TÉCNICO LISBOA

Report Project 2

Data Analysis and Statistical Modeling

Prof Isabel Rodrigues

Grupo 1

João Matos nº98949

Ana Pinto nº102949

Marina Nóbrega nº103880

Manuel Mota Dias nº96056

Maria Freitas nº96757

Introduction

For this project a datagram is given for us to analyse:

Data frame: Auto
Subset: observation 1 to 50
Variables: all except name:

- mpg – miles per gallon
- cylinders – number of cylinders between 4 and 8
- displacement – engine displacement (cu. Inches)
- horsepower – engine horsepower
- weight – vehicle weight (lbs.)
- acceleration – time to accelerate from 0 to 60 mph (sec.)
- year – model year (modulo 100)
- origin – (origin of the car (1. American, 2. European, 3. Japanese)

1

Summary statistics

```
> # Summary statistics
> summary(auto_subset)
      mpg          cylinders      displacement      horsepower         weight      acceleration        year           origin    
 Min.   : 9.00   Min.   :4.00   Min.   : 97.0   Min.   : 46.00   Min.   :1835   Min.   : 8.00   Min.   :70.00   Min.   :1.00  
 1st Qu.:14.00   1st Qu.:4.50   1st Qu.:154.5   1st Qu.: 91.25   1st Qu.:2599   1st Qu.:11.50   1st Qu.:70.00   1st Qu.:1.00  
 Median :17.50   Median :7.00   Median :280.0   Median :121.50   Median :3381   Median :13.75   Median :70.00   Median :1.00  
 Mean   :18.08   Mean   :6.48   Mean   :268.8   Mean   :135.34   Mean   :3366   Mean   :13.40   Mean   :70.42   Mean   :1.28  
 3rd Qu.:22.00   3rd Qu.:8.00   3rd Qu.:357.8   3rd Qu.:173.75   3rd Qu.:4195   3rd Qu.:15.38   3rd Qu.:71.00   3rd Qu.:1.00  
 Max.   :28.00   Max.   :8.00   Max.   :455.0   Max.   :225.00   Max.   :5140   Max.   :20.50   Max.   :71.00   Max.   :3.00  
```

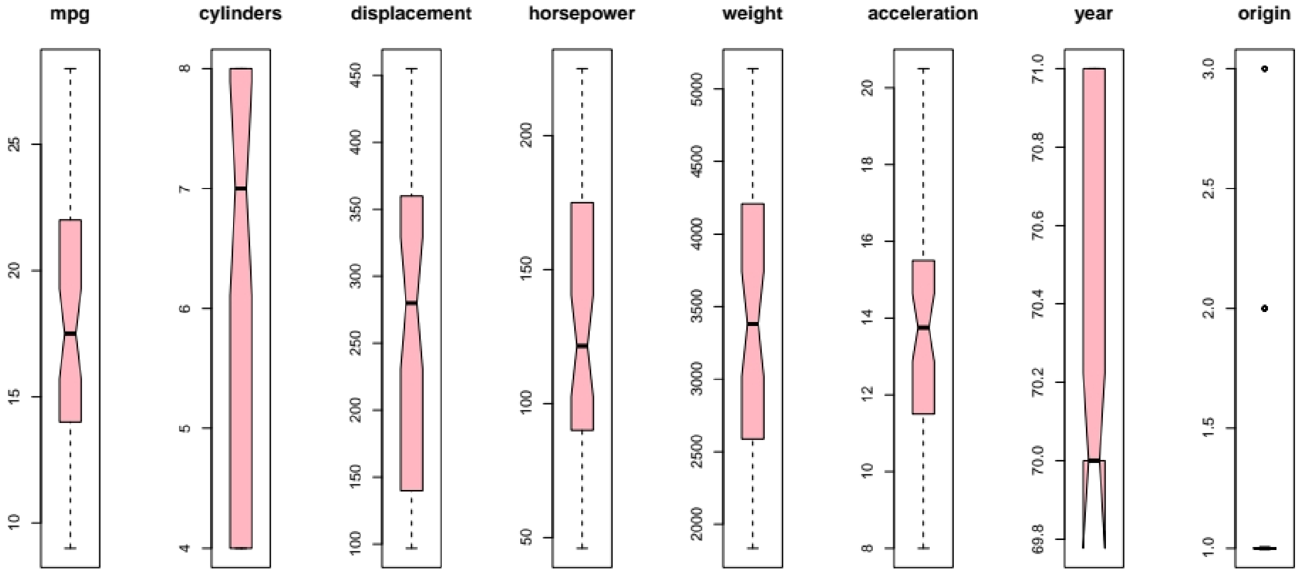
```
             vars  n          sd   trimmed        mad  min    max  range        skew   kurtosis           se Winsorized_Mean    Variance
mpg             1 50   5.2092539   17.9000    5.18910    9   28.0   19.0  0.37391779 -0.9913305   0.73669976           18.06     27.1363
cylinders       2 50   1.6932037    6.6000    1.48260    4    8.0    4.0 -0.46070716 -1.4771470   0.23945516            6.48      2.8669
displacement    3 50 115.7538839  267.5750  137.14050   97  455.0  358.0 -0.07343823 -1.3197659  16.37007125          263.80  13398.9616
horsepower      4 50  49.1945783  132.2750   49.66710   46  225.0  179.0  0.34493617 -1.2440568   6.95716398          131.16   2420.1065
weight          5 50 899.0059865 3342.4250 1202.38860 1835 5140.0 3305.0  0.13682281 -1.2067456 127.13864587         3354.26 808211.7637
acceleration    6 50   2.8193935   13.3375    2.59455    8   20.5   12.5  0.12596323 -0.3767769   0.39872245           13.40      7.9490
year            7 50   0.4985694   70.4000    0.00000   70   71.0    1.0  0.31449986 -1.9386713   0.07050836            6.48      0.2486
origin          8 50   0.6074369    1.1250    0.00000    1    3.0    2.0  1.93771682  2.4157790   0.08590455            1.00      0.3690
```

