ID5059 - P1 - Auto MPG Analysis

MSc Data-Intensive Analysis

180024570

04/03/2019

Contents

I confirm that the following report and associated code is my own work, except where clearly indicated.

1 Introduction

This practical analyse the pairwise relationships of MPG against Displacement, Horsepower, Weight and Acceleration. Linear models (section 2), bin smooth models (section 3) and B-spline models (section 5) are implemented based on each of the four predictors. The fit of these models is evaluated by RSS, AIC and a measure of generalisation error, Mean Squred Error of prediction of test data.

> Orange line is the minimum. 6 observations without hoursepower values are omitted in this analysis.

2 Linear Models

In this section, linear models are implemented using Displacement, Horsepower, Weight and Acceleration to predict MPG. The candidates models for each predictor are based on polynomial terms with degree from 1 to 15. RSS, AIC and MSE of 10-fold cross validation are used as measurement of model fitting.

2.1 Linear Models for Displacement

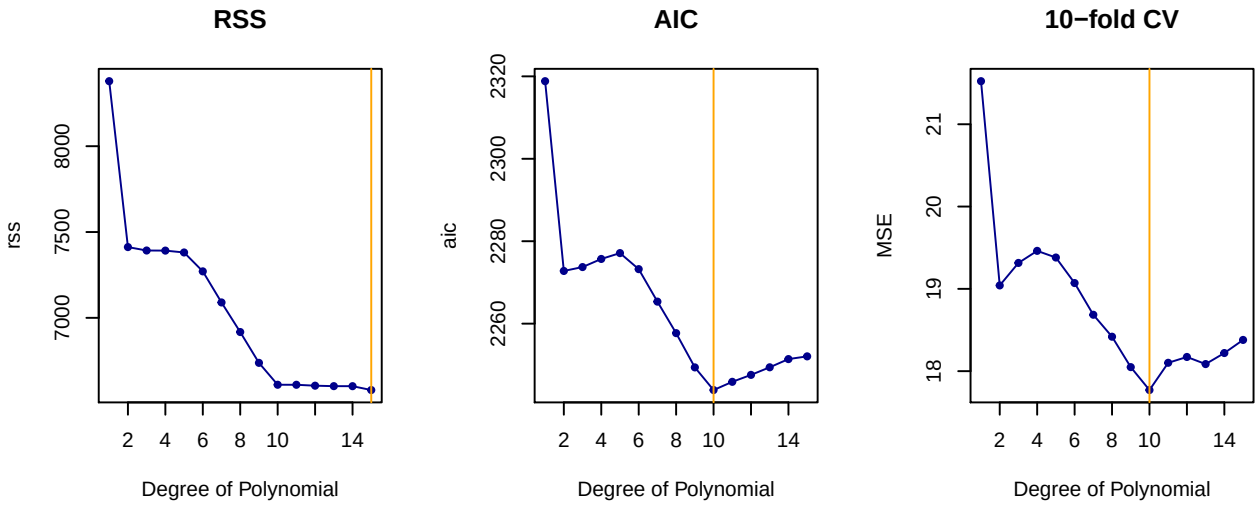

Figure 1: Candidate Linear Models for Displacement

Figure 1 shows the evaluation of 15 candidate models in terms of three measurement. *RSS*, *AIC* and *MSE* of 10-fold cross validation appear to have similar trends when the degree lower than 10. *RSS* continues to gently decline after that, but *AIC* and *MSE* start rising because of overfitting.

Therefore, the one with degree 10 is selected and shown in figure 2.

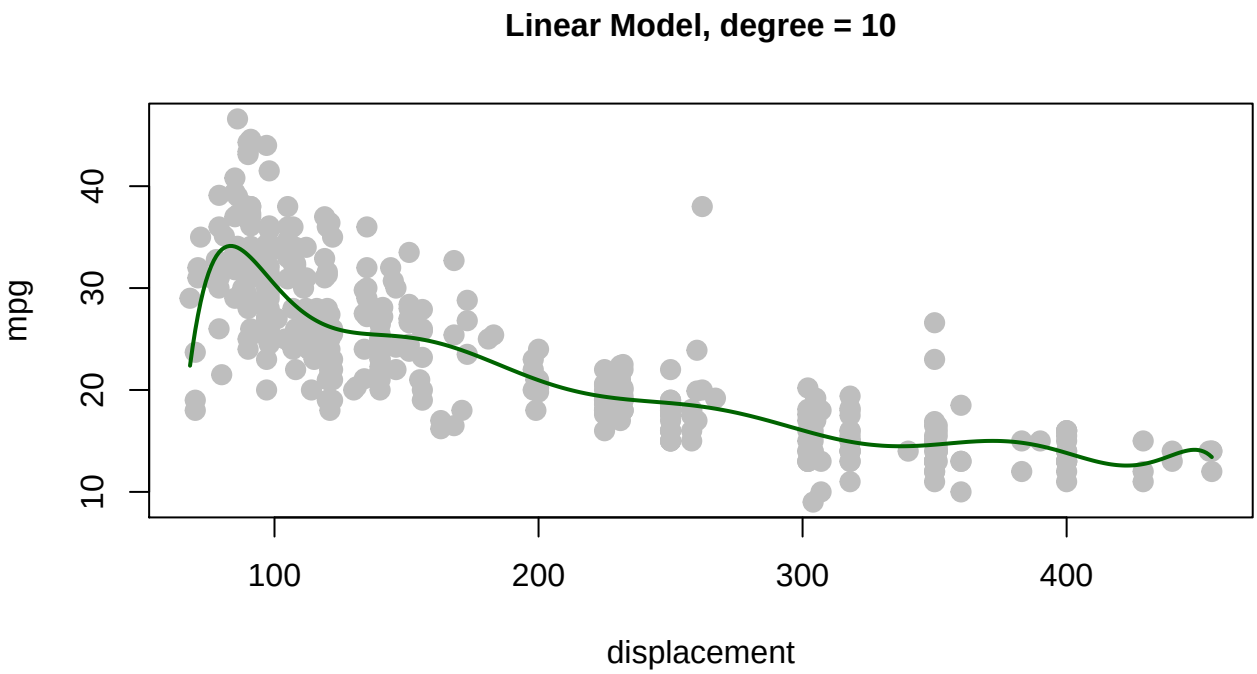

Figure 2: Selected Linear Models for Displacement

RSS: 6610.19 AIC: 2243.889 MSE: 17.68683

2.2 Linear Models for Horsepower

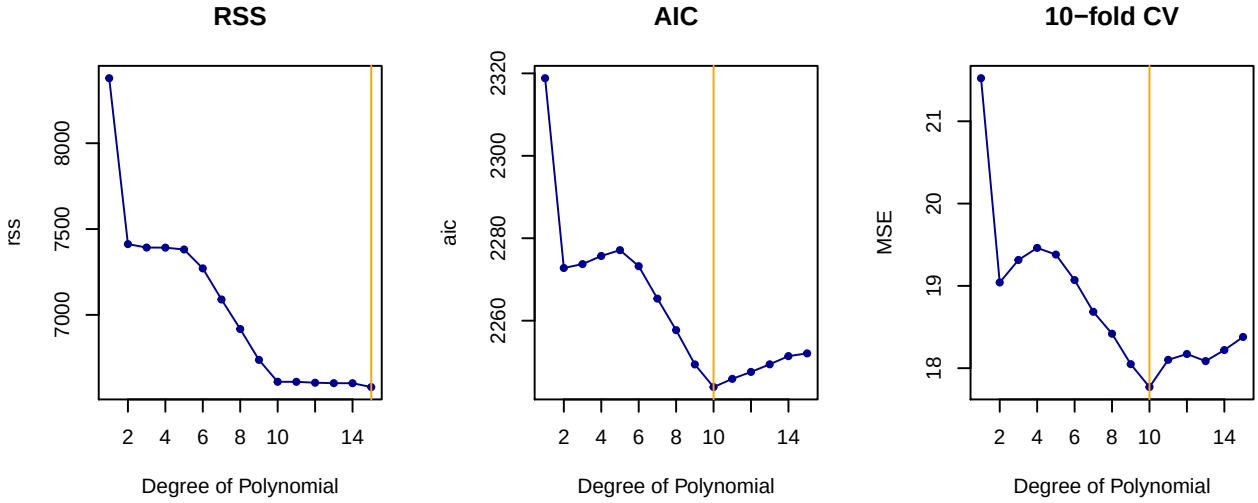

Figure 3: Candidate Linear Models for Horsepower

Similar to Displacement, figure 3 shows the optimal model for Horsepower is also with degree 10. The model is illustrated in figure 4.

