HW3_Applied

Jon Griffith

2025-03-25

```
library(ISLR2)
```

```
## Warning: package 'ISLR2' was built under R version 4.4.3
```

8

```
head(Auto)
```

```
##   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)

```
mod <- lm(mpg ~ horsepower, data=Auto)
summary(mod)
```

```
##
## 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
```

Comments:

(i) There is a statistically significant relationship between the predictor and the response, with a very low p-value observed.

(ii) The relationship between the predictor and the response is very strong with a p-value of essentially zero. The predictor estimated effect on the response is approximately 24 standard deviations away from what we'd expect under the null hypothesis.

(iii) The relationship between the predictor and the response is negative

(iv)

```
conf_int <- predict(mod, data.frame(horsepower = 98), interval = 'confidence')
pred_int <- predict(mod, data.frame(horsepower = 98), interval = 'prediction')

paste('95% Confidence Interval:')
```

```
## [1] "95% Confidence Interval:"
```

```
conf_int
```

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

```
paste('95% Prediction Interval:')
```

```
## [1] "95% Prediction Interval:"
```

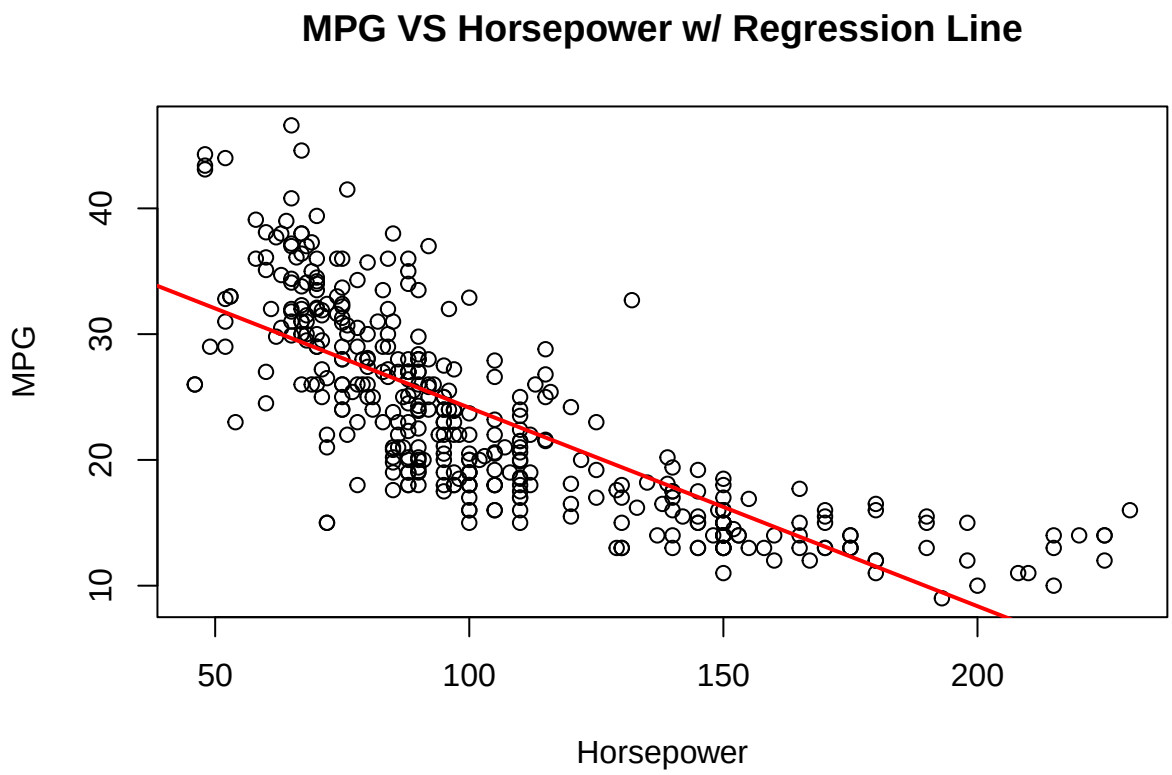
```
pred_int
```

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

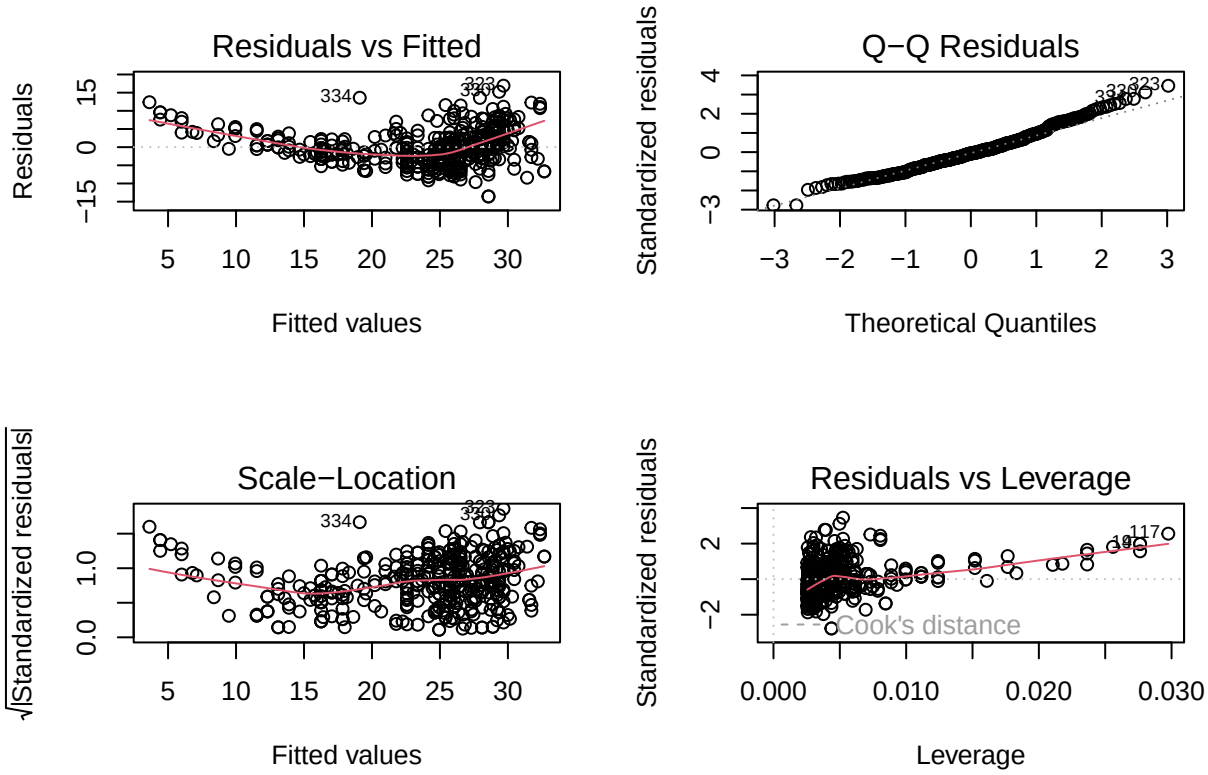
We can see the prediction above in the 'fit' column where a value of 98 for horsepower translates to a prediction of approximately 24.47 mpg. The first interval represents a 95% confidence interval while the second is a 95% prediction interval. Notice that the latter is more wide since this is an interval over a specific value prediction rather than the average value prediction.

(b)

