Homework Chapter 7

Jacob Thielemier

20 March 2024

Question 1

Part (a) Here we penalize the g and a infinite λ means that this penalty dominates. This means that the $\hat{g}$ will be 0 and the fitted curve $\hat{g}$ will be a horizontal line at the level that minimizes the vertical distances to all the data points.

Part (b) Here we penalize the first derivative (the slope) of g and a infinite λ means that this penalty dominates. Thus the slope will be 0 and the fitted curve $\hat{g}$ will be a straight line that minimizes the sum of squared residuals from the data points.

Part (c) Here we penalize the second derivative (the change of slope) of g and a infinite λ means that this penalty dominates. Thus the line will be straight (and otherwise best fitting x).

Part (d) Here we penalize the third derivative (the change of the change of slope) of g and a infinite λ means that this penalty dominates. In other words, the curve will have a consistent rate of change and the curve $\hat{g}$ will be a parabola.

Part (e) Here we penalize the third derivative, but a value of $\lambda = 0$ means that there is no penalty. The resulting $\hat{g}$ will pass through or very close to every data point, and could potentially be a high-degree polynomial that fits the data points exactly, which means it can have a very wiggly shape, overfitting the data.

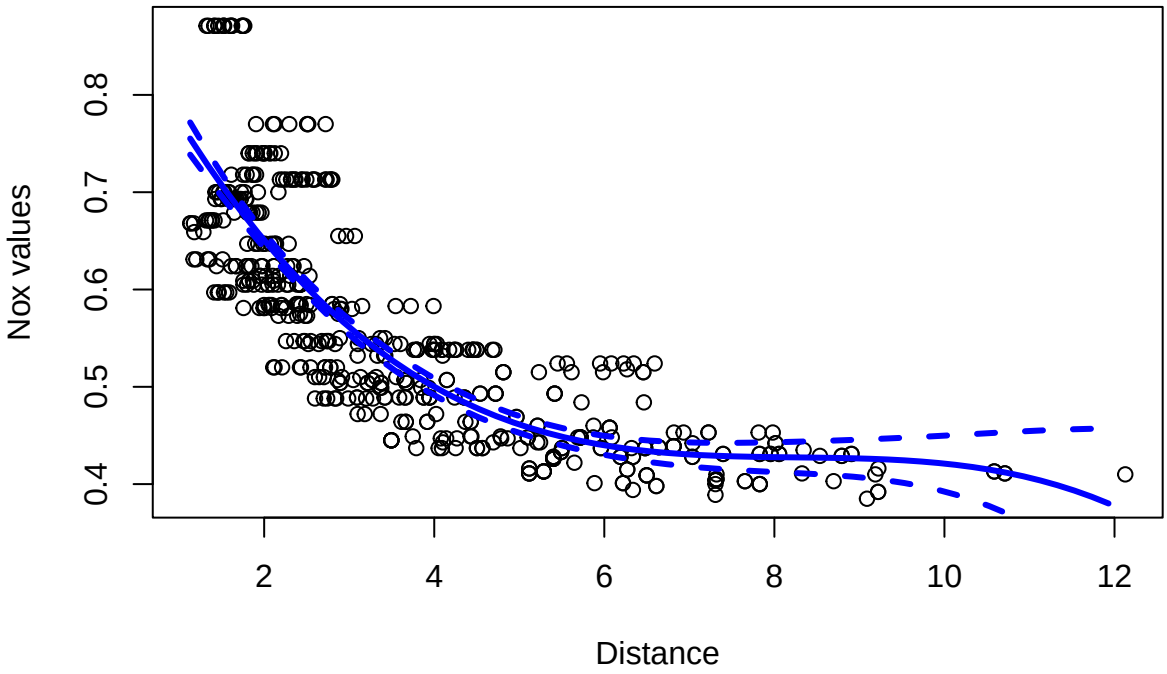
Question 2

Part (a) $\hat{g}_2$ is more flexible (by penalizing a higher derivative of g) and so will have a smaller training RSS.

