Data Analysis and Statistical Modeling

2nd Project

Data analysis through statistical methods

Group 9 :

Pedro Matias 90162

Aurel Vitelariu 96728

Duarte Flôr 98931

Profª Isabel Rodrigues

Instituto Superior Técnico

Universidade de Lisboa

Portugal, 2744-016 Porto Salvo

LETI 2023/24

Index:

Introduction

In this project for the course in Data Analysis and Statistical Modelling, we aim to examine the 'Auto' dataset using various statistical techniques. The dataset concerns characteristics of automobiles in the late 20th century. The variables, all numerical, are the following:

1. 'mpg': Miles per gallon, indicating fuel efficiency.
2. 'cylinders': Number of cylinders in the engine, ranging from 4 to 8.
3. 'displacement': Engine displacement in cubic inches.
4. 'horsepower': Engine horsepower.
5. 'weight': Vehicle weight in pounds.
6. 'acceleration': Time taken to accelerate from 0 to 60 mph (seconds).
7. 'year': Model year (modulo 100).
8. 'origin': Origin of the car (1 for American, 2 for European, 3 for Japanese).

This dataset offers an opportunity to understand the factors affecting vehicle performance and efficiency, reflecting the technological and design trends of the period.

1. Summary Statistics

mpg	cylinders	displacement	horsepower
Min. : 9.00	Min. :4.00	Min. : 97.0	Min. : 46.00
1st Qu.:14.00	1st Qu.:4.50	1st Qu.:154.5	1st Qu.: 91.25
Median :17.50	Median :7.00	Median :280.0	Median :121.50
Mean :18.08	Mean :6.48	Mean :268.8	Mean :135.34
3rd Qu.:22.00	3rd Qu.:8.00	3rd Qu.:357.8	3rd Qu.:173.75
Max. :28.00	Max. :8.00	Max. :455.0	Max. :225.00

weight	acceleration	year	origin
Min. :1835	Min. : 8.00	Min. :70.00	Min. :1.00
1st Qu.:2599	1st Qu.:11.50	1st Qu.:70.00	1st Qu.:1.00
Median :3381	Median :13.75	Median :70.00	Median :1.00
Mean :3366	Mean :13.40	Mean :70.42	Mean :1.28
3rd Qu.:4195	3rd Qu.:15.38	3rd Qu.:71.00	3rd Qu.:1.00
Max. :5140	Max. :20.50	Max. :71.00	Max. :3.00

Figure 1: Summary Statistics

Looking at the variables 'cylinders' and 'year', we can observe that measures of location such as median and mean are close together, and the range is narrow. This observation indicates that there is a low standard deviation, suggesting that the values are clustered around the mean. We can conclude that there is less variability in the number of cylinders and the model year within the dataset.

We can observe a symmetry in the 'mpg' variable since the median is almost equidistant from the first quartile and the third quartile, which suggests that there is no pronounced skewness in the 'mpg' values. We can infer that the distribution of fuel efficiency across the cars in the dataset does not show a strong skew in either direction, presenting the same behavior as a normal distribution. This contrasts with the 'displacement' and 'horsepower' variables, where both variables exhibit a larger range and IQR, signaling a wider spread of engine sizes and power. Even so, since the means are close to the median in both cases we can conclude that the extreme values do not heavily influence the distribution, despite the broader spread.

The largest spread is present within the 'weight' variable, but similarly to the 'displacement' and 'horsepower', the ample variation does not reflect a heavy influence in the distribution, seen by the proximity between the mean and median values.

The 'origin' variable seems to be predominantly American (1). Although there is limited variation, the instances of European (2) and Japanese (3) cars slightly impact the mean.

Finally, we can conclude that there are no extreme outliers, namely we do not see the observations of large scale vehicles such as cargo trucks or prototype race cars.