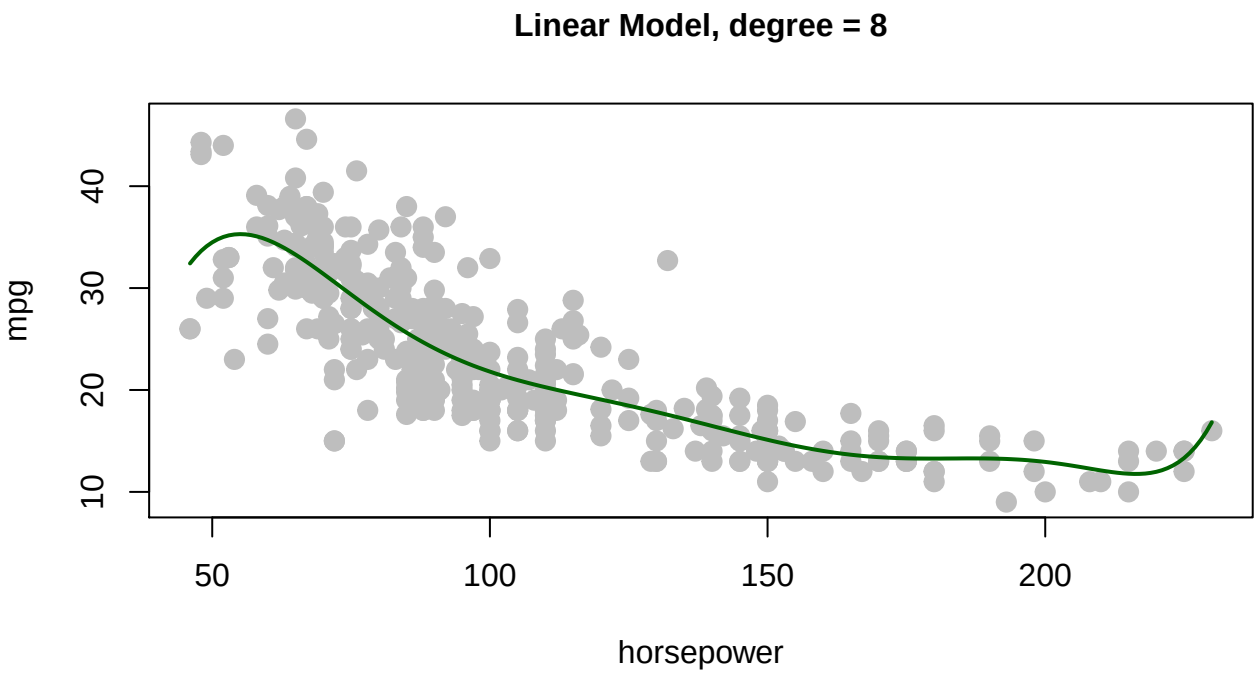

Figure 4: Selected Linear Models for Horsepower

RSS: 7081.923 AIC: 2266.911 MSE: 18.83704

2.3 Linear Models for Weight

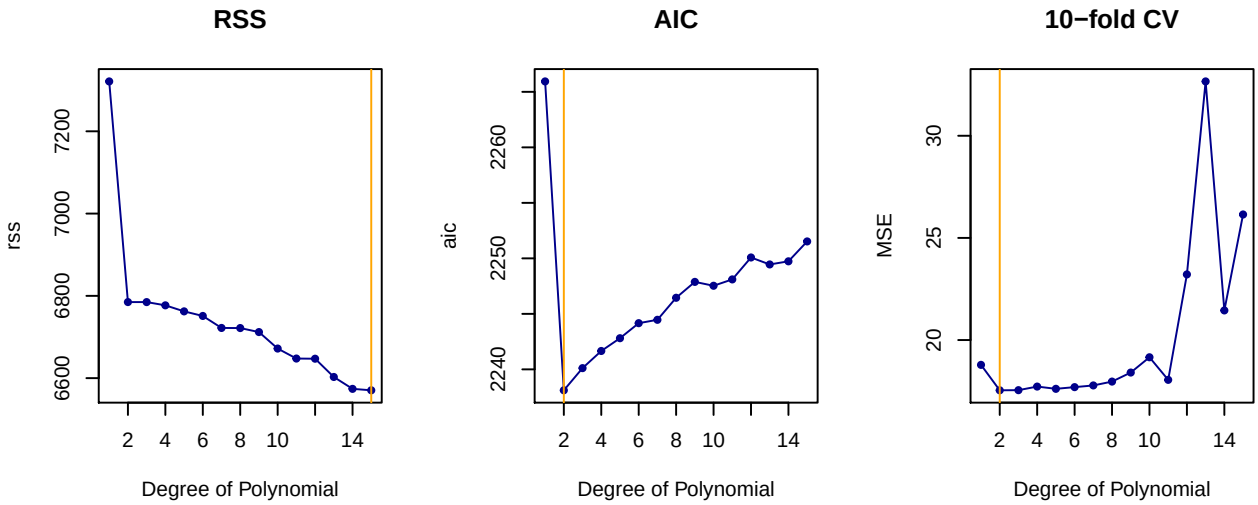

Figure 5: Candidate Linear Models for Weight

Figure 5 illustrates that the linear models created using Weight have different trends from the previous two. The generalisation errors based on 10-fold cross validation show slight variation between degree 2 and 8, while AIC score shows a steep fall from degree 1 to 2 and then steadily increace.

The optimal model for Weight, shown in figure 6, is selected to be the one with polynomial degree equal to 2.

