Elements of Machine Learning

Assigment 1 - Problem 6

Sangeet Sagar(7009050), Philipp Schuhmacher(7010127)
{sasa00001,phsc00003}@stud.uni-saarland.de

November 11, 2021

Part 1

It's inaccurate to say that relationship between all variables are linear. A variety of relationship can be visualized in the scatter plot among different variables.

1. No relationship: Some variables hold no relationship. E.g.`mpg-origin`, `cylinders-origin`, `displacement-origin`. Or rather increase or decrease of one variable is independent of other.
2. Linear Relationship: Some variables are linearly correlated like `displacement, horsepower, and weight`. These variables are **strongly correlated**.
3. Non-linear relationship: A few variables are non-linearly connected like `mpg` is non-linearly related with `weight`, `horsepower` and `displacement`. We can also call these **negatively correlated**.
4. Scattered points: These are just scattered points and the connection between the variables could not be determined. E.g. `acceleration-mpg`. These variables are **weakly correlated**. `acceleration-weight`.

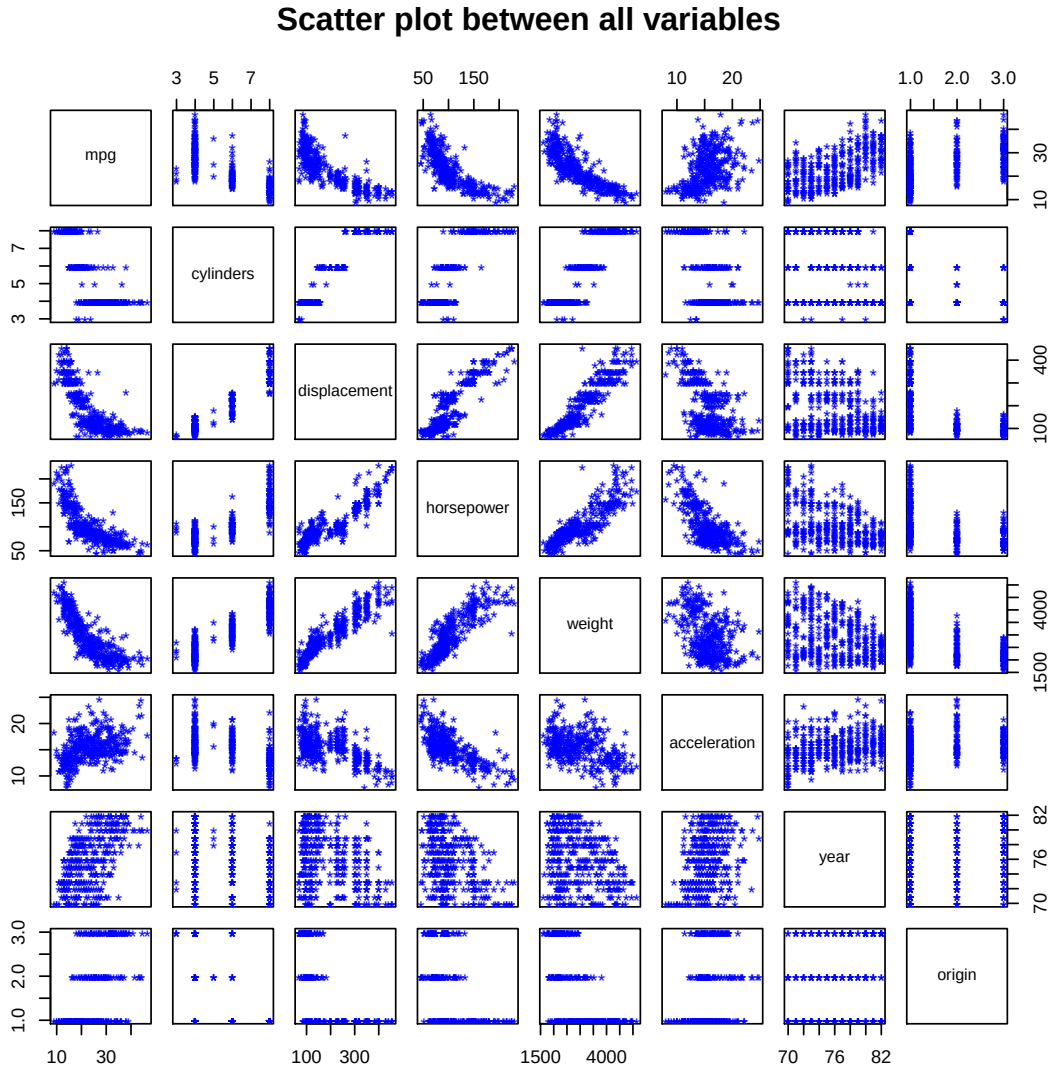

Figure 1: Scatterplot between all variables namely `mpg`, `cylinders`, `displacement`, `horsepower`, `weight`, `acceleration`, `year`, `origin`