Part (b) We cannot tell which function will produce a smaller test RSS, but there is chance that $\hat{g}_1$ will if $\hat{g}_2$ overfits the data.

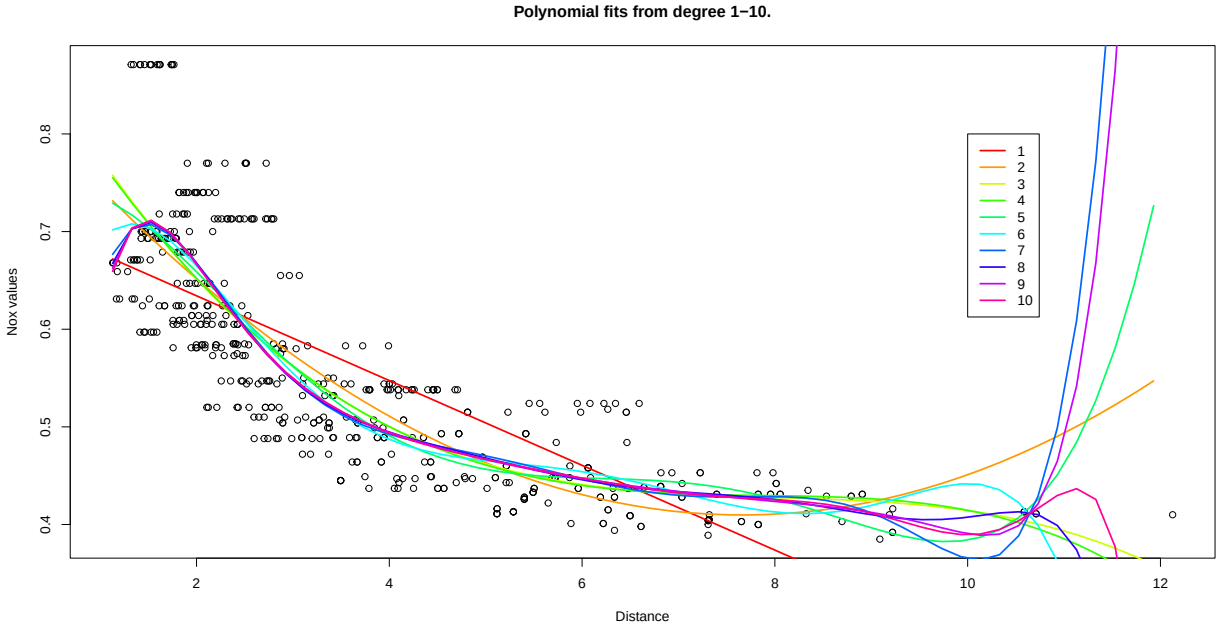
Part (c) There's no penalty for model complexity, so both $\hat{g}_1$ and $\hat{g}_2$ will fully fit the training data without considering the complexity of the model. Therefore, they might end up with the same training RSS if they can both fit the data well. For the test RSS, the model that matches the true complexity of the underlying data generation process will perform better.

