University of St Andrews | FOUNDED 1413

Essay cover sheet to be completed by student and attached to the front of essay.

STUDENT ID	:	**180009169**
STUDENT SUBJECT	:	**MSc Applied Statistics and Datamining**
COURSE	:	**ID5059 Knowledge Discovery and Datamining**
ESSAY/PROJECT TITLE	:	**Practical1**
NAME OF TUTOR(S)	:	**Tom Kelsey & Carl Donovan**
SUBMISSION DEADLINE	:	**2019/03/11**
DATE SUBMITTED	:	**2019/03/11**

DECLARATION

- I confirm that I have read and understood the University's policy on plagiarism.
- I confirm that this assignment is all my own work.
- I confirm that in preparing this piece of work I have not copied any other person's work, or any other piece of my own work.
- I confirm that this piece of work has not previously been submitted for assessment on another course.

180009169 (Student ID in place of signature)

2019-03-11 (date of submission)

1. Introduction

The study conducted in this report is based on automobile data, especially miles per gallon data. Miles Per Gallon (mpg) is an important indicator for cars, which measures the fuel efficiency. The data set used in this report is from UCI Machine Learning Repository (2019), which is a slightly modified version of the dataset published by Quinlan in 1993. The response variable used in this report is mpg, while explanatory variables are displacement, horsepower, weight and acceleration. The analysis software used in this study is RStudio and some functions are provided by "P01-code-stats-functions" file on Moodle (2019).

The first step of this report is to build three types of models (polynomial linear models, B-splines and bin smooth models). Then, this report will interpret the relationships between the response and explanatory variables and find the most effective attribute. Finally, we will compare the accuracy of these models.

2. Exploratory Data Analysis

To make us familiar with the data, some preliminary data analysis work should be done before fitting models. There are six missing values in the column of horsepower. Because the number of missing values is quite small compared to the size of dataset, the rows with missing values are simply deleted.

Table1 shows the correlation coefficients between mpg and explanatory variables. According to it, only acceleration has positive relationship with mpg, while others have negative relationships with mpg. It means when acceleration increases, mpg is likely to increase. However, when other variables increase, mpg is more likely to decrease. Furthermore, the correlation for acceleration is much smaller than others, which is an indication that acceleration might has weaker relationship with mpg. Figure1 can also prove these signs. Only the slope for acceleration is negative, the pattern between acceleration and mpg is more random as well.

Table1. Correlations between mpg and other variables

	displacement	horsepower	weight	acceleration
mpg	-0.8051269	-0.7784268	-0.8322442	0.4233285

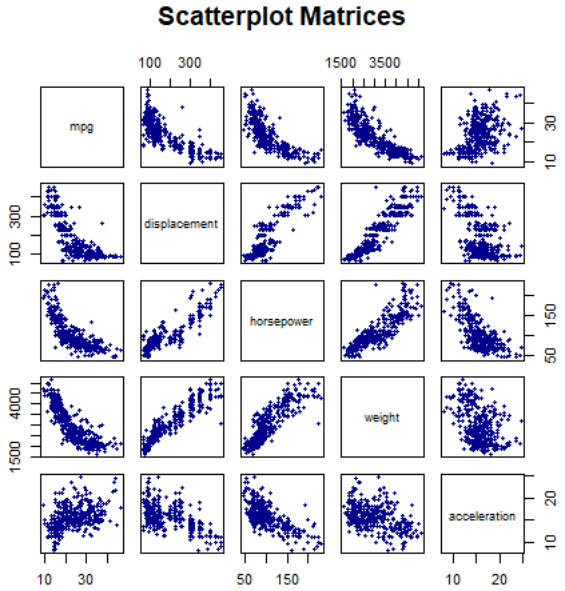


Figure1. Scatterplot Matrices

3. Model Fitting and Selection

To select the optimal model among models with different parameters, we use Akaike's Information Criterion (AIC) instead of Cross Validation in this study since AIC is easier to apply in practice. AIC is an estimator to measure the quality of models. The model with lowest AIC is most likely to be the optimal model. AIC has a penalty term to prevent overfitting. Therefore, it is a trade-off between the goodness-of-fit and complexity of models. However, AIC cannot tell the absolute quality of a model, but only the model quality compared to other candidate models.

