linear regression modelling

Justin Lo

2023-08-15

I will be using the 'Auto' dataset from the ISLR package. I aim to find relationship between the variables and build multiple linear regression models.

Loading the Auto dataset

```
options(repos = "https://cran.rstudio.com/")
install.packages('ISLR')
```

```
##
## The downloaded binary packages are in
##  /var/folders/f6/_zxsgwtd23164g9bpmgcn95r0000gn/T//Rtmpwm1Muq/downloaded_packages
```

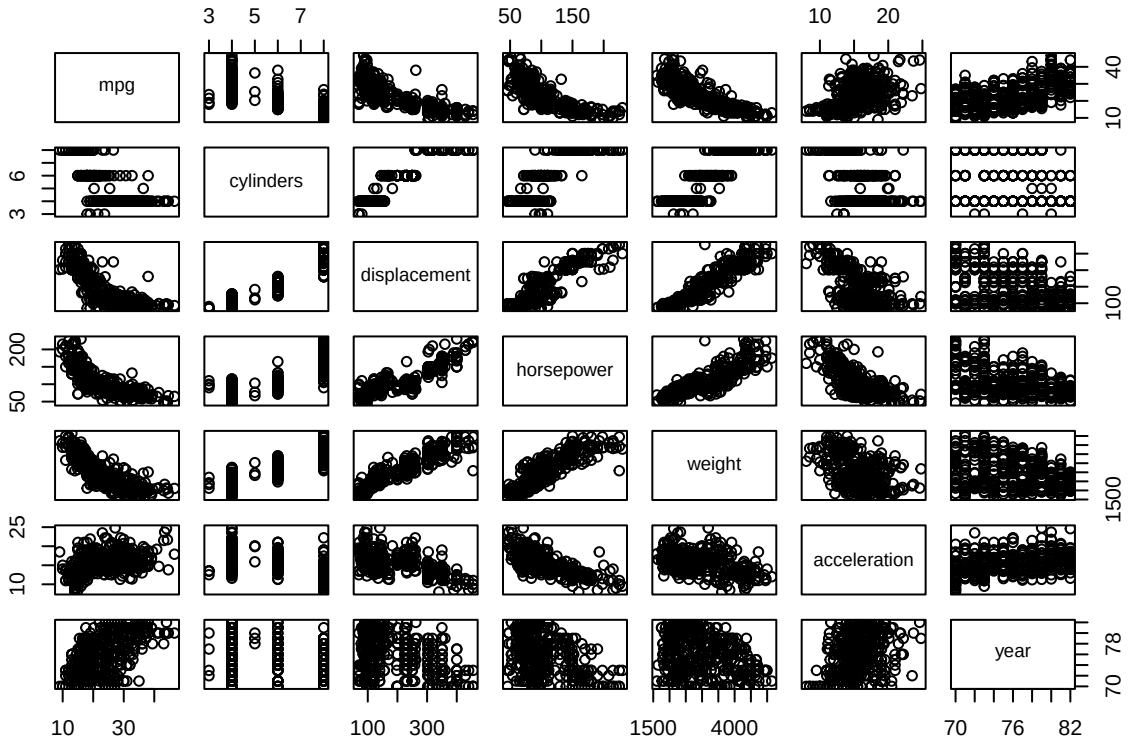
```
library('ISLR')
data("Auto", package = 'ISLR')
```

First, let's do basic visualisation and correlation findings.

```
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
```

```
pairs(Auto[,1:7])
```




```
cor(Auto[,1:7])
```

```
##                     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
##              acceleration       year
## mpg             0.4233285  0.5805410
## cylinders      -0.5046834 -0.3456474
## displacement   -0.5438005 -0.3698552
## horsepower     -0.6891955 -0.4163615
## weight         -0.4168392 -0.3091199
## acceleration    1.0000000  0.2903161
## year            0.2903161  1.0000000
```

To build a a multiple linear regression with `mpg` as the response and all other variables except `name` as the predictors.