Question 3

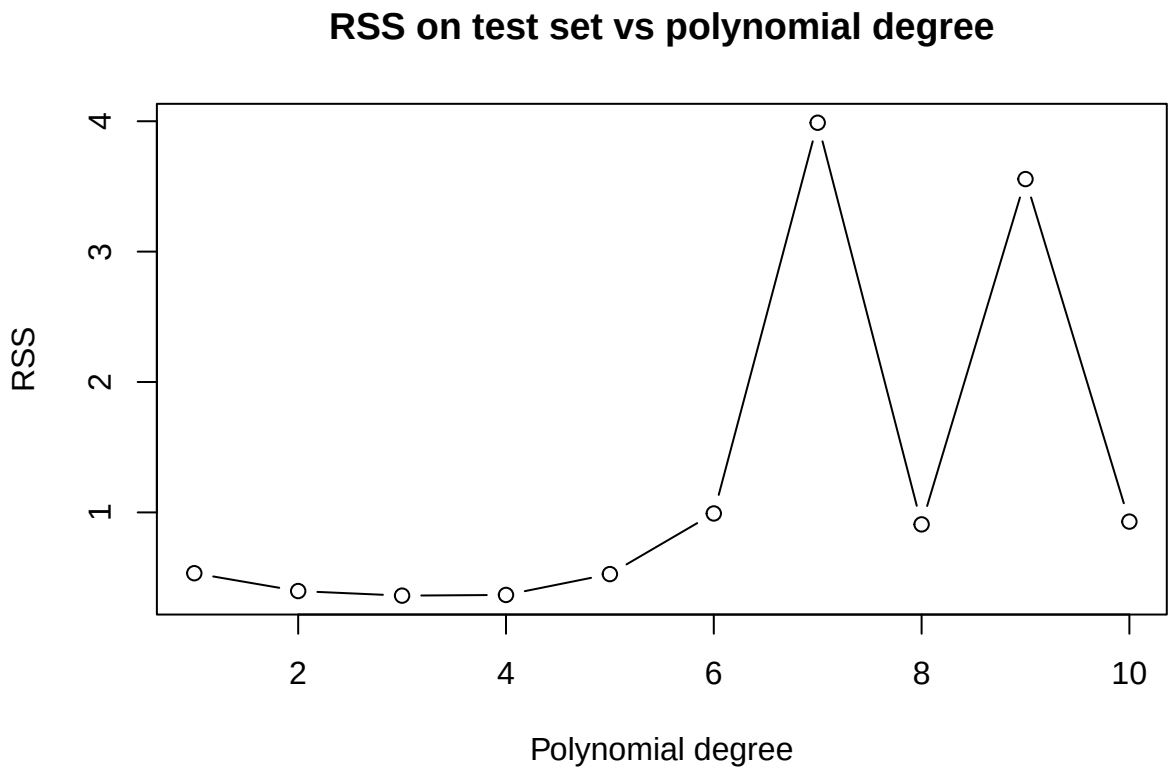


Part (a)

- The regression summary shows a cubic fit is statistically significant. The cubic fit is plotted on the chart, and does appear to match the underlying shape of the data.



- The RSS decreases from the linear (0.533) to the cubic model (0.361), and increases thereafter. This supports the argument that the cubic model provides the best fit.



Part (b)

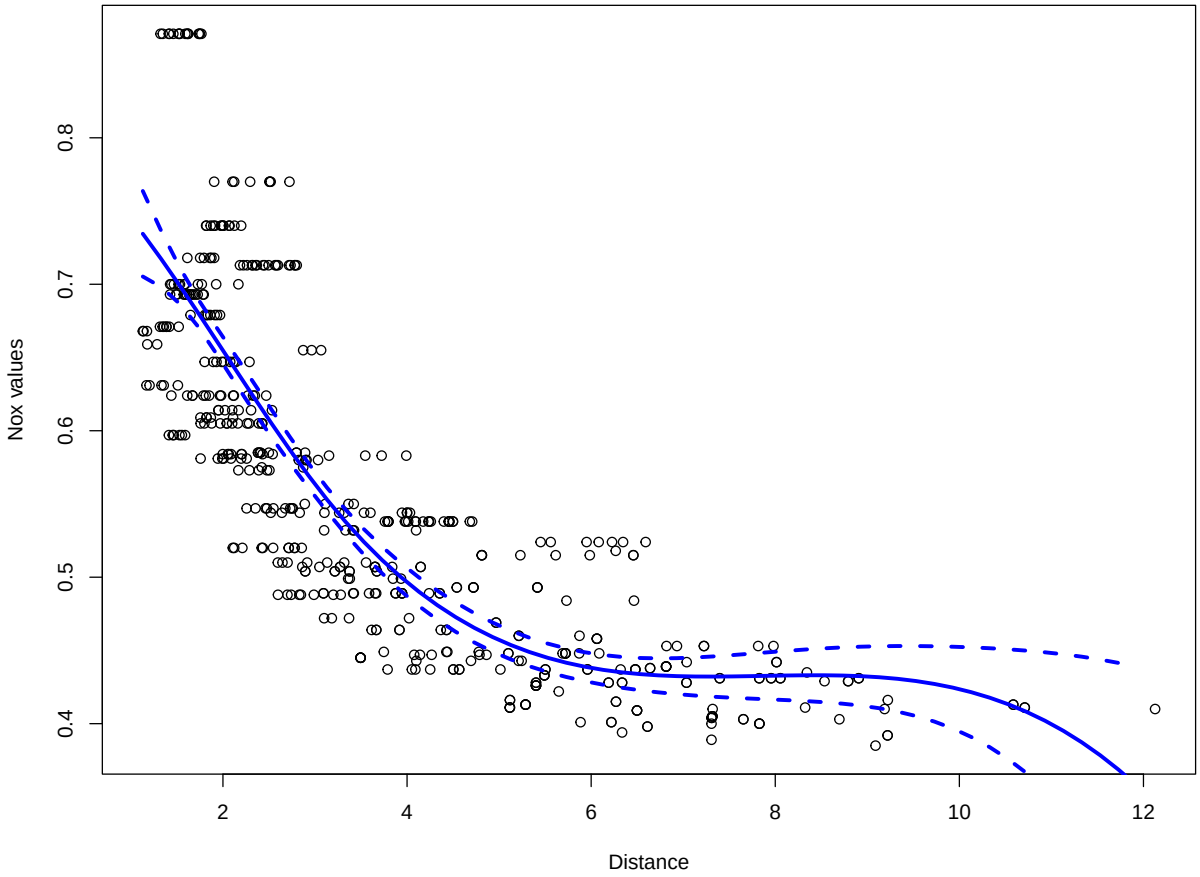
- RSS is at a minimum for the degree 3 polynomial.