```r
plot(Auto$horsepower, Auto$mpg,
     xlab = 'Horsepower',
     ylab = 'MPG',
     main = 'MPG VS Horsepower w/ Regression Line')
abline(mod, col='red', lwd=2)
```




```r
par(mfrow=c(2,2))
plot(mod)
```

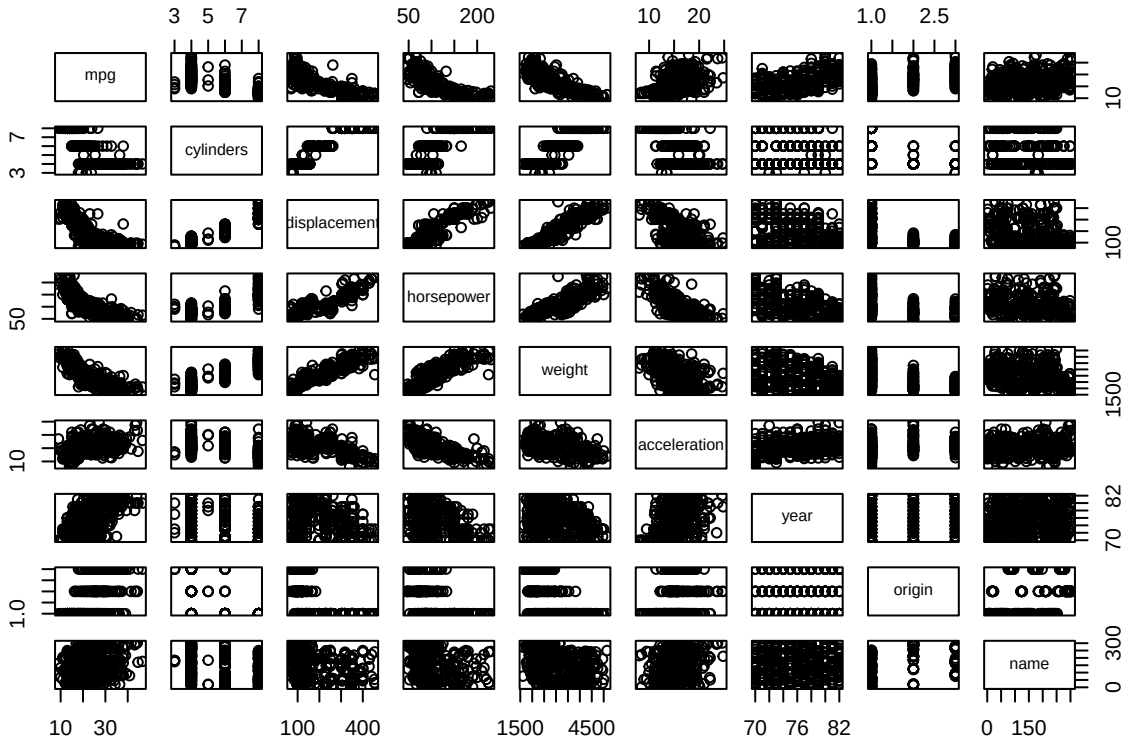


We can see that the residuals appear to be slightly heteroskedastic and nonlinear. This makes sense since the above plot shows a linear fit to a nonlinear scatter.

9 - Multiple linear regression on the Auto data set

(a)

```
pairs(Auto)
```



(b)

```
df_sub <- Auto[,-length(names(Auto))]
cormat <- cor(df_sub)
cormat
```

```
##                     mpg  cylinders displacement horsepower     weight
## mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.8322442
## cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.8975273
## displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944
## horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377
## weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.0000000
## acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.4168392
## year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.3091199
## origin        0.5652088 -0.5689316   -0.6145351 -0.4551715 -0.5850054
##              acceleration       year     origin
## mpg             0.4233285  0.5805410  0.5652088
## cylinders      -0.5046834 -0.3456474 -0.5689316
## displacement   -0.5438005 -0.3698552 -0.6145351
## horsepower     -0.6891955 -0.4163615 -0.4551715
## weight         -0.4168392 -0.3091199 -0.5850054
## acceleration    1.0000000  0.2903161  0.2127458
## year            0.2903161  1.0000000  0.1815277
## origin          0.2127458  0.1815277  1.0000000
```

(c)

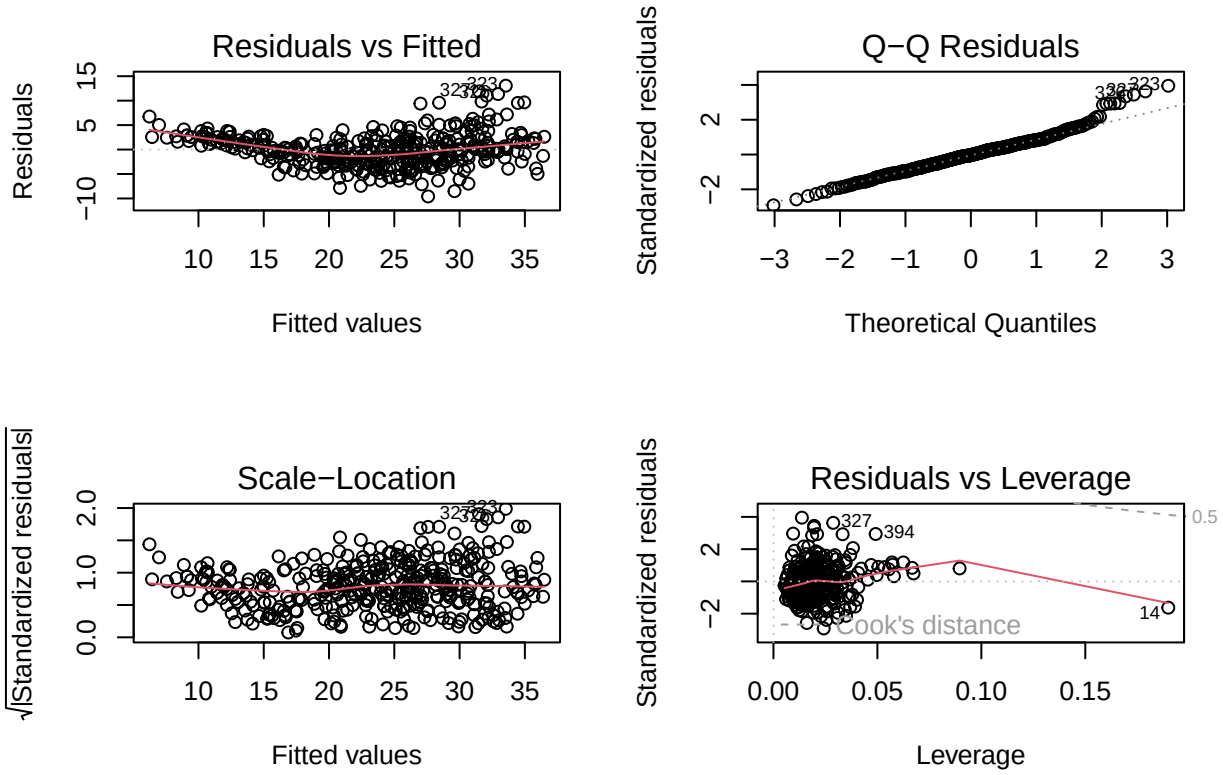
```
mod2 <- lm(mpg ~ .-name, data=Auto)
summary(mod2)
```