3a) Polynomial Linear Models

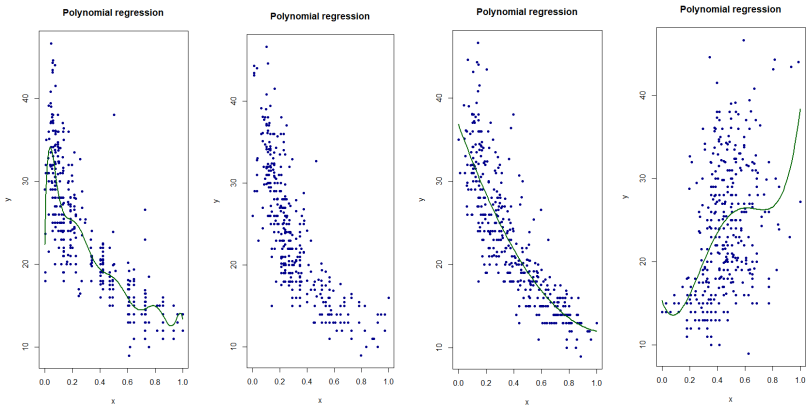
The polynomial regression is used to model the relationship between the response variable and multiple degree of explanatory variable. It is given the formula: $y = \beta_0 + \beta_1 x_1 + \beta_2 x^2 + \cdots\cdots + \beta_n x^n + \varepsilon$. Although the polynomial linear model fits a nonlinear relationship, as a statistical estimation problem it is linear. (Wikipedia, 2019)

We construct polynomial linear models with one degree to thirty degrees for each explanatory variable. Following the principle that selecting the model with minimal AIC, ten, twenty-five, two and four degrees are chosen for displacement, horsepower, weight and acceleration respectively. More detailed results such as RSS and AIC are showed in Table2. According to the Residual Sum of Squares (RSS), the fit quality of models for displacement, horsepower and weight are similar and much better than the model for acceleration. The graphs in Figure2 are also illustrate that points of acceleration are much further to the line generally.

Table2. Polynomial Regression Results

Variable	Degree	AIC	RSS	Adj.R2
Displacement	10	2243.889	6610.19	0.7224824
Horsepower	25	2255.505	6536.669	0.7068238
Weight	2	2238.115	6784.899	0.7151476
Acceleration	4	2640.19	18731.31	0.2135979

Figure2. Polynomial Regression Plots



Displacement Horsepower Weight Acceleration

3b) B-splines

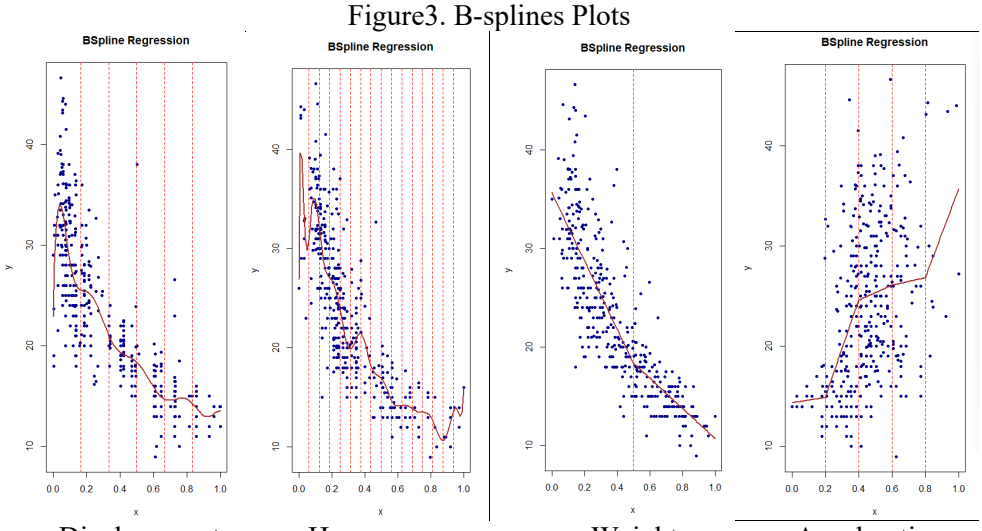
B-splines are the more flexible and local function than polynomial regression. The reason is that B-splines not only have degree parameter but also the knots. Therefore, the B-spline can increase its flexibility by placing more knots while the degree is fixed. We should place more knots in places where more fluctuate, and less knots in places more stable. However, in practice, it is common to place knots at uniform quantiles of the data. (James et al., 2013) Therefore, this report will place knots uniformly.