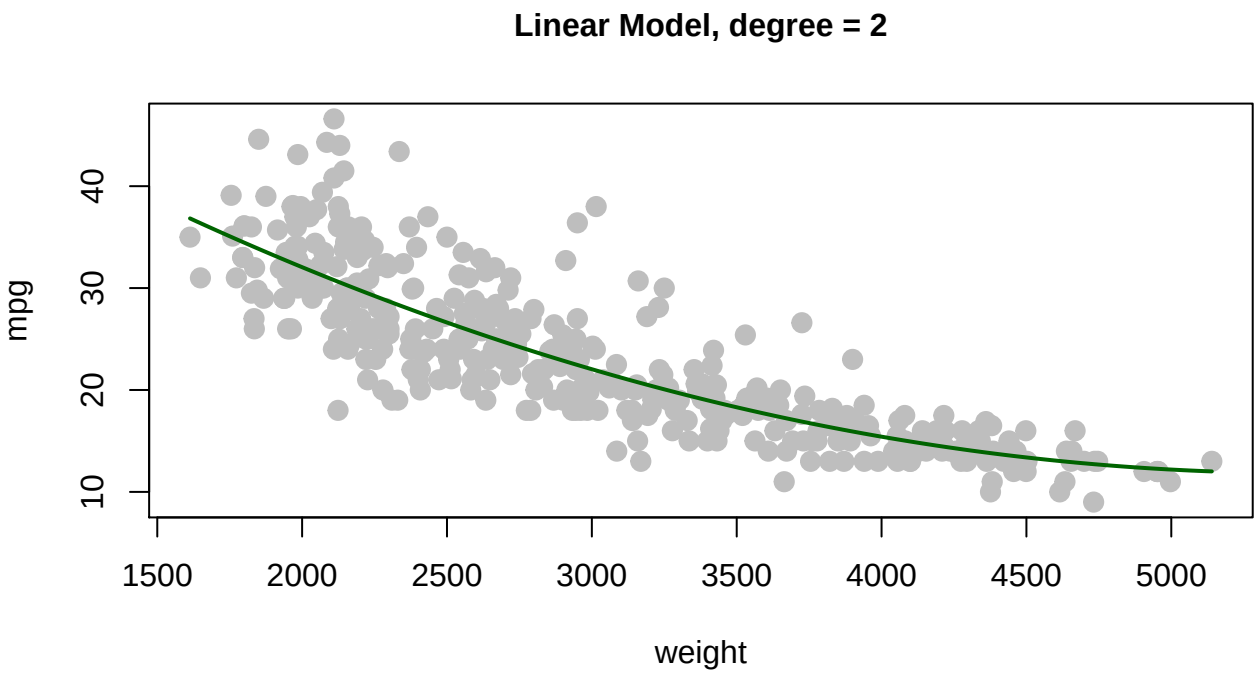

Figure 6: Selected Linear Models for Weight

RSS: 6784.899 AIC: 2238.115 MSE: 17.43871

2.4 Linear Models for Acceleration

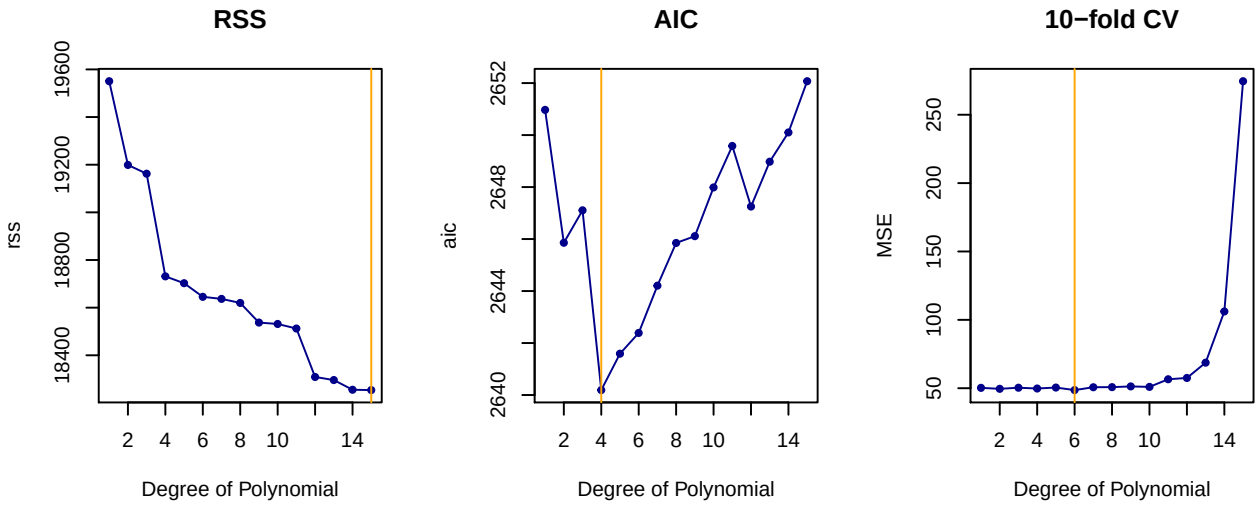

Figure 7: Candidate Linear Models for Acceleration

According to figure 7, the MSE barely varies from degree 1 to 10, but the error values are considerably larger than all the previous. The optimal model for Acceleration is selected to be the one with the lowest AIC score, degree equal to 4 and shown in figure 8.

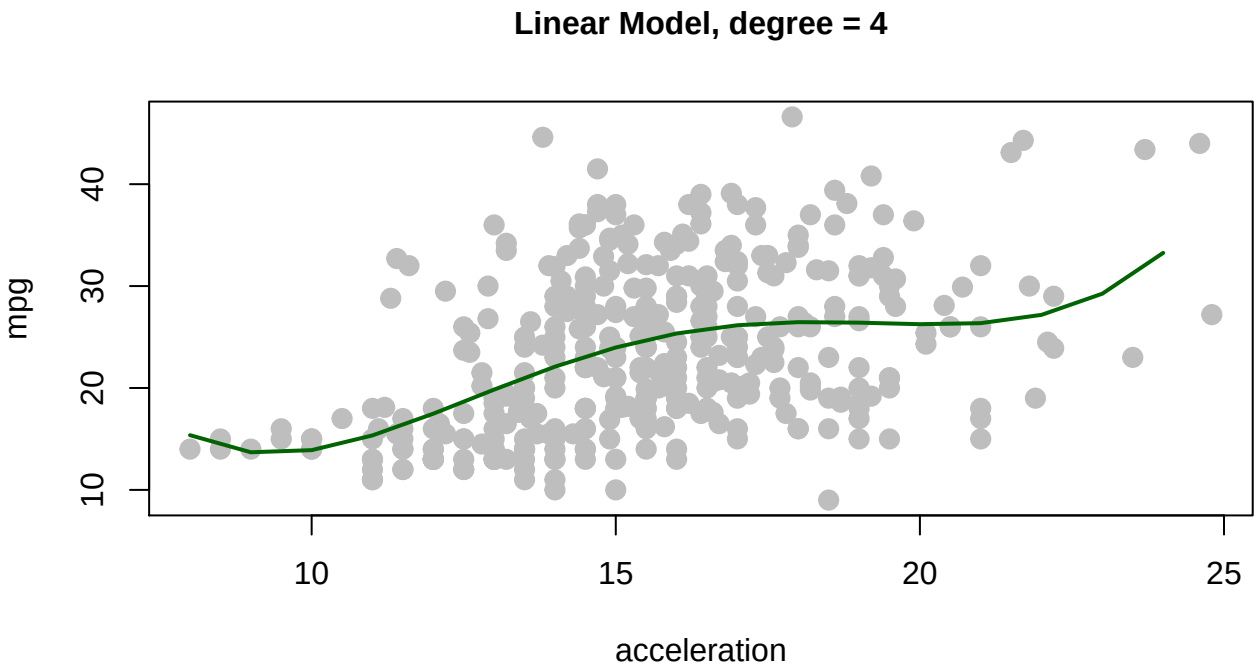

Figure 8: Selected Linear Models for Acceleration

RSS: 18731.31 AIC: 2640.19 MSE: 49.24686

2.5 Summary for Linear Models

On the basis of the linear models implemented in the section, it is easy to conclude that the *Residual Sum of Square* monotonically decreases with the increase of model comlexity, namely the degree of polynomial terms in this case. As to *Akaike information criterion*, the penalty on the number of model parameters would outweight the benefits of model complexity beyond a certain degree. *Mean Squard Error* based on cross validation will also increase when the model is overfitting.

Table 1: Summary of Model Fitting Measurement for the Selected Linear Model

	RSS	AIC	MSE
Displacement	6610.19	2243.89	17.69
Horsepower	7081.92	2266.91	18.84
Weight	6784.90	2238.12	17.44
Acceleration	18731.31	2640.19	49.25

Table 1 summarises the evaluation for the four optimal model based on Displacement, Horsepower, Weight and Acceleration. It's clear that the Acceleration model is the is far worse than the other three. Overall, the Displacement model and the Weight are as good at predition, while the Weight model with 2 degree is much simpler than degree 10 of the Displacement model.

3 Bin Smooth Models

Dataset is randomly split into two halves (each has 196 observations) in this section, one used as training data to fit models and the other as test data to obtain generalisation error. Four predictors are fitted into bin

smoorh models, and candidate models for each predictor are across bin length from 1 to 200. One optimal bin smooth model with the most reasonable bin length is selected for each predictor, and the four selected models competes with each other.

