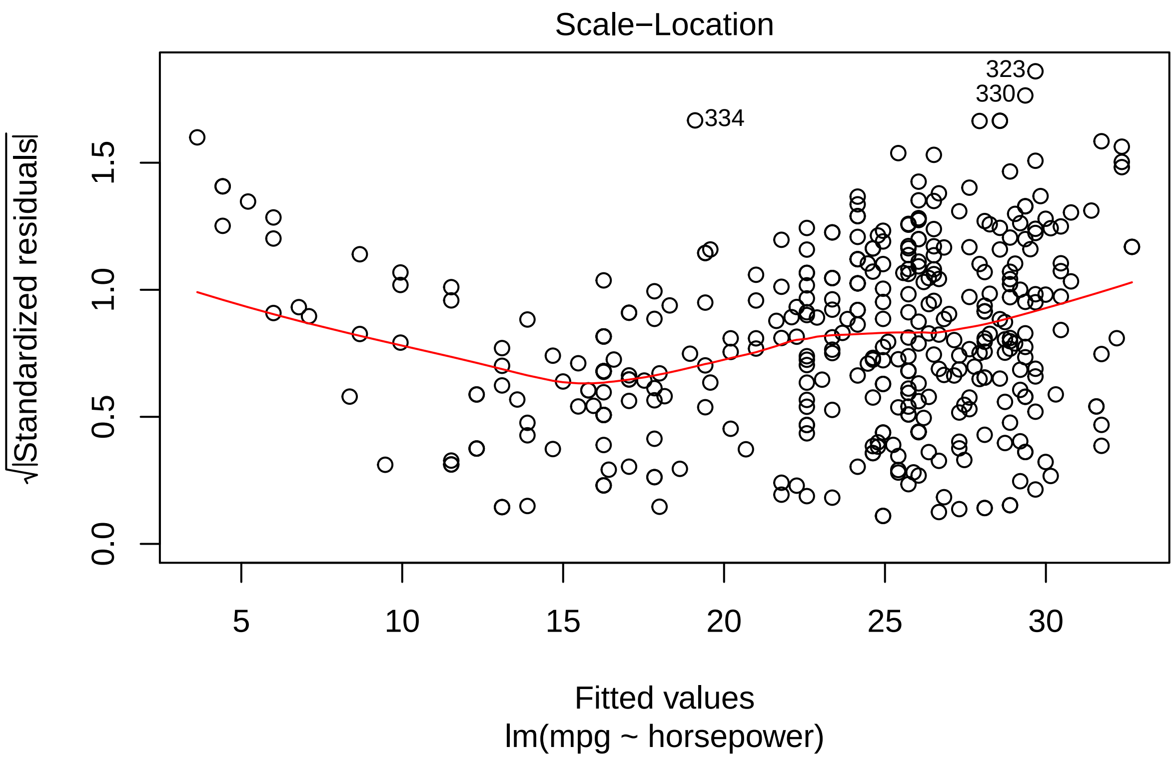


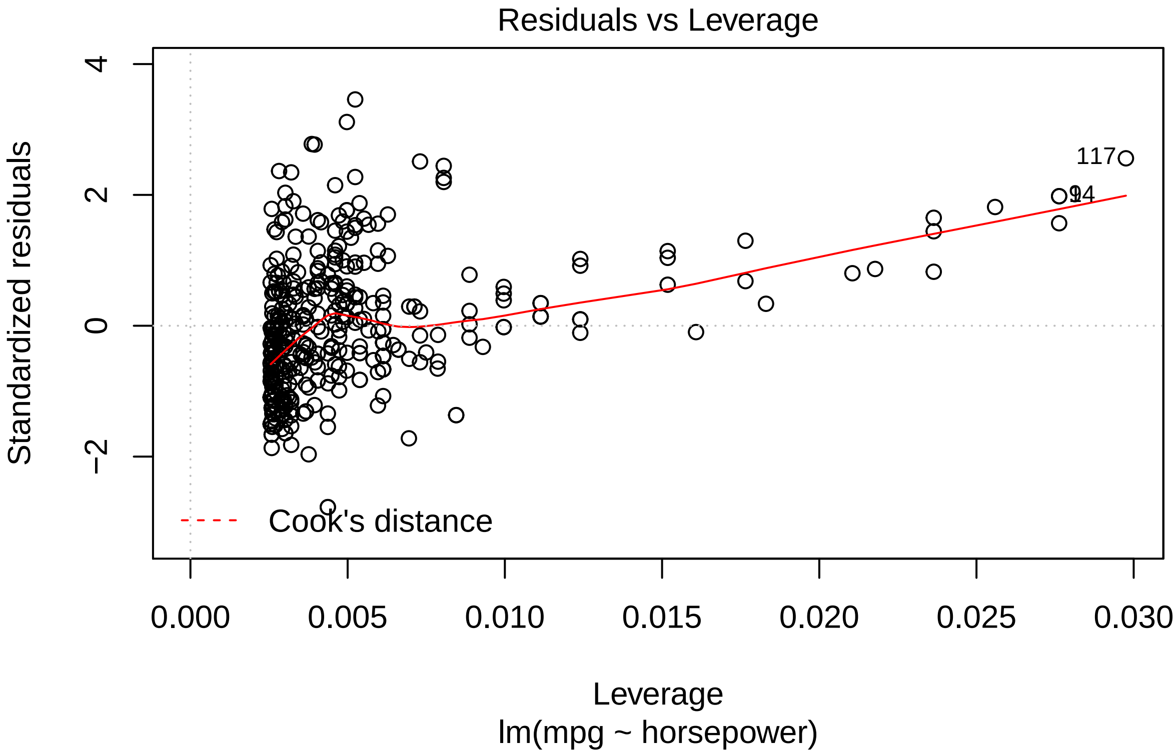
Diagnostic plots (residual analysis)

```r
plot(lm.fit)
```


Residuals vs Fitted
Residuals
Fitted values
lm(mpg ~ horsepower)
323
330
334
20
15
10
5
0
−5
−15
5
10
15
20
25
30


Normal Q−Q
Standardized residuals
4
3
2
1
0
−1
−2
−3
323
330
334
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(mpg ~ horsepower)


Scale−Location
323
330
334
1.5
1.0
0.5
0.0
√|Standardized residuals|
5
10
15
20
25
30
Fitted values
lm(mpg ~ horsepower)



- The first plot, "Residuals vs Fitted," indicates that residuals increase as `mpg` does. The strong pattern suggests non-linearity in the data in the right region of the plot.
- The plot "Normal Q-Q" further supports the possibility of non-linearity in the data. Although the diagnostic line aligns well to the left region of the a-b line, at about quantile 1.25 it diverges, which means the residuals in the region to the right do not fall on a normal distribution.
- "Scale - Location" shows that the variance of the residual have a dependency on `horsepower`. There is more variability in the residuals when `horsepower` is higher.
- "Residuals vs Leverage." This plot can be used to identify outliers with excessive influence or leverage on the regression. I think this example reveals little of that, but I need some more experience interpreting this plot.