In this case, the upper limit for degree is set as five and the upper limit for knots is thirty. We chose the optimal models with minimized AIC and get the results in Table3. The results of B-splines are similar with results of polynomial linear models. The first three variables perform much better than the last one. The plots in Figure3 illustrate that the B-splines curves might be more wiggly than polynomial curves, which likely to be the result of adding knots.

Table3. B-splines Results

Variable	Degree	Knots	AIC	RSS	Adj.R2
Displacement	5	5	2243.412	6602.15	0.7155449
Horsepower	5	15	2258.051	6512.472	0.7118456
Weight	1	1	2239.446	6807.962	0.7127098
Acceleration	1	4	2638.806	18570.27	0.2102597

Figure3. B-splines Plots



3c) Bin Smooth Models

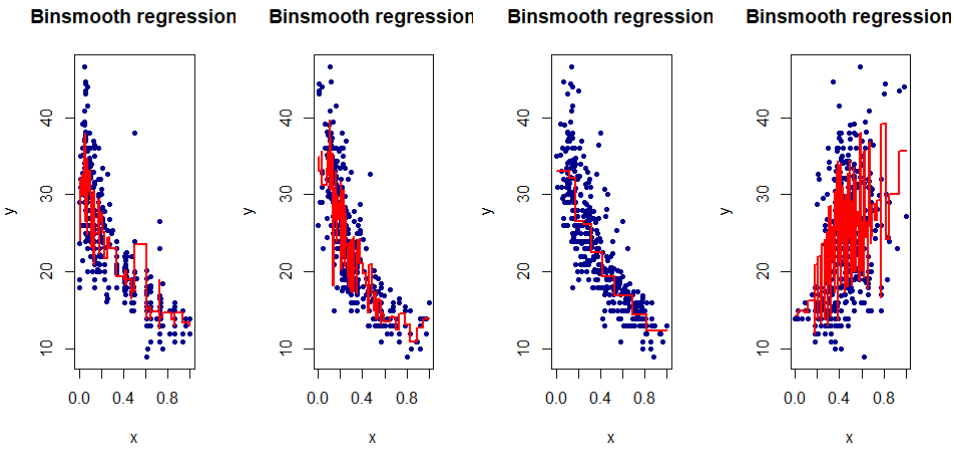
The main idea behind Bin smooth models is that dividing the x-axis into numbers of regions with same number of data points, then calculating the mean of them as the response values for the whole region. The number of data point is called the bin length. Therefore, bin smooths give us discrete values at boundaries.

The upper limit for bin length is set to be 100 in this case. After comparing the AIC, we get results as Table4. According to RSS, the goodness-of-fit for all variables, especially acceleration, when we use bin smooths. In addition, the gap between acceleration and other explanatory variables is narrowed. However, Figure4 shows that the bin size of displacement, horsepower and acceleration might be too small to cause overfitting. Selecting bin size is a trade-off between bias and variance. When the number of bins increases, there are less data in one bin. The mean tends to be more relative to each data. Therefore, the RSS is likely to decrease, and vice versa.

Table4. Bin Smooth Models Results

Variable	Bin Length	AIC	RSS	Adj.R2
Displacement	5	2192.943	4102.96	0.7848178
Horsepower	4	2254.172	4353.435	0.7569262
Weight	47	2233.07	6496.199	0.7215714
Acceleration	3	2609.204	9100.16	0.6179452

Figure4. Bin Smooth Plots



Displacement Horsepower Weight Acceleration

4. Model Comparison

Comparing these three types of models, the performance of polynomial linear models and B-splines are similar, while the outputs of bin smooths are better. However, there might be overfitting problems in bin smooth models. Moreover, although the B-spline have similar goodness-of-fit with the polynomial linear model, it reduced the degree of model by adding numbers of knots.