Part (c)

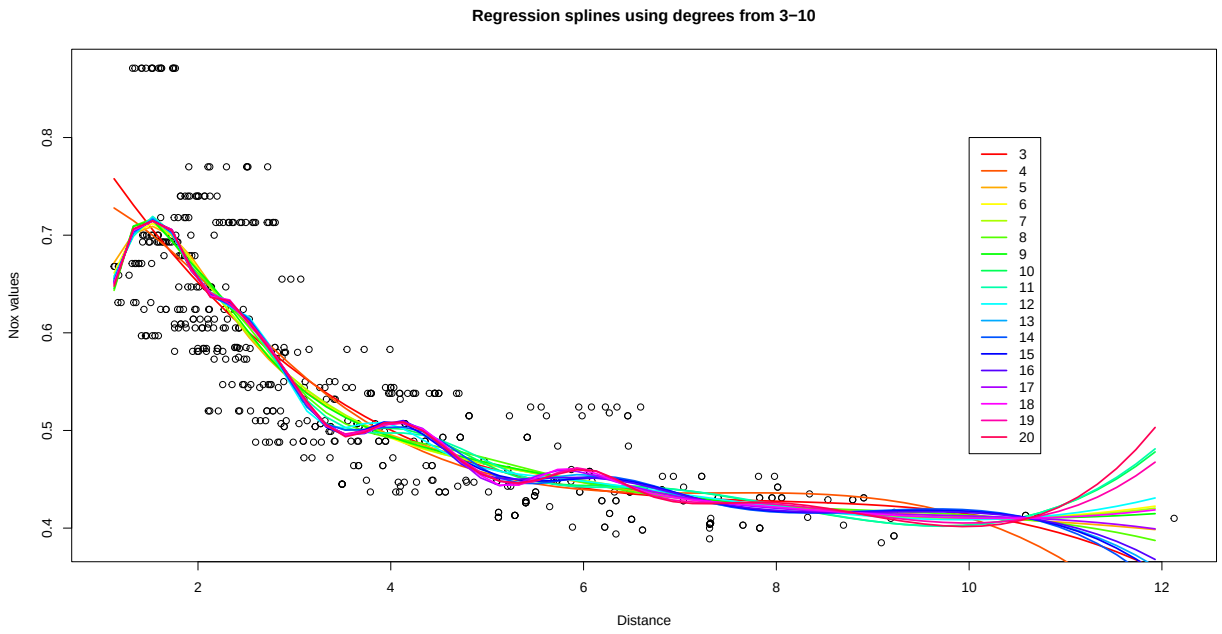
- The optimal degree is 3 based on cross-validation. Higher values tend to lead to overfitting.