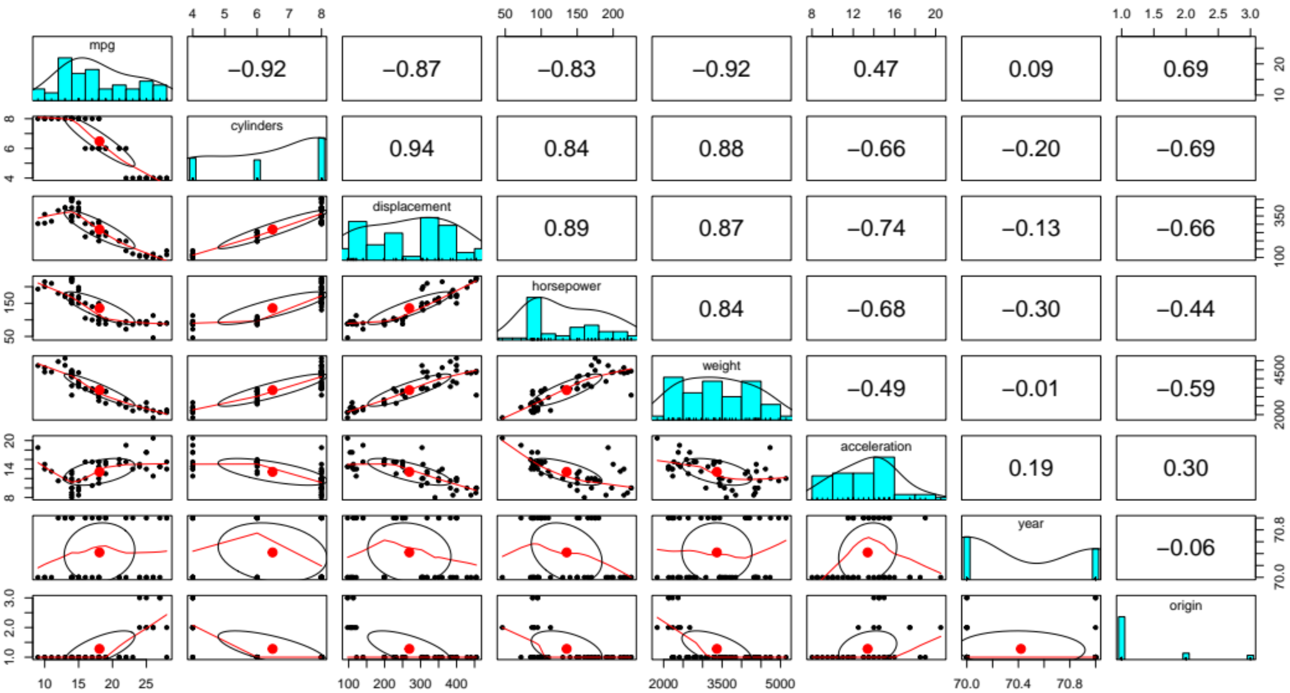
```
> #Covariance#
> auto_cov = round(cov(auto_subset), digits = 4)
> auto_cov
                   mpg cylinders displacement horsepower      weight acceleration    year    origin
mpg            27.1363   -8.1208    -527.3273  -211.8033  -4312.4865       6.8857  0.2310    2.1812
cylinders      -8.1208    2.8669     184.3216    69.9967   1341.4890      -3.1347 -0.1649   -0.7086
displacement -527.3273  184.3216   13398.9616  5050.3078  90877.8065    -241.3816 -7.4890  -46.2784
horsepower   -211.8033   69.9967    5050.3078  2420.1065  37211.8200     -94.7204 -7.3906  -13.0971
weight      -4312.4865 1341.4890   90877.8065 37211.8200 808211.7637   -1253.5347 -3.7482 -321.2539
acceleration    6.8857   -3.1347    -241.3816   -94.7204  -1253.5347       7.9490  0.2673    0.5082
year            0.2310   -0.1649      -7.4890    -7.3906     -3.7482       0.2673  0.2486   -0.0180
origin          2.1812   -0.7086     -46.2784   -13.0971   -321.2539       0.5082 -0.0180    0.3690
```