According to Adjusted R Squares (Adj.R2), the displacement is most likely to have the best predictive ability. Adj.R2 is a measurement describes how much the explanatory variables can explain the variability of response data. The displacement has the highest adj.R2 in all three types of models, which are over 0.7. It means over 70% variability of mpg data can be explained by displacement.

5. Conclusion

To sum up, there are four main conclusions can be made in this report. First, the automobile data can be fitted with polynomial linear models, B-splines and bin smooths. In addition, the degree and knot can be chosen according to the minimized AIC. Second, compared with the polynomial linear model, B-splines can reduce the degree by placing knots. Third, the RSS of the bin smooth model could decrease when the number of bins increase, but we should be wary about overfitting problem. Finally, the displacement is likely to be the best predictor in this case, which can explain over 70% variability of mpg data.

Reference:

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

ID5059 Teaching Team (2019). Practical specifications & resource**.** Available at: https://moody.st-andrews.ac.uk/moodle/mod/folder/view.php?id=489957 (Accessed: 10 March 2019).

James, G., Witten, D., Hastie, T. and Tibshirani, R., (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

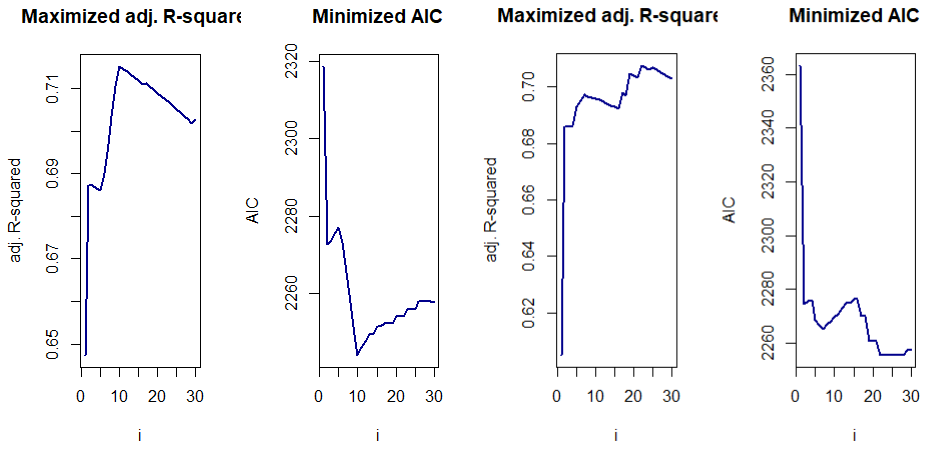
Quinlan,R. (1993). Combining Instance-Based and Model-Based Learning. In Proceedings on the Tenth International Conference of Machine Learning, 236-243, University of Massachusetts, Amherst. Morgan Kaufmann.

RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA, URL http://www.rstudio.com/.

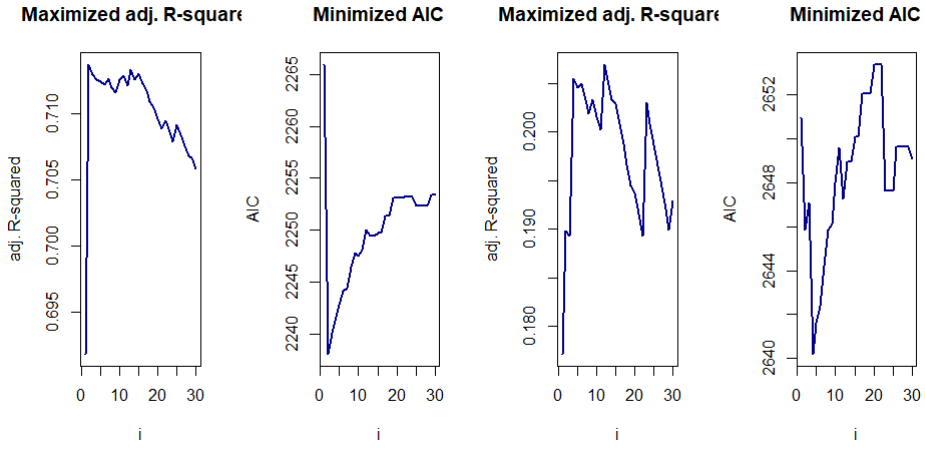
Wikipedia (2019). Polynomial Regression. Available at: https://en.wikipedia.org/wiki/Polynomial_regression#Definition_and_example (Accessed :10 March 2019).