Part (d)

- A regression spline with four degrees of freedom is statistically significant. The knots are chosen automatically when using the df() function. In this case we have single knot at the 50th percentile value.



- The resulting spline fit is very similar to that of polynomial regression using degree 3.



Part (e)

- Smaller differences between spline fits than with the polynomial fits. RSS is the lowest for the degree 12 model.

Part (f)

Cross validation:

```
##           8
## 0.003693139
```

- The minimum for the CV errors is using degree 8. This is different to the degree 12 model found using a validation set.

Cross validation using cv.glm() function:

```
## [1] 8
```

- The cv.glm() method finds a minimum at degree 8. This is the same degree found using the previous cross validation method, but different to using a validation set.

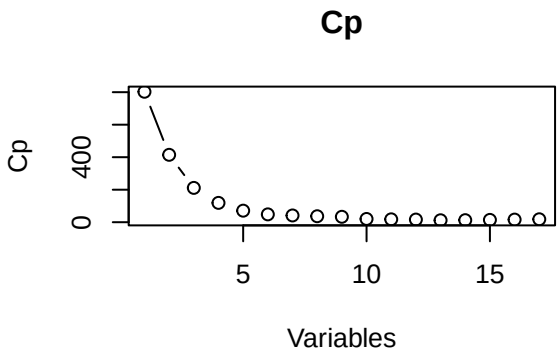
Question 4

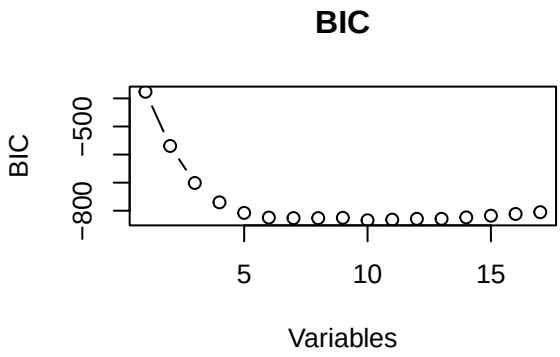
Part (a)