3.1 Bin Smooth Models for Displacement

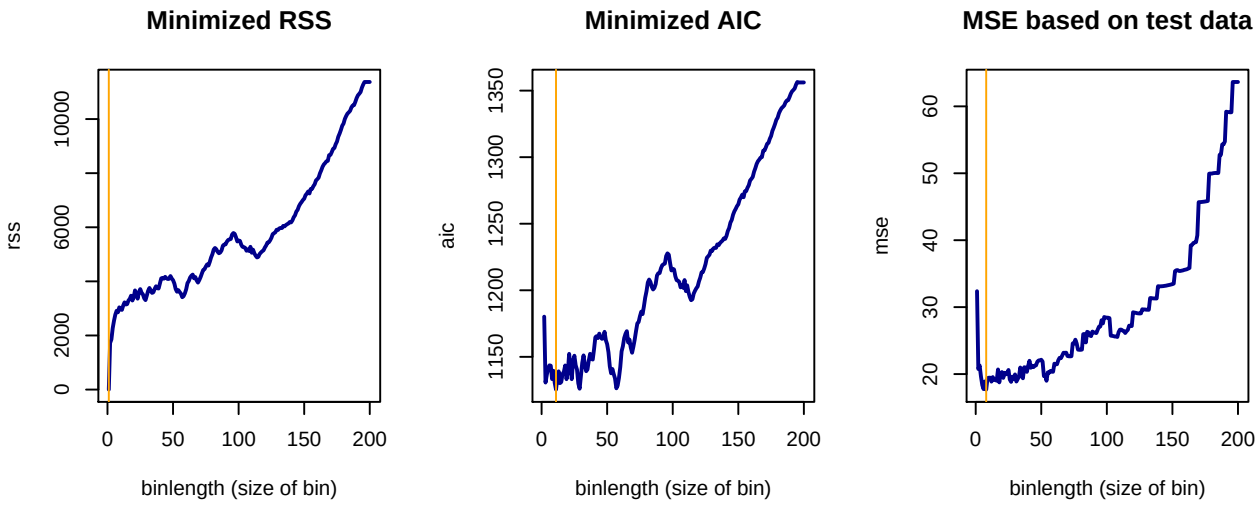

Figure 9: Candidate Bin Smooth Models for Displacement

Bin lenght: 10 Number of bins: 20 RSS: 2963.098 AIC: 1130.536 MSE: 19.44981

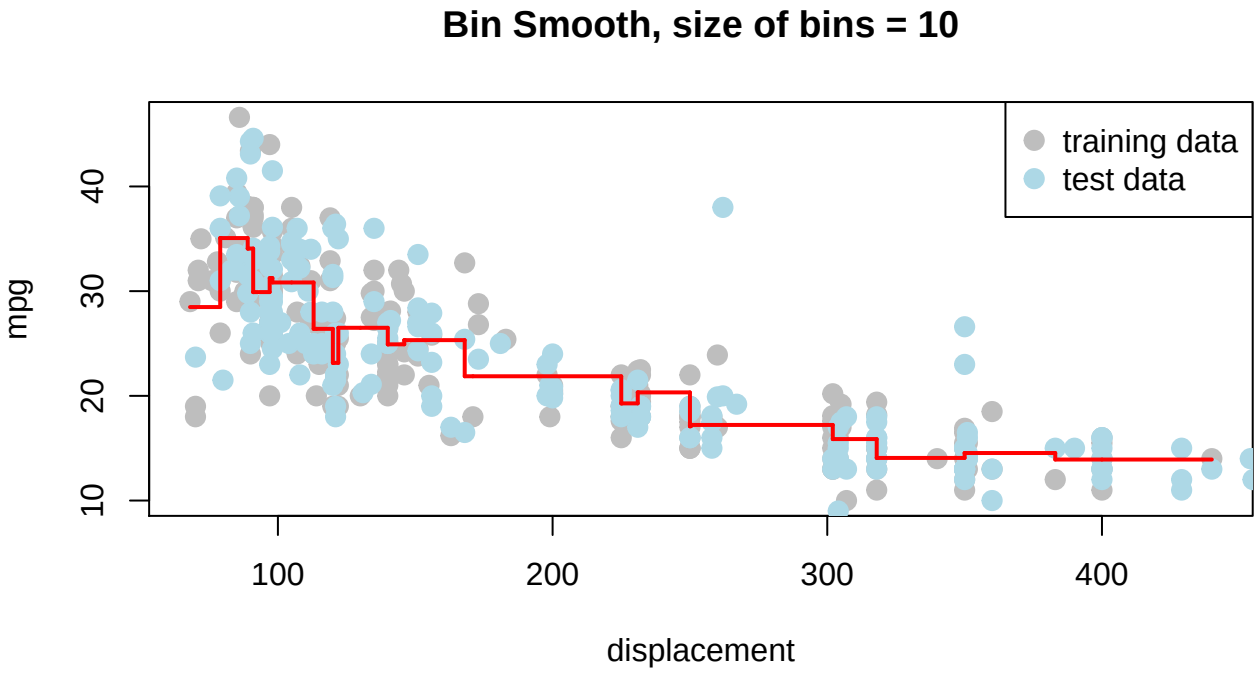

Figure 10: Selected Bin Smooth Models for Displacement

Figure 9 shows the most reasonable bin length is around 10, and figure 10 illustrates the optimal bin smooth model for predictor Displacement.

3.2 Bin Smooth Models for Horsepower

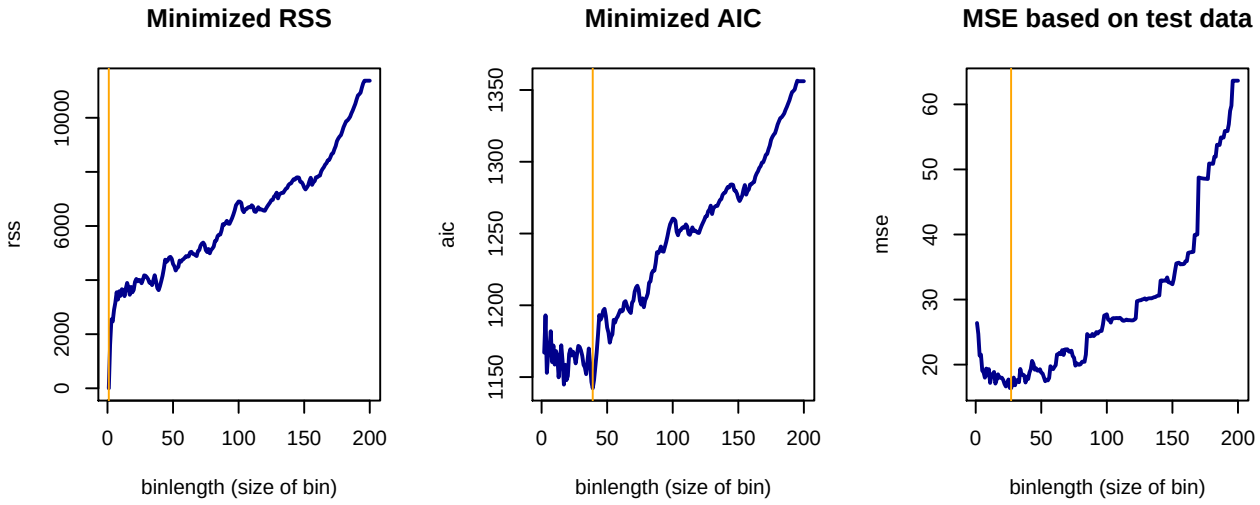

Figure 11: Candidate Bin Smooth Models for Horsepower

Bin lenght: 20 Number of bins: 10 RSS: 3415.697 AIC: 1138.396 MSE: 20.30473

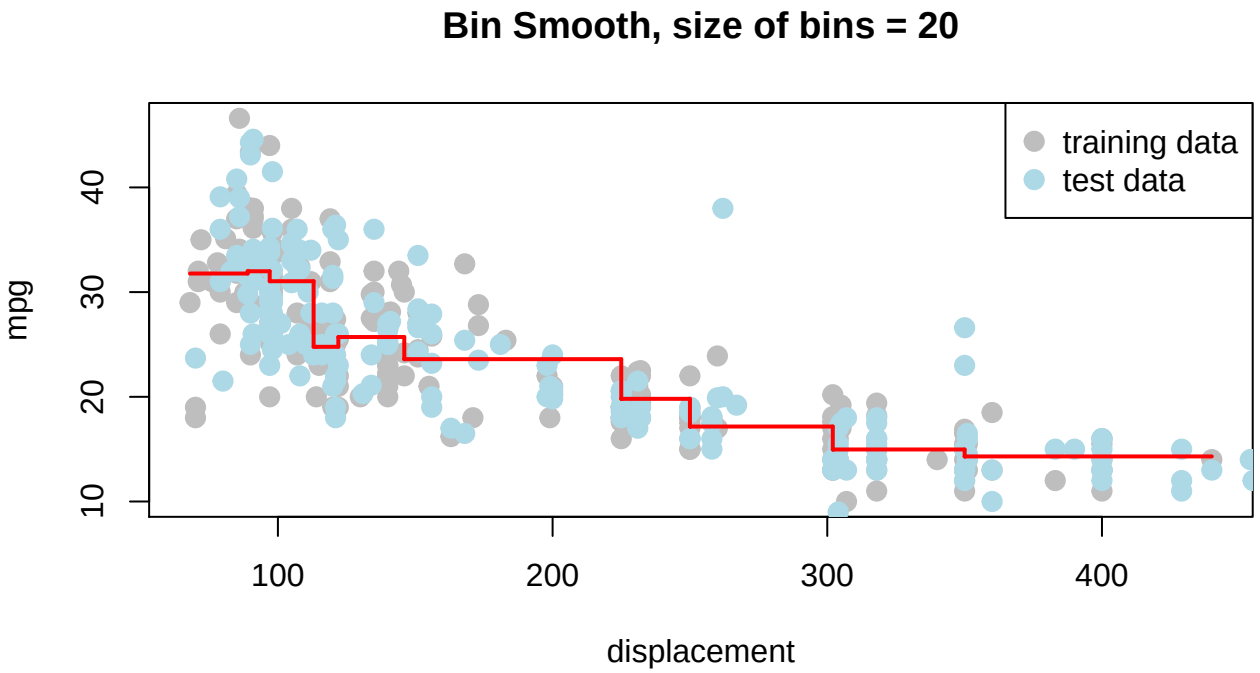

Figure 12: Selected Bin Smooth Models for Horsepower

Figure 11 shows the most reasonable bin length fall between 10 - 40, and figure 12 illustrates the selected optimal bin smooth model for predictor Horsepower, with bin length equal to 20.