Appendix:

AppendixA. The Number of Degrees against Adj.R2 and AIC for Polynomial Models

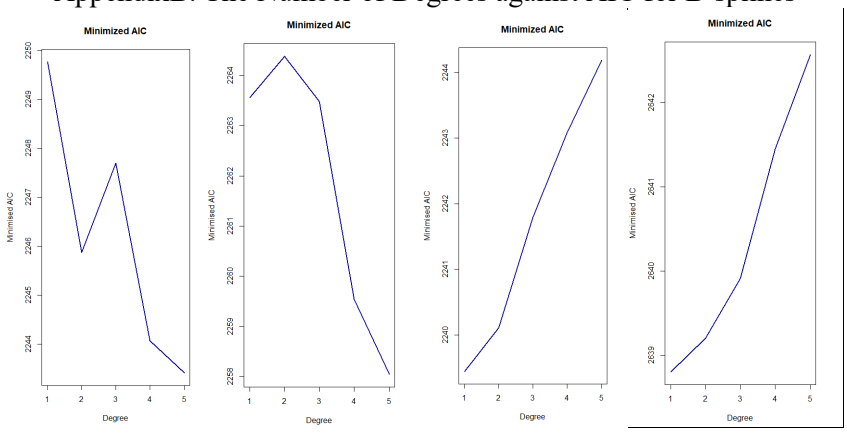


Displacement Horsepower



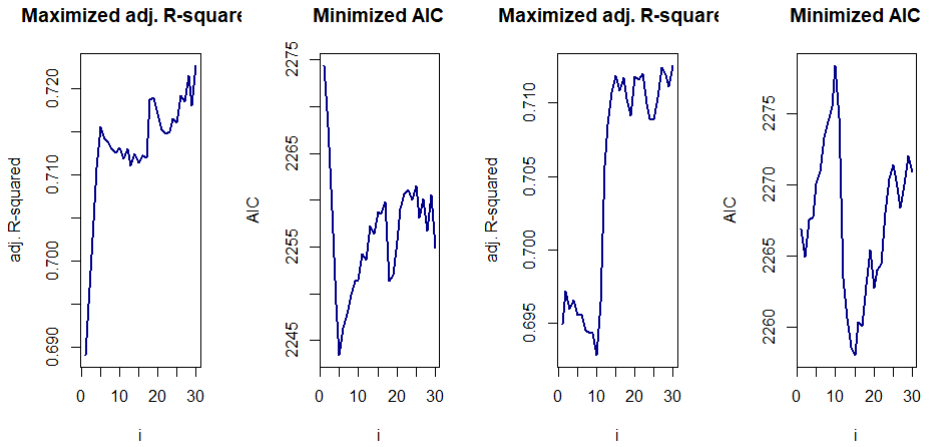
Weight Acceleration

AppendixB. The Number of Degrees against AIC for B-splines

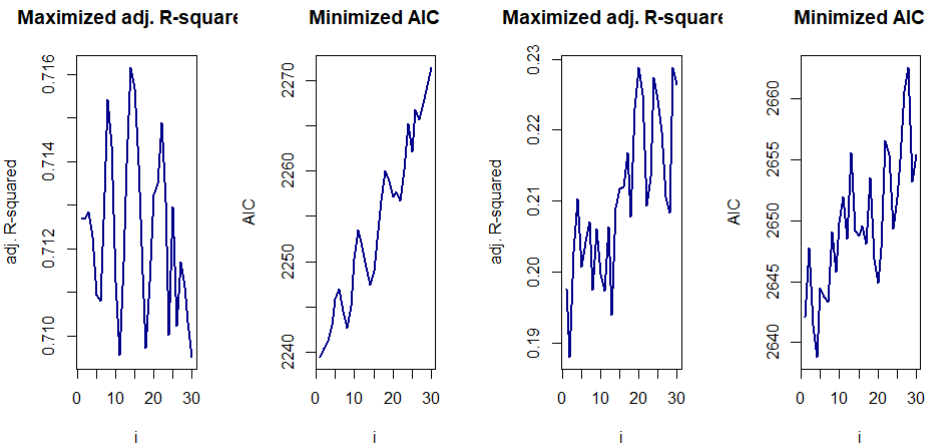


Displacement Horsepower Weight Acceleration