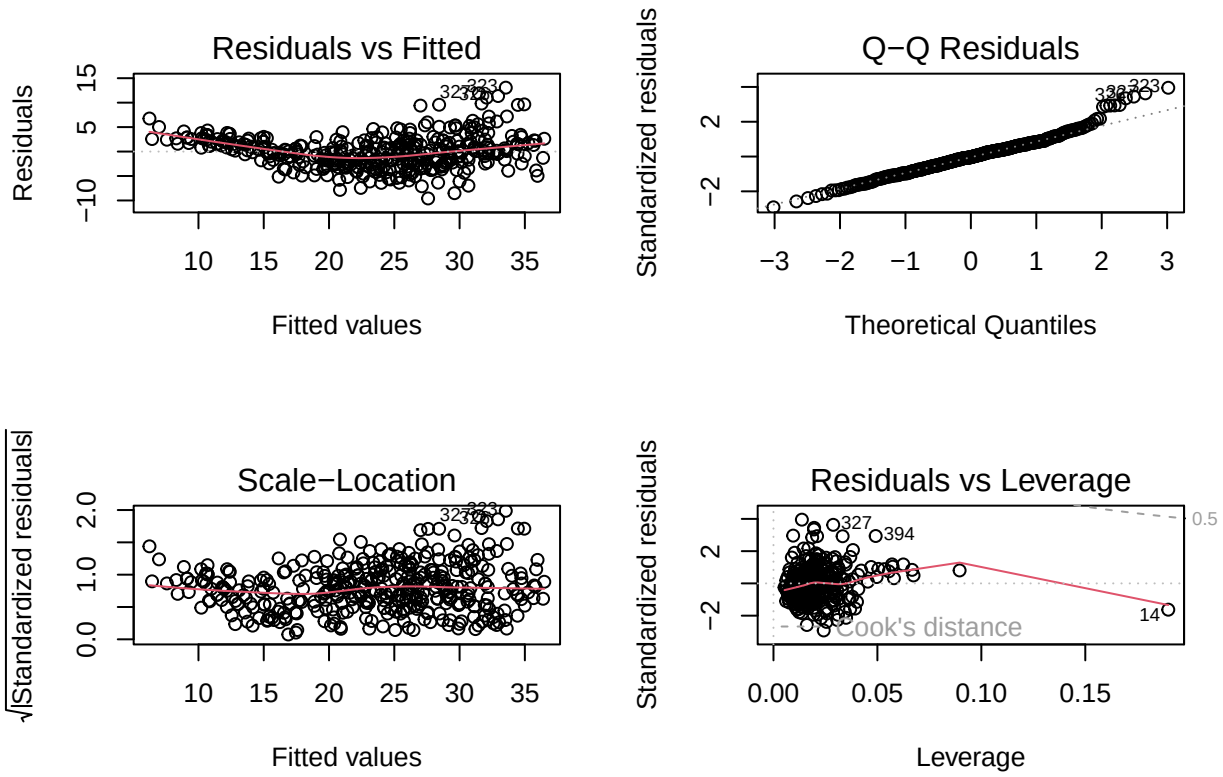
```
lm.fit1 <-  lm(mpg ~ . - name, data = Auto)
summary(lm.fit1)
```

```
##
## 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
```