3.3 Bin Smooth Models for Weight

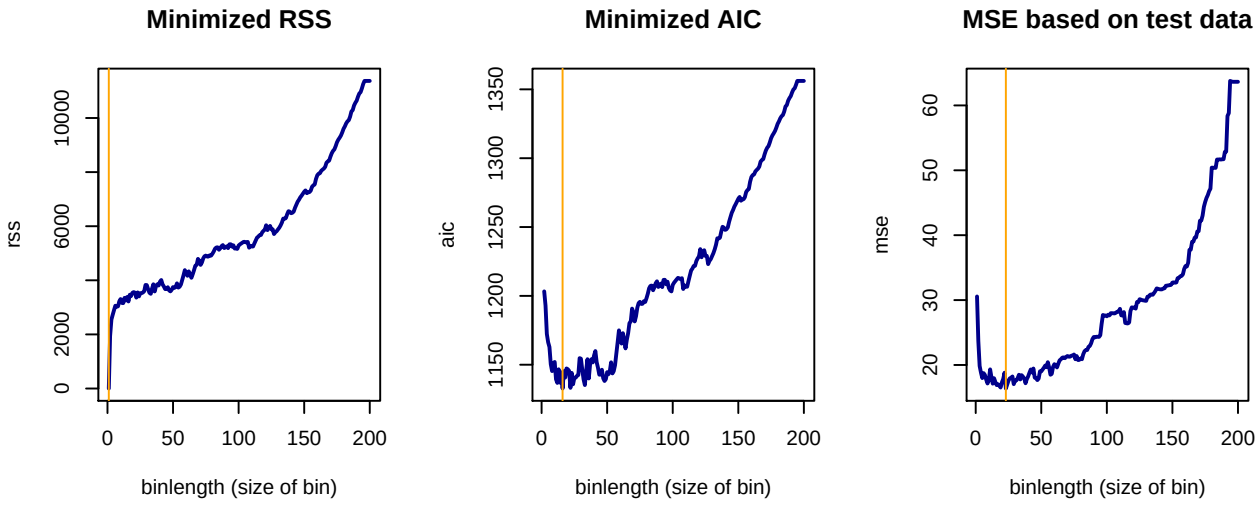

Figure 13: Candidate Bin Smooth Models for Weight

Bin lenght: 20 Number of bins: 10 RSS: 3563.595 AIC: 1146.704 MSE: 17.32378

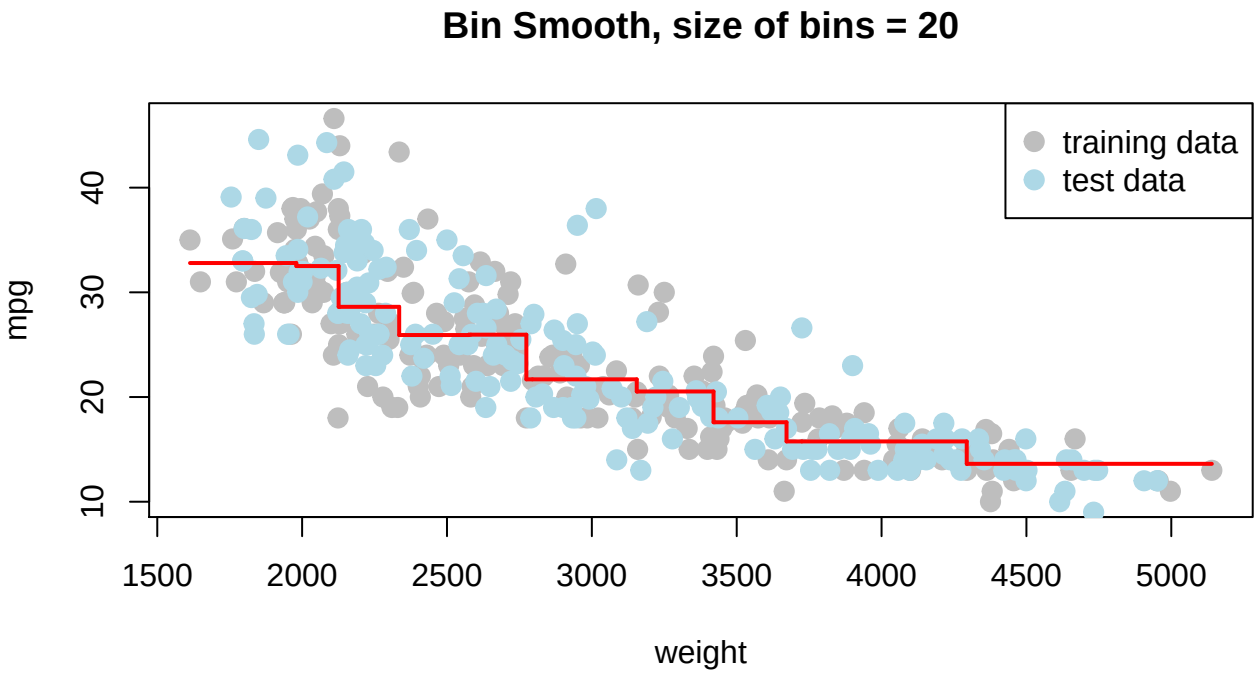

Figure 14: Selected Bin Smooth Models for Weight

Figure 13 shows the most reasonable bin length fall between 10 - 30, and figure 14 illustrates the selected optimal bin smooth model for predictor Weight, with bin length equal to 20.

3.4 Bin Smooth Models for Acceleration

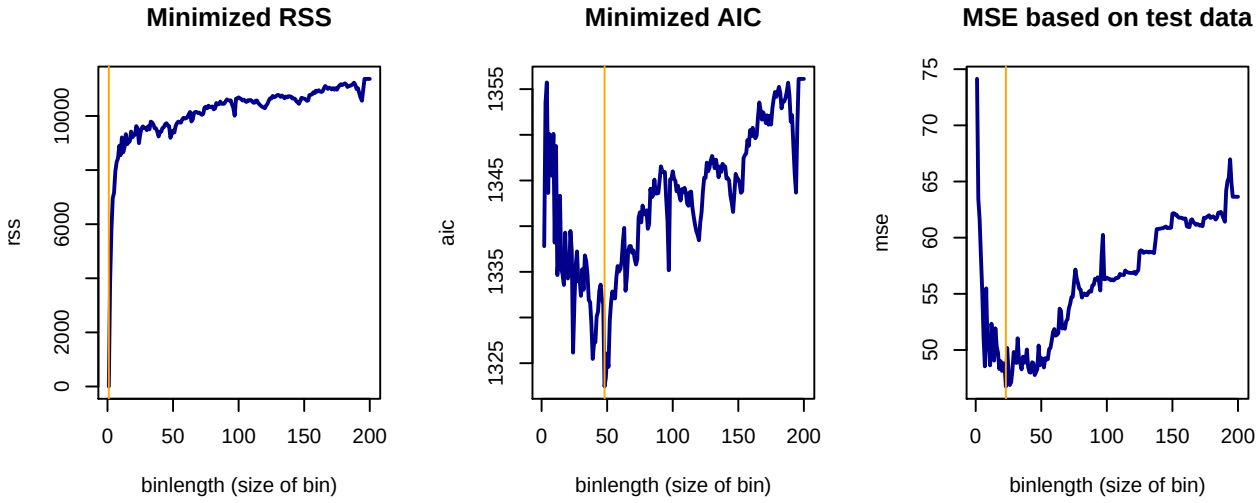

Figure 15: Candidate Bin Smooth Models for Acceleration

Bin lenght: 25 Number of bins: 8 RSS: 9339.012 AIC: 1331.537 MSE: 48.09347

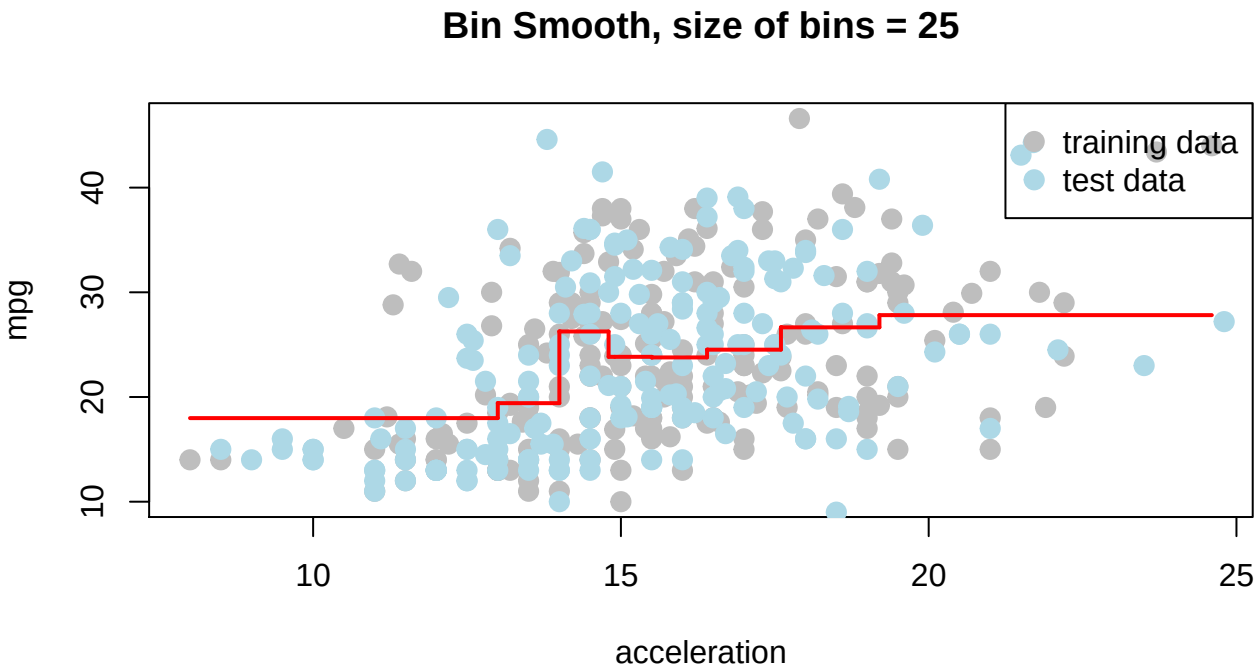

Figure 16: Selected Bin Smooth Models for Acceleration

Figure 15 shows that the *AIC* and *MSE* lines are way more wiggly than all the previous, which indicates the smooth models for Acceleration have much uncertainty than others. And similar to linear model section, the *AIC* and *RSS* are also reletively higher than models based on other three predictors.