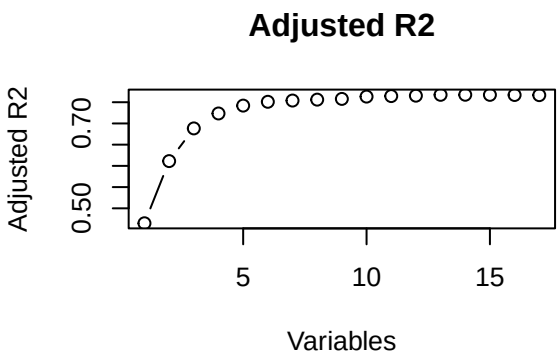
```
## [1] 13
```

```
## [1] 10
```

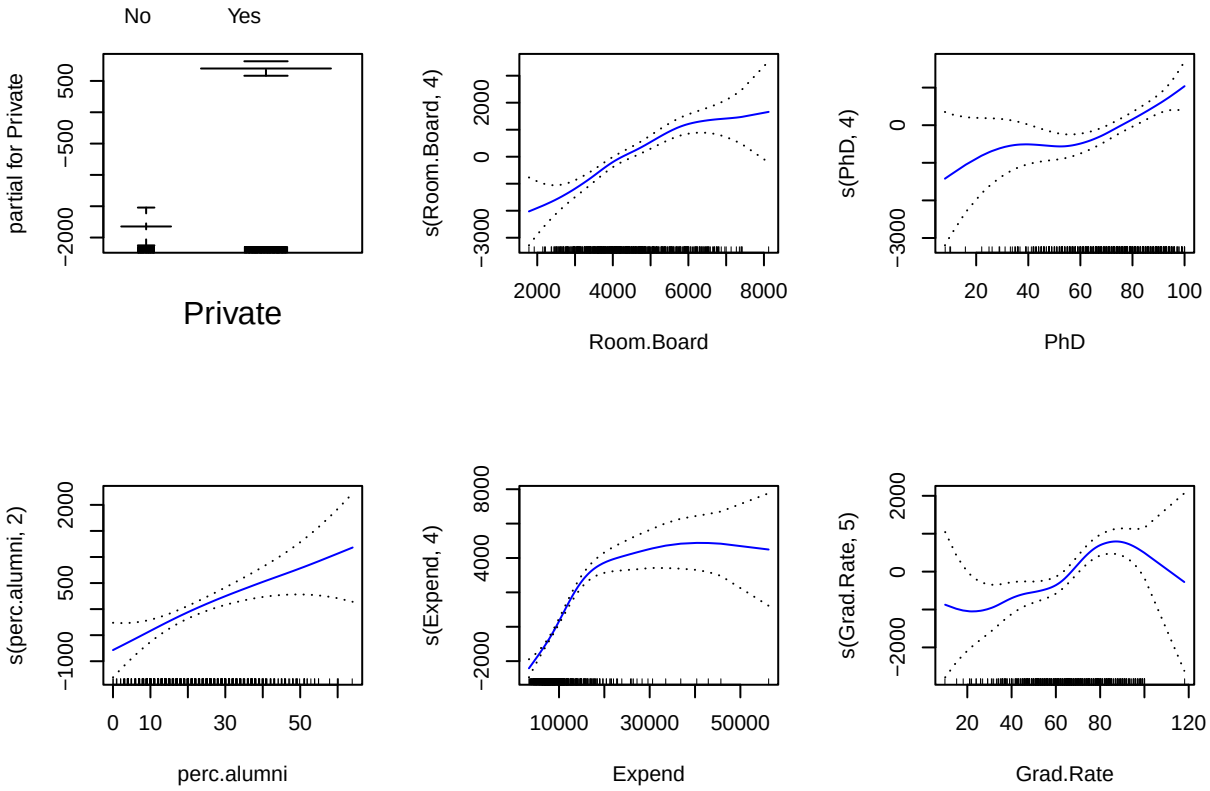
```
## [1] 14
```







- The Cp, BIC and Adjusted R^2 all identify minimums and a maximum for models with a different number of variables. As can be seen from the charts, the metrics change rapidly as the number of variables increase, but there are only small improvements after a model with 6 variables. The model with 6 variables is selected because it appears to be better at describing this relationship.



Part (b)

- Holding other variables fixed, out of state tuition increases as room and board costs get higher. Similarly, out of state tuition increases as the proportion of alumni who donate increase.

Part (c)

```
## [1] 3753134
```

```
## Analysis of Deviance Table
##
## Model 1: Outstate ~ Private + s(Room.Board, 4) + s(PhD, 4) + s(perc.alumni,
##     2) + s(Expend, 4)
## Model 2: Outstate ~ Private + s(Room.Board, 4) + s(PhD, 4) + s(perc.alumni,
##     2) + s(Expend, 4) + Grad.Rate
## Model 3: Outstate ~ Private + s(Room.Board, 4) + s(PhD, 4) + s(perc.alumni,
##     2) + s(Expend, 4) + s(Grad.Rate, 4)
## Model 4: Outstate ~ Private + s(Room.Board, 4) + s(PhD, 4) + s(perc.alumni,
##     2) + s(Expend, 4) + s(Grad.Rate, 5)
##   Resid. Df Resid. Dev     Df Deviance       F    Pr(>F)
## 1       605 2185197963
## 2       604 2088077461 1.0000 97120502 28.3753 1.415e-07 ***
## 3       601 2059357222 3.0002 28720240  2.7968   0.03949 *
## 4       600 2053630774 0.9995  5726448  1.6739   0.19623
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The results provide strong evidence that a GAM which includes `Grad.Rate` as a non-linear function is needed (p=0.03939).