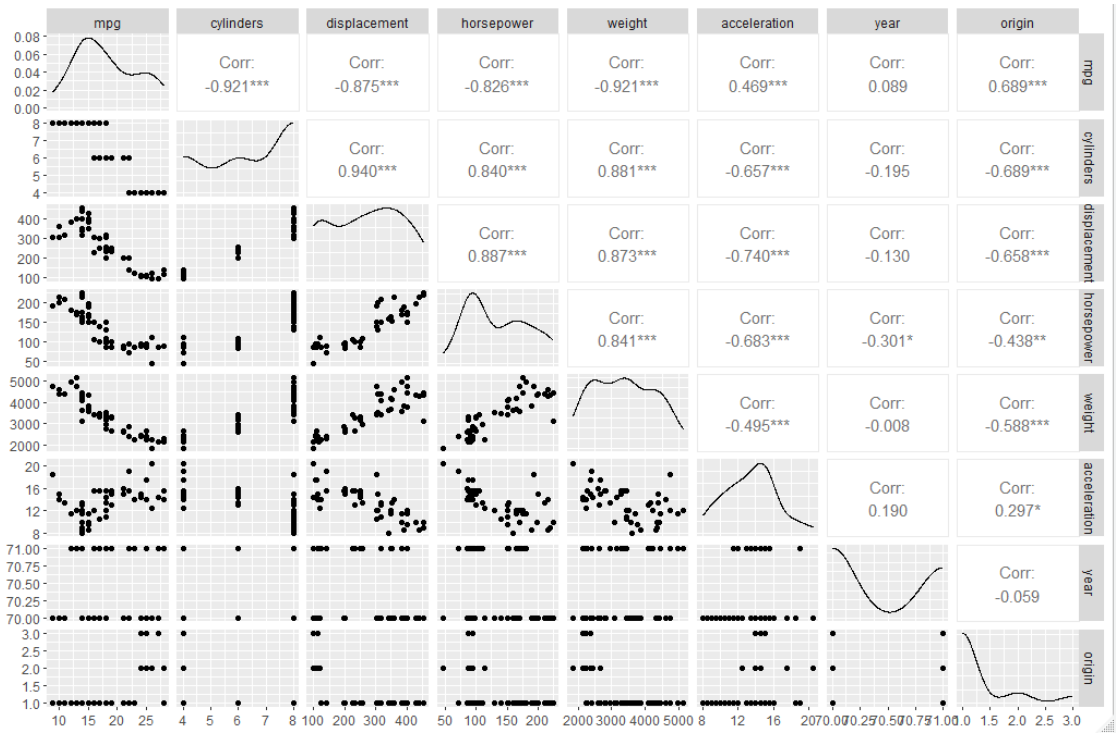

Figure 2: Pairs Panels

	mpg	cylinders	displacement	horsepower	weight	acceleration	year	origin
mpg	27.1363265	-8.1208163	-527.32735	-211.803265	-4312.486531	6.8857143	0.23102041	2.18122449
cylinders	-8.1208163	2.8669388	184.32163	69.996735	1341.488980	-3.1346939	-0.16489796	-0.70857143
displacement	-527.3273469	184.3216327	13398.96163	5050.307755	90877.806531	-241.3816327	-7.48897959	-46.27836735
horsepower	-211.8032653	69.9967347	5050.30776	2420.106531	37211.820000	-94.7204082	-7.39061224	-13.09714286
weight	-4312.4865306	1341.4889796	90877.80653	37211.820000	808211.763673	-1253.5346939	-3.74816327	-321.25387755
acceleration	6.8857143	-3.1346939	-241.38163	-94.720408	-1253.534694	7.9489796	0.26734694	0.50816327
year	0.2310204	-0.1648980	-7.48898	-7.390612	-3.748163	0.2673469	0.24857143	-0.01795918
origin	2.1812245	-0.7085714	-46.27837	-13.097143	-321.253878	0.5081633	-0.01795918	0.36897959