Part 2

Two variables that appear to be the most highly

1. **Correlated**: `cylinder-displacement`, `weight-displacement`. However a few other pairs like `horsepower-displacement` and `weight-cylinder` also appear to be equally highly co-related. Its difficult to differentiate visually. By looking at graph, the data points in all of these pairs strictly follow a positive linear behaviour.

2. **Anti-correlated**: `mpg-displacement` and `mpg-weight` appear to be most highly anti-correlated variables because of their inverse relationship. `mpg-cylinder`, `mpg-horsepower` also appear to be anti-correlated on a similar extent, but the negative non-linearity was more obvious in the former.

Using the `cor()`, we find that

1. `cylinder-displacement`, `weight-displacement` are the most highly correlated variables with scores of 0.950 and 0.93 respectively.

2. `mpg-displacement` and `mpg-weight` are the most anit-correlated variables with with scores of -0.80 and -0.832 respectively.

Thus, the results from `cor()` are almost same as what can be visualized with slight variations. However, a close inspection of the scatter plot makes it clear. The goal here is to observe for variable pairs that show strong positive linear behaviour (highly correlated) and those that show strong linear/non-linear behaviour (highly anti-correlated).

Part 3

In order to determine a statistically significant predictors, we need to look for variables with high absolute t-value as well as small p-value. The small p-value indicates a stronger evidence that we should reject the null hypothesis and that there is an relation that holds between predictor and response.

	Estimate	Std. Error	t value	Pr(>|t|)
cylinders	−3.558078	0.1456755	−24.42468	1.311384e−80
displacement	−0.06005143	0.002240043	−26.80815	1.660642e−90
horsepower	−0.1578447	0.006445501	−24.48914	7.031989e−81
year	1.230035	0.08735748	14.08048	1.075794e−36

Which predictors appear to have a statistically significant relationship to the outcomes?
Each predictor appears to hold statistically significant relationship with the response for the reason that p-value for each predictor is less that 5%.

	R2 score
cylinders	0.604689
displacement	0.6482294
horsepower	0.6059483
year	0.3370278

How good are the resulting models?
All of the resulting model fit the data except one. The extent of fit is similar among predictors like `cylinders, displacement, horsepower` fitting almost 60% of the data. However `year` gives a fit with only 33% of the data, hence the resulting model with this predictor is not good.

Part 4

How is the model fit? (using R^2)?
The R^2 statistic is a measure of the linear relationship between the response and the predictor. Using multiple linear regression, the R^2 value is observed to be 0.809, which suggests that the 80% of the data fits the regression model. This score is, in fact, higher than each model in part 3. Therefore, the multiple linear regression fits the given data well than a linear regression using a single predictor.

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	−1.453525e+01	4.7638818940	−3.05113578	2.437741e−03
cylinders	−3.298591e−01	0.3321041332	−0.99323994	3.212169e−01
displacement	7.678430e−03	0.0073577361	1.04358597	2.973318e−01
horsepower	−3.913556e−04	0.0138365215	−0.02828425	9.774501e−01
weight	−6.794618e−03	0.0006700227	−10.14087727	1.416190e−21
acceleration	8.527325e−02	0.1020355670	0.83572081	4.038303e−01
year	7.533672e−01	0.0526181480	14.31763010	1.410428e−37

R–sq value:[1] 0.8092553

What can you observe in the different models concerning the significance of the relationship between response and individual predictors?
Predictors like `cylinders, displacement, horsepower, acceleration` have p-value more than 5%; hence they are non-significant, and there does not exist any substantial relationship between the response and these individual predictors. However, other predictors like `weight, year` have p-value less than 5%, thus, supporting the evidence that the null hypothesis be rejected since there is some significant association between the response and these individual predictors.

What does the sign of the coefficient (i.e. of the estimate) tell you about the relationship between the predictor and the response?
The sign of the coefficient, i.e. t-value, tells us about the direction of the effect the predictor is having on the response `mpg`. If the sign is positive, it means that the response will increase with the increase in that individual predictor (keeping other predictors as constant) and vice-versa.