Figure 16 illustrates the selected optimal bin smooth model for predictor Acceleration, with bin length equal to 25.

3.5 Summary for Bin Smooth Models

The most obvious pattern within bin smooth models is that decreasing the number of bins (increasing the size) would increase *RSS*, and increasing the number of bins (reducing the size) would lessen *RSS*. Generally, four groups of models all show better predictive ability at smaller bin length than at large bin length. However, excess of bins (namely bin length too small) usually leads to overfitting, despite the lower *RSS*.

Table 2: Summary of Model Fitting Measurement for the Selected Bin Smooth Model

	RSS	AIC	MSE	Adjusted R2
Displacement	2963.10	1130.54	19.45	0.71
Horsepower	3415.70	1138.40	20.30	0.69
Weight	3563.60	1146.70	17.32	0.67
Acceleration	9339.01	1331.54	48.09	0.15

The measurements of four optimal models are summarized below in table 2. *AIC* scores of bin smooth models are around 1000, clearly lower than those of linear models (around 2000), because bin smooth model is simpler than high degree linear regression model. Acceleration is still the worst predictor, the three other models have similar predictive abilities. Displacement model outdo other three models in terms of *RSS*, *AIC* and adjusted R^2, but the generalisation error (*MSE*) is slightly higher than of Weight model.

4 B-spline Models

In this section, four predictors are respectively fitted to four group of b-spline models. There are 30 candidate models in each group, with the number of knots from 1 to 10 and degree from 1 to 3. The knots are placed in a uniform fashion. One optimal model is selected for each group/predictor, and four optimal b-spline model compete. The training data and test data are the same as used in the last section. The generalisation error is the *MSE* of prediction on test data.

4.1 B-spline Models for Displacement

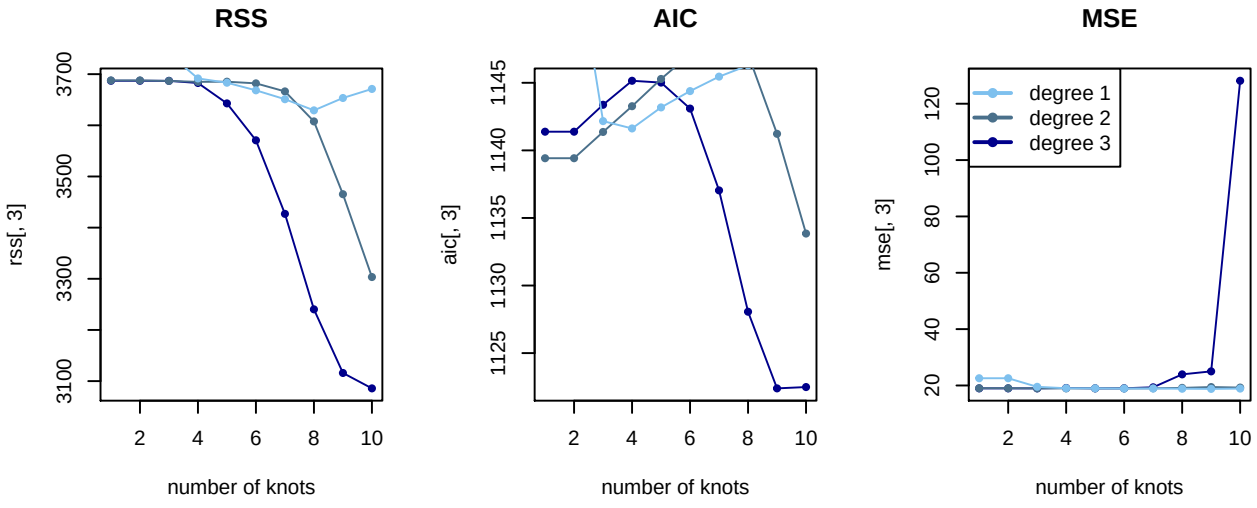

Figure 17: Candidate B-spline Models for Displacement

Figure 17 shows that degree 3 models have loewr *RSS* than models with degree 1 and 2, and *AIC* scores are also lower when the number of knots exceeding 6. However, degree 3 models highly diverge with large number of knots, resulting extreme generalisation errors.

The optimal b-spline model for Displacement is selected to be the one with 7 knots and degree equal to 3, shown in figure 18. The vertical dotted lines indicate positions of the knots. (hereafter the same)