AppendixC. The Number of Knots against Adj.R2 and AIC for B-splines with Fixed Degree

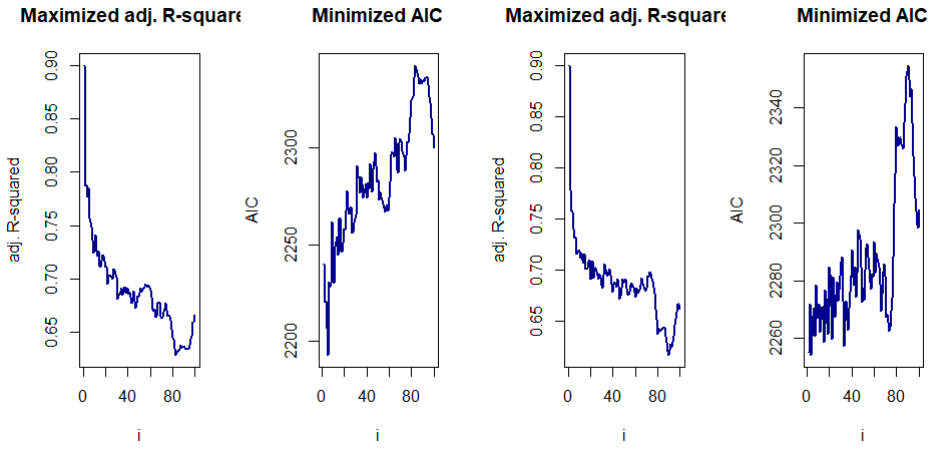


Displacement (degree = 5) Horsepower (degree = 5)

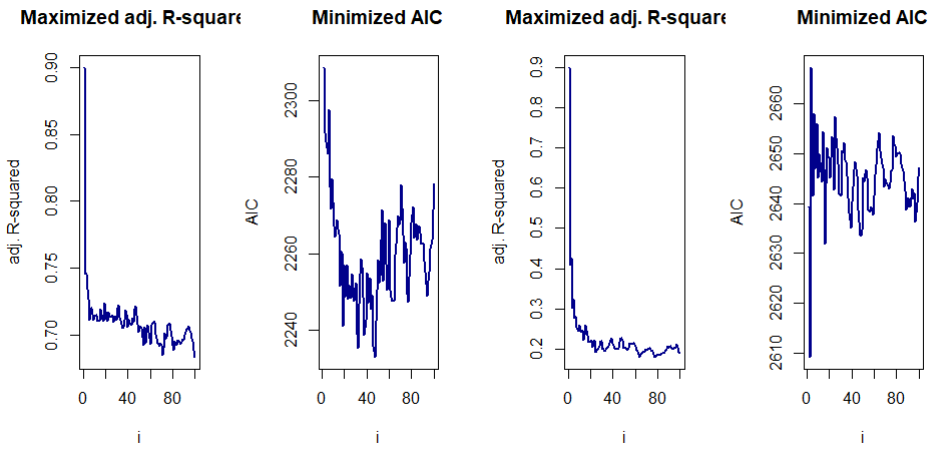


Weight (degree = 1) Acceleration (degree = 1)

AppendixD. The Size of Bins against Adj.R2 and AIC for Bin Smooth Models



Displacement Horsepower



Weight Acceleration