Part 5

Identify the residual plot.
First plot i.e Residual vs Fitted is the residual plot

Does the residual plot suggest any non-linearity in the data (provide an explanation)? Does the residual plot suggest an unusually large outliers?
Looking the residual plot, it can be concluded that there is no specific pattern, While a curve in the pattern can been but the residuals are scattered. Hence, it indicates of the non-linearity in the data.
Outliers can be inferred form the Residual vs Leverage plot. The red dashed line at the top-right suggests that there exists an unusually large outlier but could not be detected in the plot. Hence their inclusion or exclusion will affect the model.

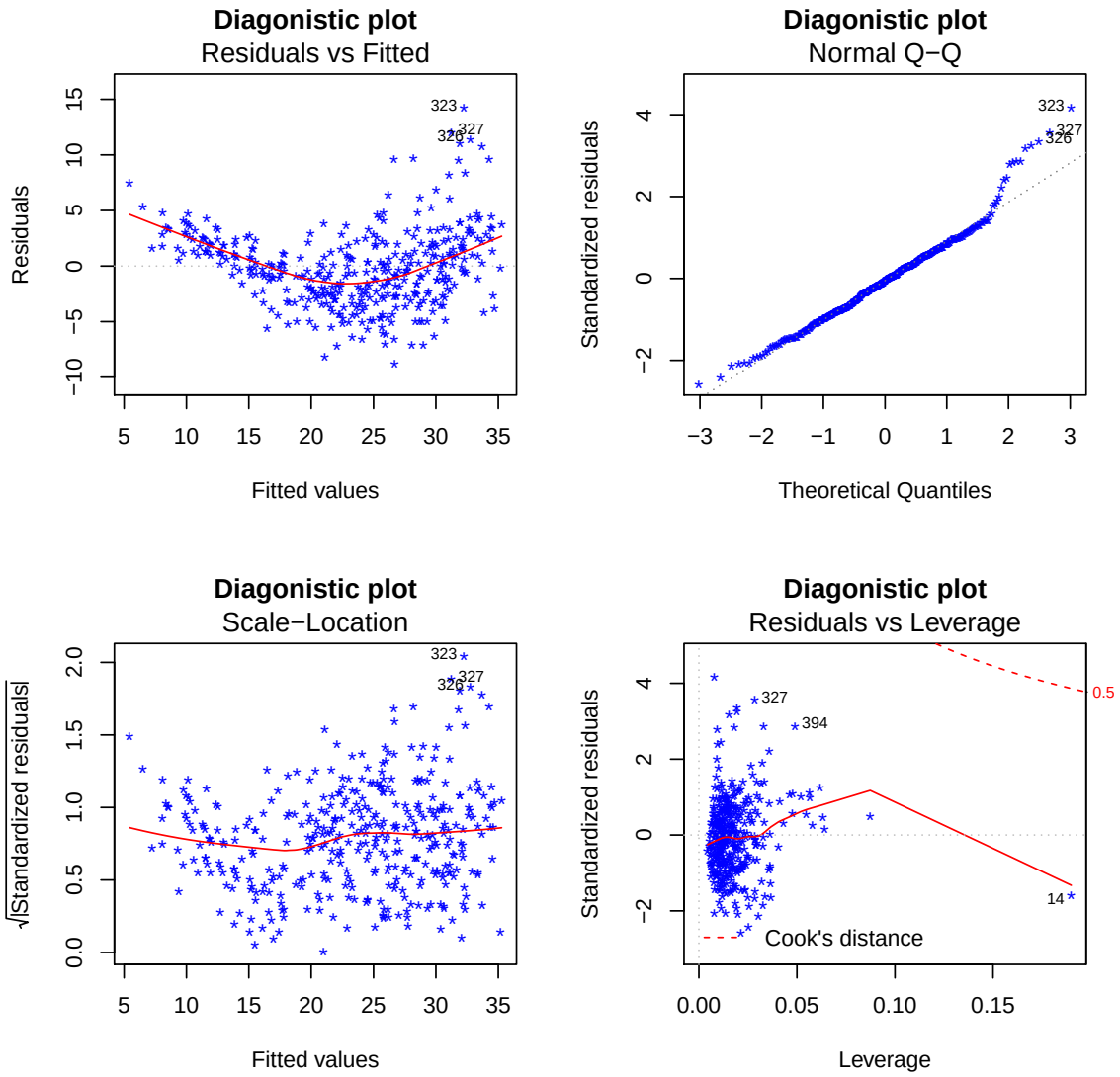

Figure 2: Diagonistic plots