Figure 3: Variance Matrix

	mpg	cylinders	displacement	horsepower	weight	acceleration	year	origin
mpg	1.00000000	-0.9206932	-0.8745189	-0.8264943	-0.920851584	0.4688326	0.088950668	0.6893244
cylinders	-0.92069323	1.0000000	0.9404409	0.8403328	0.881282845	-0.6566444	-0.195335180	-0.6889271
displacement	-0.87451889	0.9404409	1.0000000	0.8868803	0.873292400	-0.7396273	-0.129766165	-0.6581749
horsepower	-0.82649434	0.8403328	0.8868803	1.0000000	0.841397246	-0.6829213	-0.301326669	-0.4382866
weight	-0.92085158	0.8812828	0.8732924	0.8413972	1.000000000	-0.4945590	-0.008362388	-0.5882807
acceleration	0.46883257	-0.6566444	-0.7396273	-0.6829213	-0.494558971	1.0000000	0.190192734	0.2967197
year	0.08895067	-0.1953352	-0.1297662	-0.3013267	-0.008362388	0.1901927	1.000000000	-0.0593007
origin	0.68932442	-0.6889271	-0.6581749	-0.4382866	-0.588280729	0.2967197	-0.059300699	1.0000000

Figure 4: Correlation Matrix

By observing the correlation matrix we can observe which and how variables are related to one another with values ranging from -1 to 1. The most noticeable correlations are the following:

- **'mpg' and 'weight':** A high negative correlation (-0.9208) indicates that heavier cars have a lower fuel efficiency or a higher fuel consumption.
- **'mpg' and 'displacement', 'mpg' and 'horsepower':** Also high negative correlations (-0.8745 , -0.8264) suggest that cars with larger engine size and more horsepower are less fuel efficient.
- **'displacement', 'horsepower', and 'weight':** High positive correlations with each other, leading to the conclusion that heavier vehicles tend to have larger engines with more horsepower.
- **'cylinders' and 'displacement':** Highest positive correlation (0.9404) that confirms that engines with more cylinders are indeed larger in volume.

	VarName	TrimmedMean
1	mpg	18.075128
2	cylinders	6.489048
3	displacement	268.760000
4	horsepower	135.173368
5	weight	3366.460000
6	acceleration	13.381218
7	year	NaN
8	origin	NaN

Figure 5: Winsorized Means

By comparing the winsorized means to the original means we notice that they are similar which leads us to conclude that there are no extreme outliers impacting the mean and consequently the dataset and other methods of calculation.

2a. Fit a regression model to the data set and test its significance

2.a.1 Testing the overall significance of a regression:

To evaluate the overall significance of a regression, we must determine whether we should reject or accept the following null hypothesis:

$$H_0 : \beta_1 = \beta_2 = \ \ldots \ = \beta_{p-1} = 0$$

To do so, we can look at the value of the F-statistic. The higher the F-statistic, the more statistical evidence we have to reject the null hypothesis. For our evaluation of the different regression models, we will be using a significance level of alpha=0.01 for the F-statistic.

2.a.2 Testing the individual significance of each explanatory variable:

Rejection of the null in the F-test means that at least one of the explanatory variables is useful, but it has no meaning when it comes to single variables, since the F-test is a test on the entire model. In our analysis, we also want to ascertain whether any given single variable is contributing to the quality of the model. To do so, we use the following hypothesis:

$$H_0 : \beta_1 = \beta_2 = \ \ldots \ = \beta_r = 0 \quad r < (p-1)$$

To decide whether we should reject a given variable, we settled on a significance level alpha=0.05 for the p-value of the t-test of that variable.