```
##
## Call:
## lm(formula = mpg ~ . - name, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5903 -2.1565 -0.1169  1.8690 13.0604
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
## cylinders     -0.493376   0.323282  -1.526  0.12780
## displacement   0.019896   0.007515   2.647  0.00844 **
## horsepower    -0.016951   0.013787  -1.230  0.21963
## weight        -0.006474   0.000652  -9.929  < 2e-16 ***
## acceleration   0.080576   0.098845   0.815  0.41548
## year           0.750773   0.050973  14.729  < 2e-16 ***
## origin         1.426141   0.278136   5.127 4.67e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.328 on 384 degrees of freedom
## Multiple R-squared:  0.8215, Adjusted R-squared:  0.8182
## F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

The summary shows that we have four statistically significant variables that have an effect on the response. These variables are displacement, weight, year, and origin. Now we see that horsepower is not statistically significant, despite being so in the simple linear regression. This is due to horsepower being highly correlated with other variables and essentially being a proxy for those variables in the single regression. Now that those other variables are included, we can see that horsepower does not appear to be causal.

(d)

```
par(mfrow=c(2,2))
plot(mod2)
```



We can see that the residuals appear to be more random than before with no obvious correlations. Though the variance still seems slightly heteroskedastic. We don't observe any high leverage points according to Cook's distance. All points fall within the Cook's distance bounds.

12 - Simple linear regression w/o an intercept

(a)

We will see $\hat{\beta}$ be equal between regressing X onto Y and vice versa when

$$\sum x_i^2 = \sum y_i^2.$$

(b)

```
n <- 100

x <- rnorm(n)
y <- 5*x + rnorm(n)

B1 <- sum(x*y) / sum(x^2)
B2 <- sum(x*y) / sum(y^2)

paste('Beta of Y regressed on X:', B1)
```

```
## [1] "Beta of Y regressed on X: 5.09226767975261"
```

```
paste('Beta of X regressed on Y:', B2)
```

```
## [1] "Beta of X regressed on Y: 0.188182605716224"
```

(c)

```
set.seed(560)
n <- 100

x <- rnorm(100)
y <- sample(x, n)

B1 <- sum(x*y) / sum(x^2)
B2 <- sum(x*y) / sum(y^2)

paste('Beta of Y regressed on X:', B1)
```

```
## [1] "Beta of Y regressed on X: 0.0724530301177476"
```

```
paste('Beta of X regressed on Y:', B2)
```

```
## [1] "Beta of X regressed on Y: 0.0724530301177476"
```