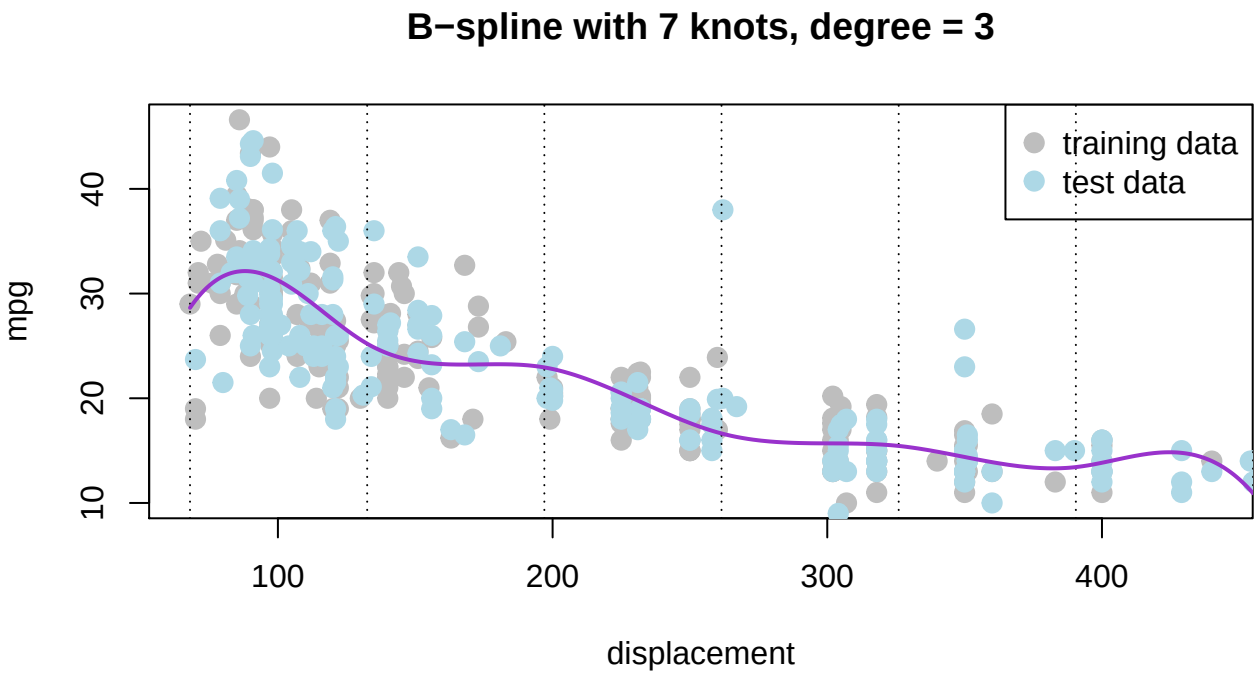

Figure 18: Selected B-spline Models for Displacement

RSS: 3426.912 AIC: 1137.039 MSE: 19.3498

4.2 B-spline Models for Horsepower

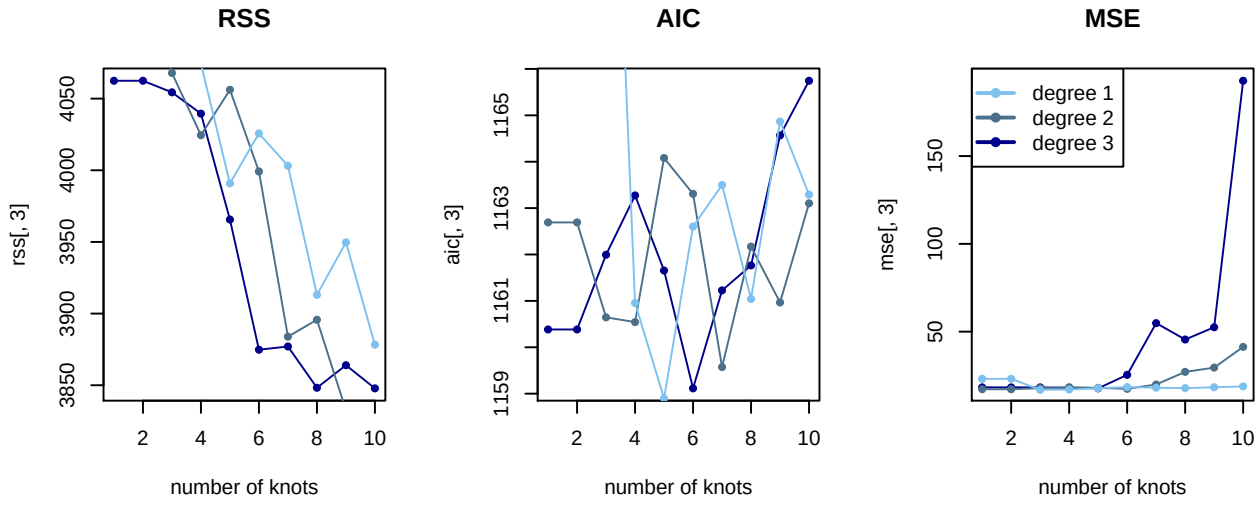

Figure 19: Candidate B-spline Models for Horsepower

Figure 19 shows that *RSS* of degree 3 models are generally lower than others but *MSE* diverge since the number of knots surpass 5. The *AIC* plot seems jumbled up, but the fluctuation range is generally within 5, not problematic.

The optimal b-spline model for Horsepower is selected to be the one with 5 knots and degree equal to 3, shown in figure 20.

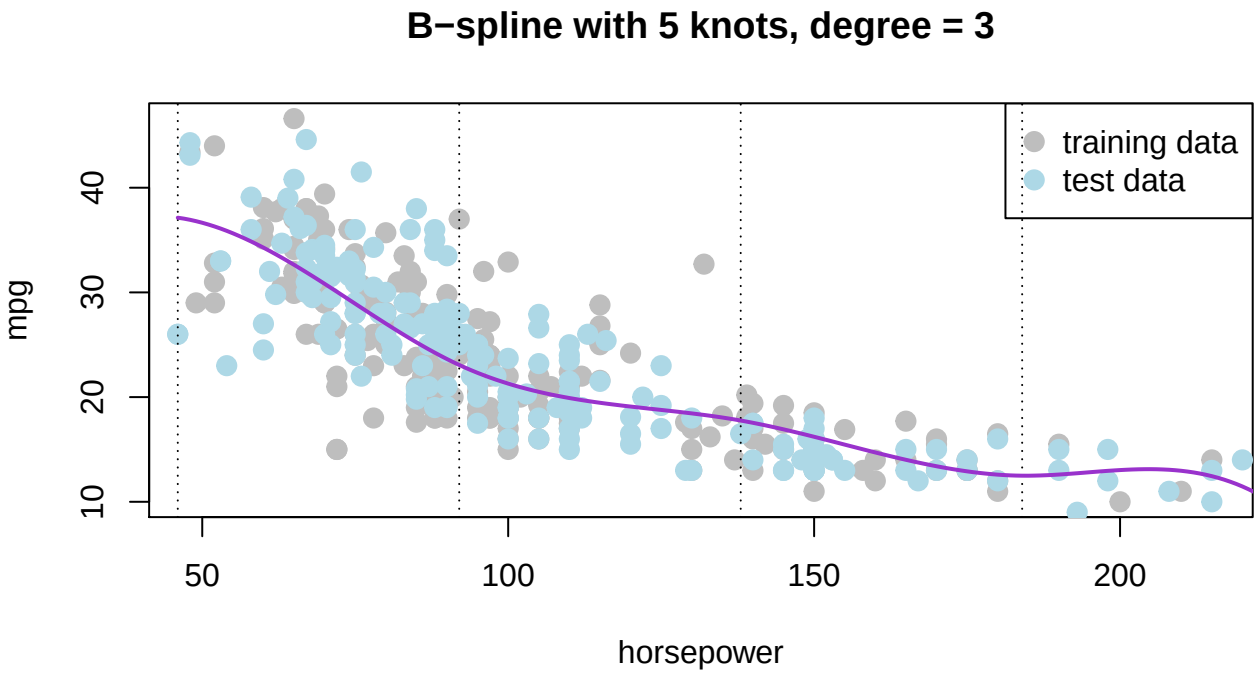

Figure 20: Selected B-spline Models for Horsepower

RSS: 3965.61 AIC: 1161.655 MSE: 17.69047

4.3 B-spline Models for Weight

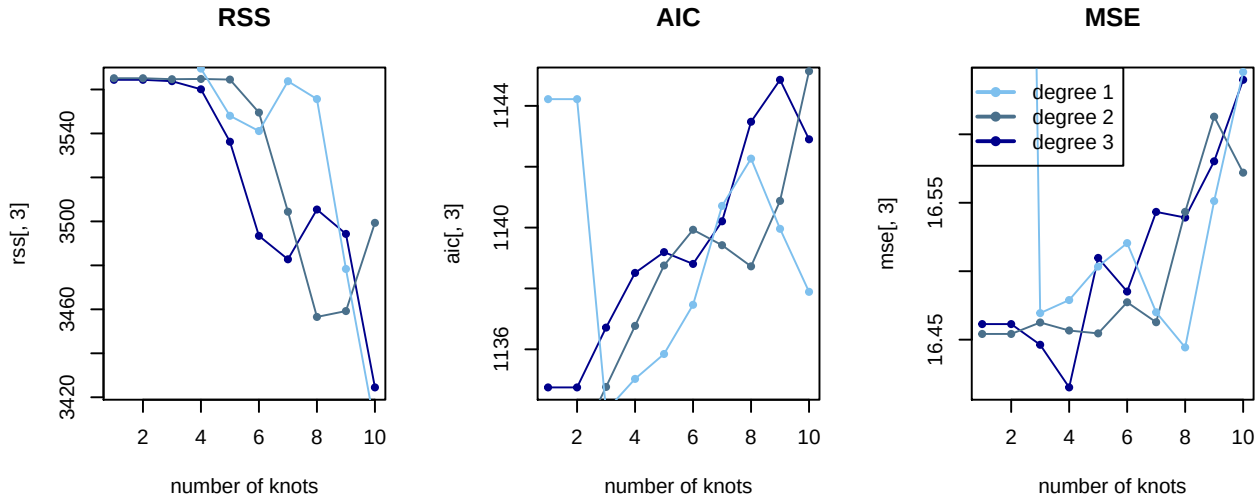

Figure 21: Candidate B-spline Models for Weight

Figure 21 shows that *AIC* and *MSE* trend to increase after number of knots surpass 3 and 4. The optimal b-spline model for Weight is selected to be the one with 4 knots and degree equal to 3, shown in figure 22.