2.a.3 Finding the best subset of regressors:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  88.1914525 40.9568090   2.153 0.037085 *
cylinders    -1.8554467  0.4601097  -4.033 0.000228 ***
displacement  0.0036874  0.0079945   0.461 0.647005
horsepower   -0.0313082  0.0127249  -2.460 0.018074 *
weight       -0.0014804  0.0007306  -2.026 0.049118 *
acceleration -0.3973985  0.1403159  -2.832 0.007070 **
year         -0.6492297  0.5674529  -1.144 0.259056
origin        0.9263427  0.5777388   1.603 0.116343
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.474 on 42 degrees of freedom
Multiple R-squared:  0.9313,    Adjusted R-squared:  0.9199
F-statistic: 81.38 on 7 and 42 DF,  p-value: < 2.2e-16
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  84.6660928 39.8672768   2.124  0.03949 *
cylinders    -1.7606663  0.4078873  -4.317 9.14e-05 ***
horsepower   -0.0287331  0.0113296  -2.536  0.01492 *
weight       -0.0014294  0.0007155  -1.998  0.05211 .
acceleration -0.4288316  0.1215273  -3.529  0.00101 **
year         -0.5934321  0.5493096  -1.080  0.28602
origin        0.8277284  0.5317864   1.557  0.12692
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.461 on 43 degrees of freedom
Multiple R-squared:  0.931,     Adjusted R-squared:  0.9214
F-statistic: 96.67 on 6 and 43 DF,  p-value: < 2.2e-16
```

Figure 6 & 7 : First iteration (Left) and Second iteration (Right)

Having defined the significance levels for the statistics we will analyze, we can now start to search for the best subset of regressors. We used a backwards selection.

First iteration

A first but naive approach to fit a regression model is to use all the variables apart from the response variable as predictors.

As we can see, the p-value of the F-statistic is lower than 2.2e-16 and thus lower than the significance level of 0.01 that we defined earlier, which means that we should reject the null hypothesis

It is also of note that the coefficients of determination, adjusted and non adjusted are both close to 1. In the case of the Multiple R-squared, it means that 93.13% of the variability in mpg can be explained by the set of regressors of this model. This indicates a strong linear dependency between the dependent variable and the chosen predictors.

However, we can see multiple variables, namely displacement, year and origin where the p-value of the t-statistic is higher than the significance level of 0.05, which suggests that they have little statistical significance. Thus, we can improve our model if we remove some of them.

Second iteration

For this second iteration, we removed the displacement variable, as it had the highest p-value for its t-test. The F-statistic is still significant, since the p-value is lower than the significance level we defined for it.

As for the coefficients of determination, we observe a slight reduction in the R^2 and a slight increase in the adjusted R^2. The R^2 artificially decreases with a decrease in the number of explanatory variables, so this was expected. On the other hand, the adjusted R^2 was designed to

account for models with different degrees of freedom. The increase in its value indicates that the displacement variable did not contribute a lot of information to explain mpg, as we expected.

There are still individual explanatory variables with a high p-value for the t-statistic, namely year, origin and weight and so there is still room for improvement in the model.

Final iteration

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  43.9154963  2.3155947  18.965  < 2e-16 ***
cylinders    -2.0388707  0.3260589  -6.253 1.21e-07 ***
weight       -0.0024559  0.0005329  -4.609 3.23e-05 ***
acceleration -0.3250775  0.1064735  -3.053  0.00376 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.54 on 46 degrees of freedom
Multiple R-squared:  0.918,     Adjusted R-squared:  0.9126
F-statistic: 171.5 on 3 and 46 DF,  p-value: < 2.2e-16
```

Figure 8 : Final Iteration

We continued with the same method of removing the single explanatory variable with the highest p-value associated with its t-statistic (all of the iterations are available in our source code), and got to this model, which is the first one where there is no variable with a p-value bigger than our significance level of 0.05.

The p-value of the F-statistic is lower than 0.01 and thus the overall regression model is statistically significant, and we can reject the null hypothesis that none of the explanatory variables has a non zero coefficient.

The R^2 and adjusted R^2 are slightly smaller than the ones of the second iteration, but not significantly so. The values we obtained in this final iteration are still close to 1, suggesting a very strong linear dependency between mpg and the variables cylinders, weight and acceleration. We also know that even the adjusted R^2 doesn't penalize enough the models with too many variables. Combining this with the fact that the model is much simpler than that of the second iteration, as we have removed 3 explanatory variables since then, following the Principle of Parsimony, we conclude that this model uses the best subset of regressors.

2b Checking model adequacy

2b.1 Residual analysis:

There are some assumptions underneath the multiple linear regression model that we should confirm to be true for the model we obtain. Having now decided on the subset of regressors, we are in a condition to verify if these assumptions hold for our model. It is assumed that the error has an expected value of 0, constant variance and is normally distributed.