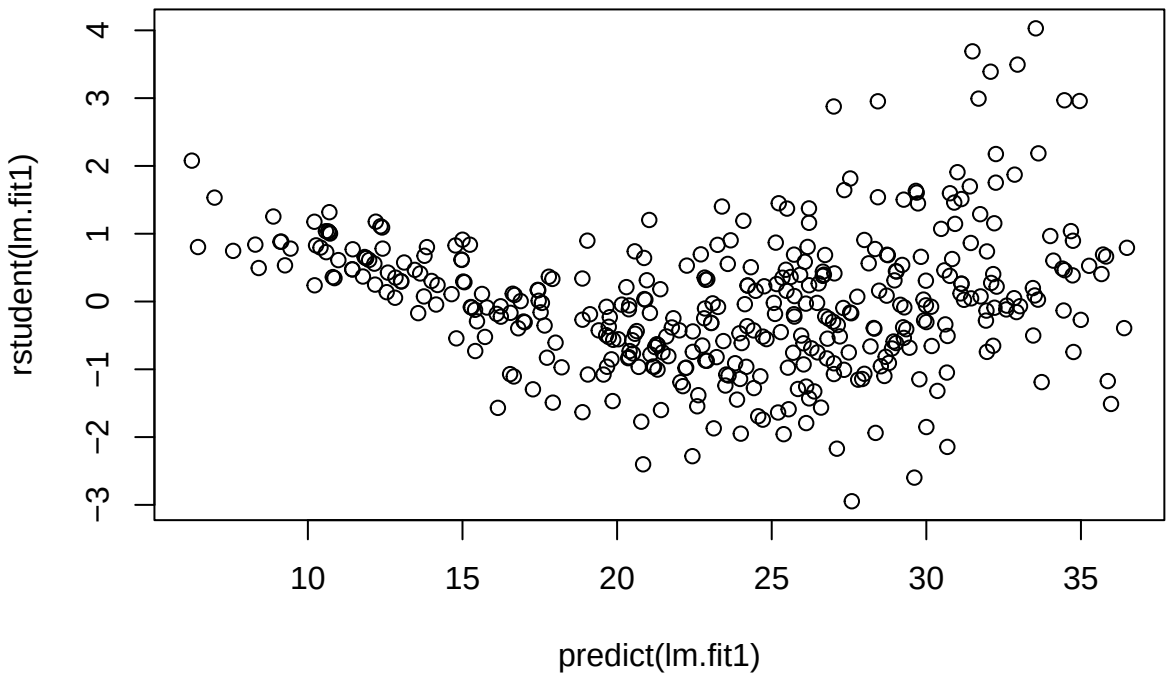
There is a relatioship between the predictors and the response by testing the null hypothesis of whether all the regression coefficients are zero. The F-statistic is far from 1 (with a small p-value), indicating evidence against the null hypothesis. Also, looking at the p-values associated with each predictor's t-statistic, we see that displacement, weight, year, and origin have a statistically significant relationship, while cylinders, horsepower, and acceleration do not.

To further determine whether the model is suitable for our dataset, we produce diagnostic plots of the linear regression fit to check linear assumptions.

```
par(mfrow = c(2, 2))
plot(lm.fit1)
```




```
par(mfrow = c(1,1))
plot(predict(lm.fit1), rstudent(lm.fit1))
```



From the residuals vs leverage plot, we see no points lie outside the cook's distance, meaning no particular observation carries significant influence.

From the scale - lcation plot, we can see that the red line is roughly horizontal, meaning that the assumption of equal variance is met.

From the QQ plot, we can see that most observations lie on the line, meaning that the assumption of error following a normal distribution is met.

From the residuals vs fitted plot, we can see that the red line isn't quite horizontal, meaning that perhaps the residuals do not follow a linear pattern. Which tells us that linear model may not be the best fit.

From the scatterplot of prediction of the model and the standardized residuals, we can see that there are data with a value greater than 3 which means possible outliers.

To further improve the model, the interactions between the predictors are considered.

```
lm.fit2 <-  lm(mpg ~ cylinders * displacement + displacement * weight, data = Auto)
summary(lm.fit2)
```

```
##
## Call:
## lm(formula = mpg ~ cylinders * displacement + displacement *
##     weight, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.2934  -2.5184  -0.3476   1.8399  17.7723
##
```

```
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              5.262e+01  2.237e+00  23.519  < 2e-16 ***
## cylinders                7.606e-01  7.669e-01   0.992    0.322
## displacement            -7.351e-02  1.669e-02  -4.403 1.38e-05 ***
## weight                  -9.888e-03  1.329e-03  -7.438 6.69e-13 ***
## cylinders:displacement -2.986e-03  3.426e-03  -0.872    0.384
## displacement:weight     2.128e-05  5.002e-06   4.254 2.64e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.103 on 386 degrees of freedom
## Multiple R-squared:  0.7272, Adjusted R-squared:  0.7237
## F-statistic: 205.8 on 5 and 386 DF,  p-value: < 2.2e-16
```

Interaction between displacement and weight is statistically signifcant, while the interaction between cylinders and displacement is not.