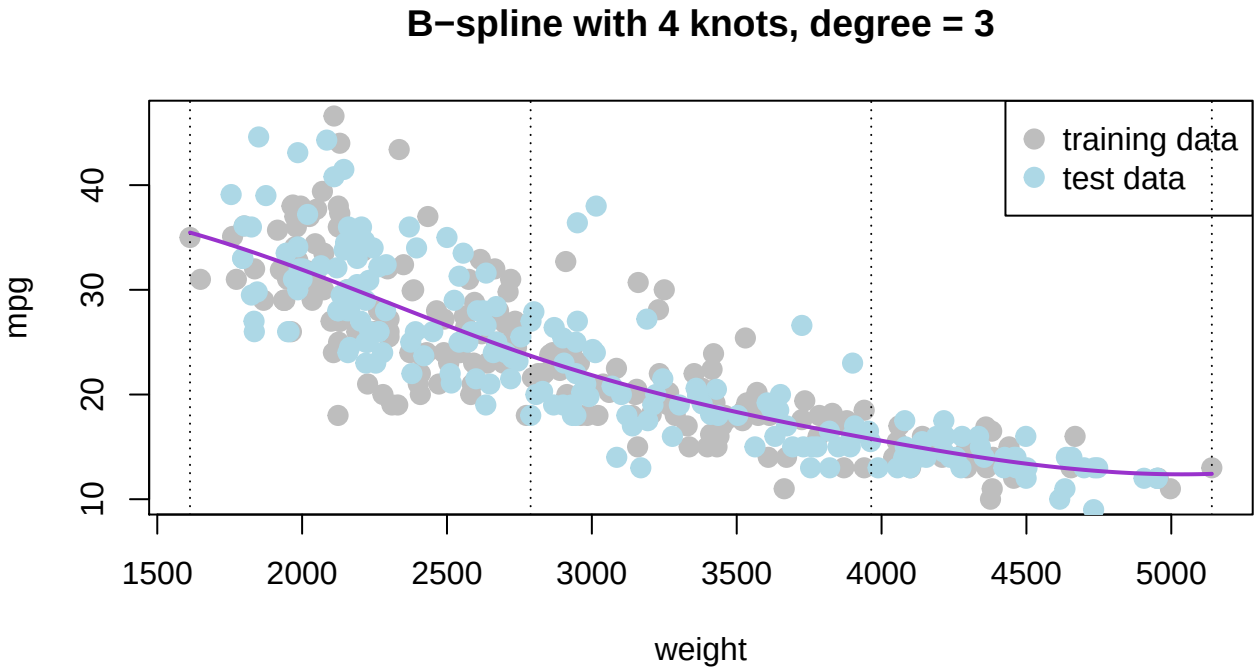

Figure 22: Selected B-spline Models for Weight

RSS: 3560.09 AIC: 1138.512 MSE: 16.41512

4.4 B-spline Models for Acceleration

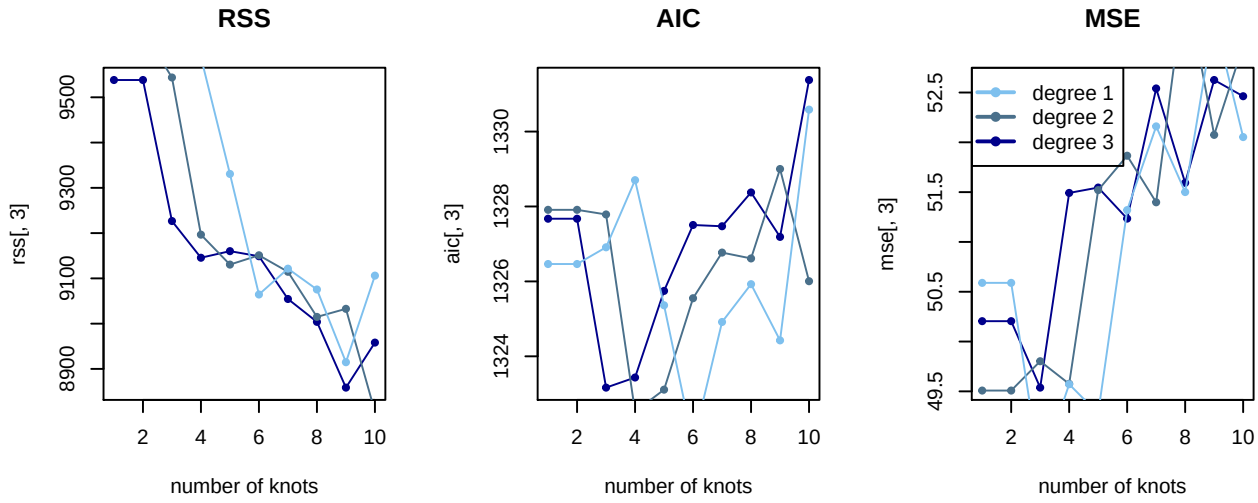

Figure 23: Candidate B-spline Models for Acceleration

Figure 23 shows that there is not much difference between models with three degrees. The optimal b-spline model for Weight is selected to be the one with 4 knots and degree equal to 2, shown in figure 24.

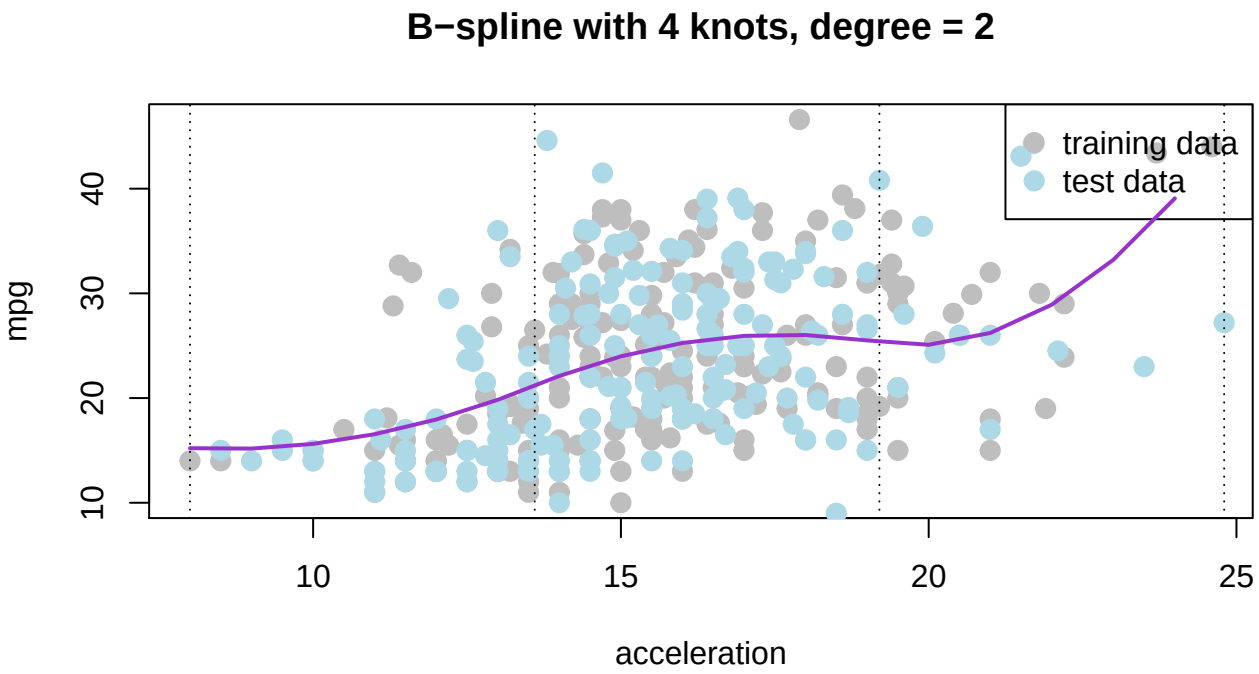

Figure 24: Selected B-spline Models for Acceleration

RSS: 9196.567 AIC: 1322.524 MSE: 49.57517

4.5 Summary for B-spline Models

Generally, the number of knots and degree of polynomial determine the flexibility of the model: the more the knots, the more flexible the model; the higher the degree, the more flexible the model. It'd better to mannually place knots at where we feel the underlying function might vary rapidly, but in practice it is more feasible to place knots evenly across the range of the covariate. Very importantly, severe diverge can happen when fitting with too many knots and higher degree.

The measurements of four selected b-spline models are summarized in table 3. The Acceleration model is still the worst, and the three other models are as good in general. Similar to the previous section, the Displacement model has the best *RSS*, *AIC* and adjusted R^2, and the Weight model has the best generalisation error (*MSE*).

Table 3: Summary of Model Fitting Measurement for the Selected B-spline Model

	RSS	AIC	MSE	Adjusted R2
Displacement	3426.91	1137.04	19.35	0.69
Horsepower	3965.61	1161.65	17.69	0.64
Weight	3560.09	1138.51	16.42	0.68
Acceleration	9196.57	1322.52	49.58	0.17

5 Conclusion

Weight is regarded to have the best predictive ability, because three kinds of models fitted with it have the lowest generalisation errors in comparison to models fitted with other predictors. Displacement is almost

as good as Weight with slightly higher generalisation errors, and the models fitted with it have the lowest RSS, AIC and the highest adjusted R^2. Acceleration is incompetent at predicting MPG, because it is barely correlated to mpg, as shown in the scatter plots.

6 Reference

Canty, A. & Ripley, B.D. (2017) R package version 1.3-20. *Boot: Bootstrap r (s-plus) functions.*

R Core Team (2018) *R: A language and environment for statistical computing.* [Online]. Vienna, Austria, R Foundation for Statistical Computing. Available from: https://www.R-project.org/.

Wickham, H. (2017) R package version 1.2.1. *Tidyverse: Easily install and load the 'tidyverse'.* [Online]. Available from: https://CRAN.R-project.org/package=tidyverse.

Xie, Y. (2018) R package version 1.20. *Knitr: A general-purpose package for dynamic report generation in r.* [Online]. Available from: https://yihui.name/knitr/.