When we calculated the mean for the residuals we obtained the value of 4.450867e-17 which is indeed very close to 0. This is an indication that our model does not break the theoretical assumptions.

If we plot the residuals against the fitted model and predictor variables we can check that our model also has constant variance around 0. If we can detect any pattern in the plot it might mean that model does not describe the data correctly. In our residual plots, we drew a red line at y=0 (the expected value of the error) to facilitate the observation.

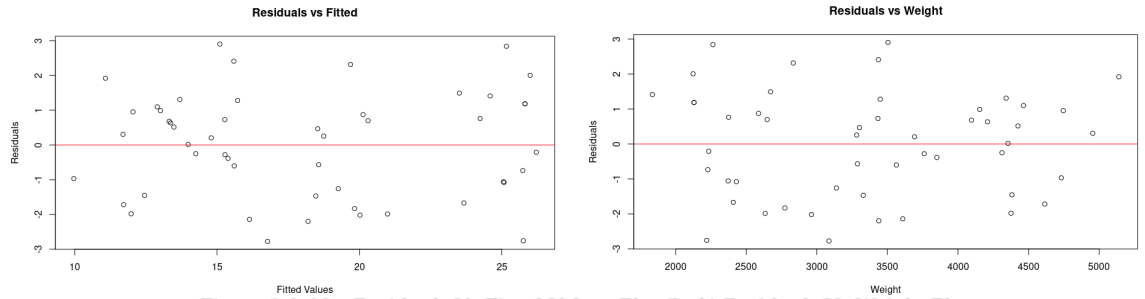

Figure 9 & 10 : Residuals Vs Fitted Values Plot (Left) Residuals Vs Weight Plot (Right)

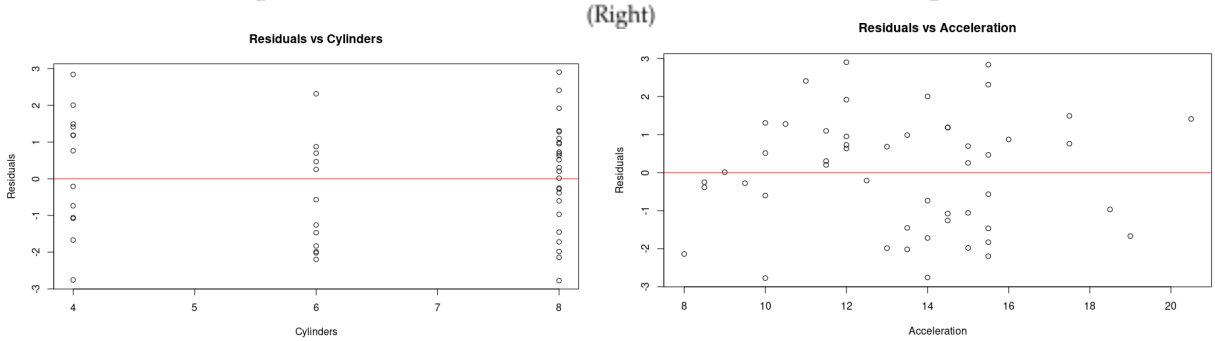

Figure 11 & 12 : Residuals Vs Cylinders Plot (Left) Residuals Vs Acceleration Plot (Right)

The residuals seem to be randomly scattered around the zero line in all the plots, which indicates that the model is appropriate for the data and that the error has constant variance. The only potentially unusual looking plot is that of the residuals against the cylinders. We note that the cylinders variable represents the number of cylinders of the car, so the vertical lines we see are merely a result of all the cars in our data set having either 4, 6 or 8 cylinders.

We can look at the histogram of the residuals and see if it resembles a symmetrical bell shape, which might help us identify whether the error has a normal distribution.