Total Variance

```
> auto_vartot
[1] 194011
```

Summary Plots





2.A)

To find the best subset of regressors, we applied the regression model until we got the ones that we considered useful.

```
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
```

```
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
```

```
Coefficients:
              Estimate  Std. Error  t value Pr(>|t|)
(Intercept)  41.7214469  3.0398687  13.725  < 2e-16 ***
cylinders    -1.5852955  0.3749030  -4.229 0.000117 ***
horsepower   -0.0229062  0.0099822  -2.295 0.026579 *
weight       -0.0018702  0.0005889  -3.176 0.002729 **
acceleration -0.3882040  0.1157813  -3.353 0.001652 **
origin        0.9604299  0.5183869   1.853 0.070636 .
```

```
Coefficients:
              Estimate  Std. Error  t value Pr(>|t|)
(Intercept)  45.3408068  2.3911834  18.962  < 2e-16 ***
cylinders    -1.9702777  0.3203725  -6.150 1.87e-07 ***
horsepower   -0.0179892  0.0098797  -1.821 0.075285 .
weight       -0.0018947  0.0006044  -3.135 0.003026 **
acceleration -0.4238977  0.1172125  -3.616 0.000752 ***
```

```
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 **
```

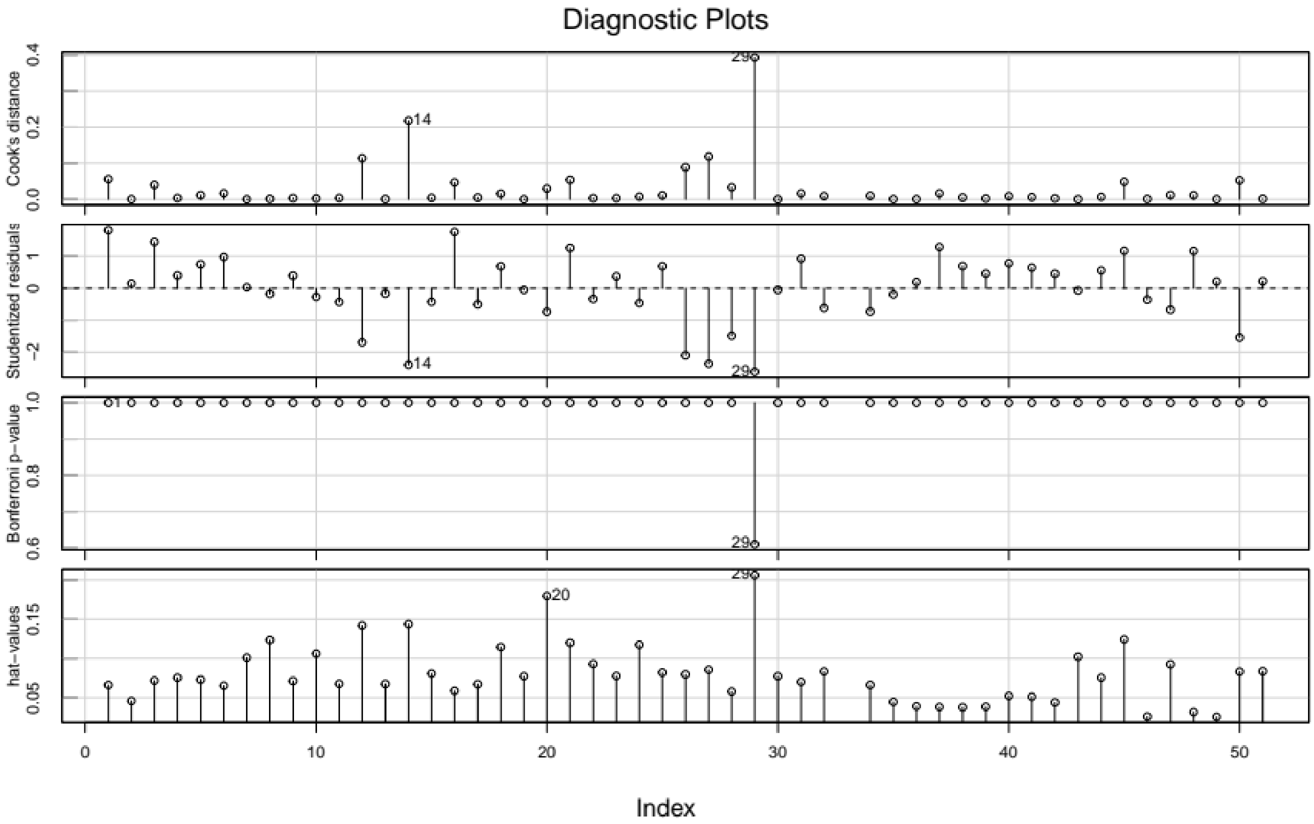
In the end, cylinders, weight, and acceleration are the selected ones. We chose these because it would allow us to work with less predictors. The adjusted **r^2** value of the last two iterations is similar and lower in the last iteration.
After fitting a regression model to explain the mpg variable using the predictors we just selected we get the values:

r^2 = 0.891092607413735
r^2adj = 0.8839899513755

Given that our **r^2** and **r^2**adj values are relatively high (>0.8), it suggests that the current model explains a significant portion of the variability in the response variable.

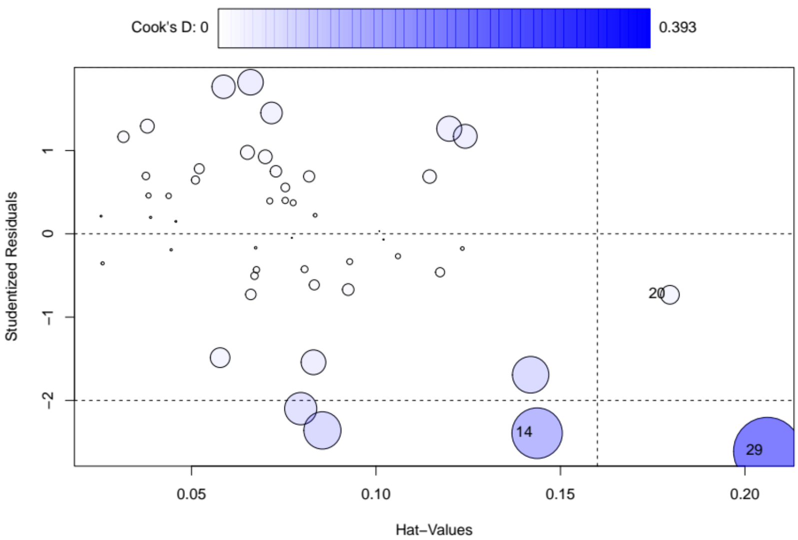
<u>2.B)</u>
For this regression we are using p = 3 predictors for n = 50 observations
Searching for possible influential/leverage observations we get this:

influenceIndexPlot(reg_mpg)



influencePlot(reg_mpg)

	StudRes	Hat	CookD
14	-2.3937542	0.1436755	0.2179395
20	-0.7317671	0.1796745	0.0296206
29	-2.6115400	0.2060838	0.3928817



For mpg, the possible leverage observations are 29, 20 and 14. The two observations with highest cook's distance are 14 and 29, so the more possible influential observations are 29 and 14.

<u>2.C)</u>

Calculating the 97.5% Confidence Interval and Prediction Interval for the expected values of the responses for observations **14** and **31** we get:

```
                         Obs_14                                         Obs_31
CI ] 2.77279693700092  ,  2.92437757683037 [ ] 3.19053386221619  ,  3.29633578732302 [
PI ] 2.63475473062577  ,  3.06241978320552 [ ] 3.03660472892808  ,  3.45026492061112 [
```

The prediction interval (PI) for Obs_14 is noticeably wider than the confidence interval (CI). This wider width in the prediction interval reflects the additional uncertainty associated with predicting individual observations, considering both the uncertainty in estimating the mean and the variability of individual observations.

Similar to Obs_14, the prediction interval (PI) for Obs_31 is wider than the confidence interval (CI). This wider width suggests a higher level of uncertainty when predicting individual observations, considering both the variability in estimating the mean and the variability of individual data points.

In both cases, the prediction intervals are wider than the corresponding confidence intervals. This is a common characteristic, as prediction intervals need to account for the variability in individual observations, making them more conservative and wider. The confidence intervals, on the other hand, primarily focus on the uncertainty around estimating the mean.

In summary, the widths of the prediction intervals highlight the increased uncertainty when making predictions for individual observations compared to estimating the mean.

Conclusion

To finish, this analysis provided a comprehensive understanding of the relationships within the Auto dataset and helped us learn more about regression models and their uses.