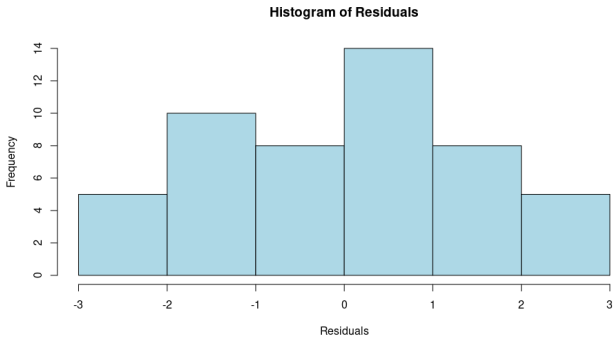

Figure 13: Histogram of Residuals

We can see that there is a central peak, as expected. The histogram is not perfectly symmetrical, but is also not extremely asymmetrical, so further analysis must be performed to check the normality assumption. We did this with a QQ-plot of the residuals. In the QQ-plot, if the residuals are normally distributed, the points will be close to the line y=x.

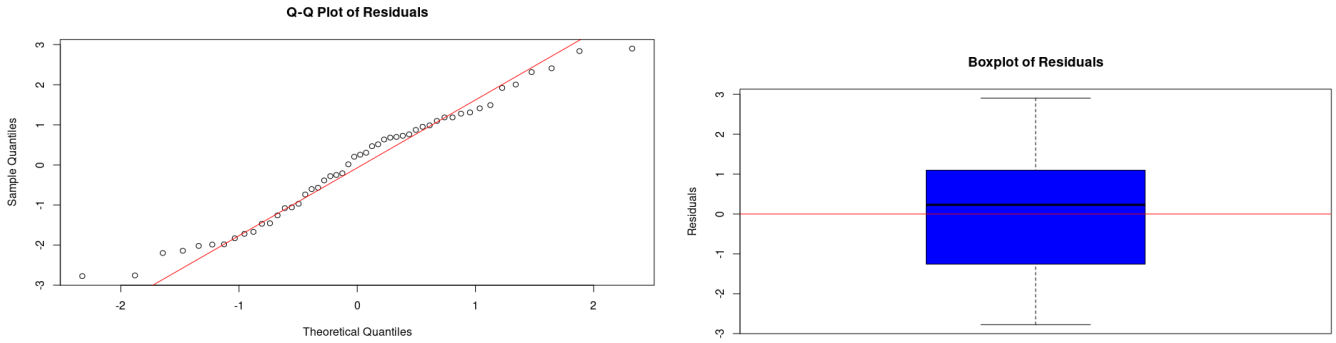

Figure 14 & 15 : Q-Q Plot of Residuals (Left) Boxplot of Residuals (Right)

The bulk of the data seems to approximately lie on the red line, albeit with slight deviations especially in the tails, so it seems like the residuals follow a normal distribution.

A boxplot of the residuals can be helpful in identifying outliers.

In our case, we found no outliers in the boxplot.

2.b.2 Finding high leverage points:

The leverage is a measurement of how far an observation's values of the predictor variables are from those of the other observations. This means that a point with high leverage is an outlier in the x-direction. It is important to identify high-leverage points because they can potentially distort the results of the regression analysis, leading to misleading interpretations.

Having high leverage is considered a necessary but not sufficient condition for an observation to be influential.

To determine high leverage points, we used the rule that if an observation has a hat value higher than 2p/n, where p is the number of predictors in the model and n is the sample size, it is a high leverage point.

The observations with indexes (20, 29, 45) have high leverage, with respective values (0.1907017, 0.2562940, 0.1687479). The threshold above which an observation was considered to have high leverage was 0.16.

2.b.3 Finding regression outliers:

While the leverage is an indication of how much an observation is an outlier in the x-direction, it is also important to verify regression outliers: points with a very large absolute value of the standardized residual. The commonly used criterion to determine whether a point is a regression outlier is that if $|d_i|>2$, then that point is an outlier.

For our model, we found no outliers using the above criterion, which is consistent with the results obtained from the box plot. We note that the points with maximum absolute value for the standardized residual are observations 1 and 31, with standardized residuals of 1.944671 and 1.902460 respectively.

2.b.4 Finding influential points:

The influential points are those that have a significant effect on the inferences drawn from the data. To measure influence, it is often determined how much the coefficients of the regression would change if a given observation were to be removed from the data. The most common of these methods is the Cook's distance. The higher the Cook's distance of a given observation, the more influence it has on the coefficients, and observations with a Cook's distance bigger than 4/(n-p) are considered influential.

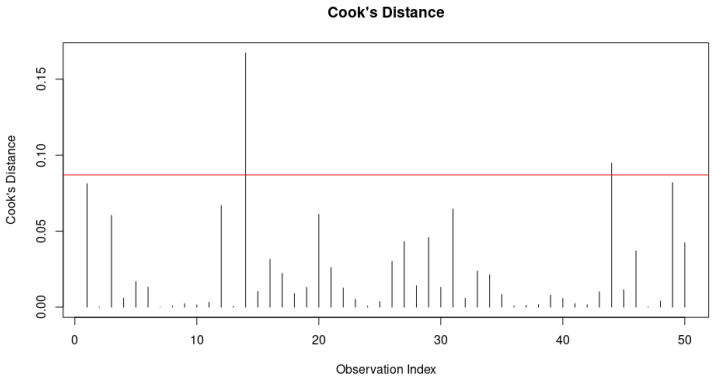

Figure 16: Cook's Distance

Using the above criterion, observations 14 and observation 45 are influential data points, with respective Cook's distances of 0.16725603 and 0.09485169.

(in red is the high Cook's distance threshold for our model)
As was earlier discussed, influential data points must also have high leverage.
Observation 45 is an influential data point, as it fits both the criteria for high leverage and high influence.
We did not identify observation 14 as a high leverage point, nor as an outlier. Upon more careful examination, observation 14 was found to have a leverage of 0.14917058, which is below the high leverage threshold of 0.16 for our model but just barely. We also inspected its standard residual and found that it was -1.859691587, which is also below but very close to the threshold for a regression outlier. This leads us to conclude that data point 14 is also an influential data point, especially given how much bigger its Cook's distance is than that of all the other observations.

2.c Confidence and Prediction Intervals

The observed value for mpg in observation 14 was 14. The 97.5% confidence interval for the expected value of observation 14 is [15.3967, 18.15321]. This means that we are 97.5% confident that the expected value of observations with the same X-value (cylinders, weight and acceleration) as observation 14 lies in that range.
The one for observation 31 is [24.24188, 26.08058], and its observed value was 28.
In both cases, the observed value lies outside the confidence interval. Going back to our previous analysis, both 14 and 31 had a high standardized residual, and 14 was even an outlier, that also had high leverage and was an influential point. Being so, what this means is that our model concludes that it is not likely that the observed values for the X-values in observations 14

and 31 are close to the true expected value. This is natural, because our model is not a very good fit for these observations, because as we said previously they have high standardized residuals. For observation 14, the 97.5% prediction interval is [12.94952, 20.60038]. This means that we are 97.5% confident that a future observation with the same X-value as observation 14 will lie within this range. For observation 31, it is[21.47619, 28.84627].
Now, we can see that both observed values lie within the respective prediction intervals. This suggests that the model is capturing the variability of individual data points effectively, since even though these observations have high standardized residuals and thus the model considers them to be far from the expected value, they are still considered to be within the interval for future observations. If this were not the case, it would be a problem for our model, because it is not inconceivable that having been observed once, these values would be observed again in the future.
As was expected, the prediction intervals are wider than the confidence intervals. This is because the reason for the uncertainty in confidence intervals is the sampling variability, but for prediction intervals we take into account sampling variability and variability within individual data points.
The range of the intervals for observation 31 is smaller, which indicates that for that X-value the model is more certain about the results. This is quite natural given that observation 14 has high leverage while observation 31 does not. Since the predictor variables for observation 14 are unusual, the model does not have many similar counterparts in the data set, and thus the model is less certain about its results.

Bibliography

Checking assumptions of the linear model