RSM8512 Assignment - Bias/Variances Tradeoff

Yanbing Chen 1009958752 2023/10/02

Question 1

For each of parts (a) through (d), indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.

(a) The sample size n is extremely large, and the number of predictors p is small.

Answer: A flexible method would generally be better. With a large sample size, flexible methods have ample data to learn from, reducing the risk of overfitting. Since the number of predictors is small, a flexible approach can adapt to more complex relationships between predictors and the response without overfitting.

(b) The number of predictors p is extremely large, and the number of observations n is small.

Answer: An inflexible method would generally be better. When the dimensionality (p) is much larger than the sample size (n), flexible methods run a high risk of overfitting. They could potentially fit to the noise rather than the true underlying pattern. Inflexible methods would be more conservative and less likely to overfit in such scenarios.

(c) The relationship between the predictors and response is highly non-linear.

Answer: A flexible method would generally be better. With more degrees of freedom, a flexible model would obtain better fit.

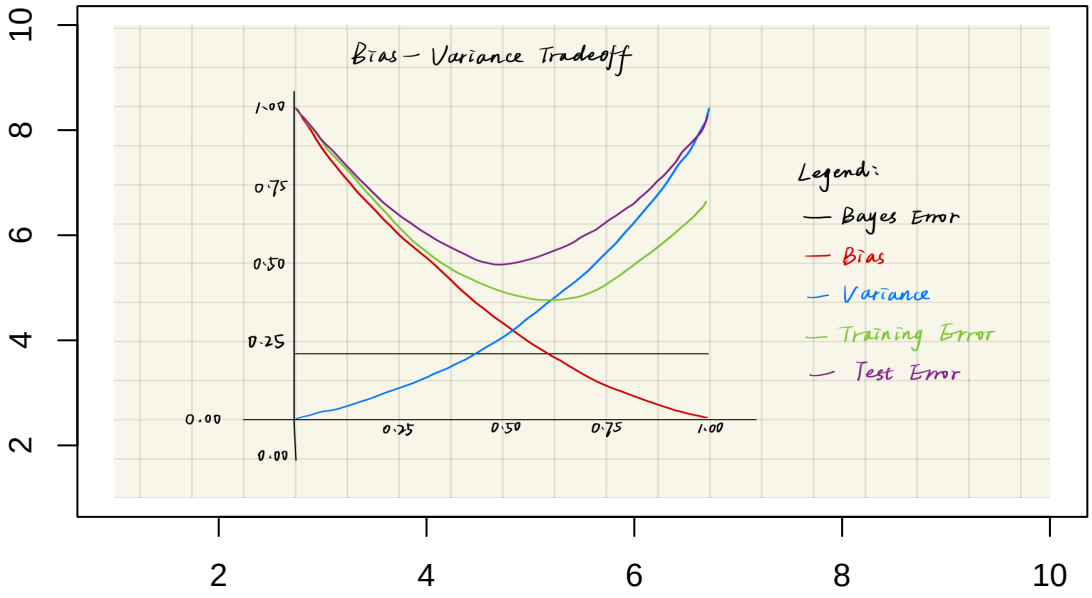
(d) The variance of the error terms, i.e. $\sigma^2 = \text{Var}(\epsilon)$, is extremely high.

Answer: An inflexible method would generally be better. High error variance means that there's a lot of noise in the relationship between predictors and the response. Flexible methods might adapt too closely to this noise, effectively overfitting to the random errors in the training data. An inflexible method, on the other hand, would not adapt as closely to the noise, making it more robust in the presence of high-variance error terms.

Question 3

We now revisit the bias-variance decomposition.

(a) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, and the y-axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(b) Explain why each of the five curves has the shape displayed in part (a).

Answer: (Squared) Bias: Inflexible models make strong assumptions about the shape of the true relationship between predictors and response. As flexibility increases, models make fewer assumptions, reducing bias.

Variance: Inflexible models don't change much with different training datasets. But as models become more flexible, they start to capture not just the true relationship but also the noise in the training data. This means their predictions can vary widely with different training sets, leading to increased variance.

Training Error: More flexible models will always fit the training data better, leading to a continuous decrease in training error as flexibility increases.

Test Error: The balance between bias and variance influences test error. Initially, reducing bias has a more significant effect on reducing test error. But after a certain point, increasing variance starts increasing the test error.

Bayes (or Irreducible) Error: This represents the error due to noise in the data. No matter how good our model is, we can't predict random noise, so this error remains constant.

Question 9

This exercise involves the Auto data set studied in the lab. Make sure that the missing values have been removed from the data.

(a) Which of the predictors are quantitative, and which are qualitative?

Answer:

Quantitative: mpg, displacement, horsepower, weight acceleration, year

Qualitative: name, cylinders, origin

(b) What is the range of each quantitative predictor? You can answer this using the range() function.

```
range1 = apply(Auto[,c(1,3:7)],2, range)
Qunatitative_predictor = c('mpg','displacement', 'horsepower','weight','acceleration','year')
Range = c('9.0-46.6','68-455','46-230','1613-5140','8.0-24.8','70-82')
des_tab<-cbind(Qunatitative_predictor,Range)
knitr::kable(des_tab, "pipe")
```

Qunatitative_predictor	Range
mpg	9.0-46.6
displacement	68-455
horsepower	46-230

Qunatitative_predictor	Range
weight	1613-5140
acceleration	8.0-24.8
year	70-82

(c) What is the mean and standard deviation of each quantitative predictor?

```
mean1 = apply(Auto[, c(1,3:7)], 2, mean)
sd1 = apply(Auto[, c(1,3:7)],2,sd)
Qunatitative_predictor = c('mpg','displacement', 'horsepower','weight','acceleration','year')
Mean = c(23.445918,194.411990,104.469388,2977.584184,15.541327,75.979592)
SD = c(7.8050075, 104.6440039,38.4911599,849.4025600,2.7588641,3.6837365)
des_tab<-cbind(Qunatitative_predictor, Mean, SD)
knitr::kable(des_tab, "pipe")
```

Qunatitative_predictor	Mean	SD
mpg	23.445918	7.8050075
displacement	194.41199	104.6440039
horsepower	104.469388	38.4911599
weight	2977.584184	849.40256
acceleration	15.541327	2.7588641
year	75.979592	3.6837365

(d) Now remove the 10th through 85th observations. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains?

```
df <- Auto[-(10:85), ] # remove the 10th through 85th observations
#dim(df)
range2 = apply(df[,c(1,3:7)],2,range)
mean2 = apply(df[,c(1,3:7)],2,mean)
sd2 = apply(df[,c(1,3:7)],2,sd)
```

```
Qunatitative_predictor = c('mpg','cylinders', 'displacement', 'horsepower','weight','accelerati
Range = c('11.0-46.0','68-455', '46-230', '1649-4997', '8.5-24.8','70-82')
Mean = c(24.404430,187.240506, 100.721519, 2935.971519, 15.726899, 77.145570)
SD = c(7.867283, 99.678367, 35.708853, 811.300208, 2.693721, 3.106217)
des_tab<-cbind(Qunatitative_predictor, Range, Mean, SD)
```

```
## Warning in cbind(Qunatitative_predictor, Range, Mean, SD): number of rows of
## result is not a multiple of vector length (arg 2)
```

```
knitr::kable(des_tab, "pipe")
```

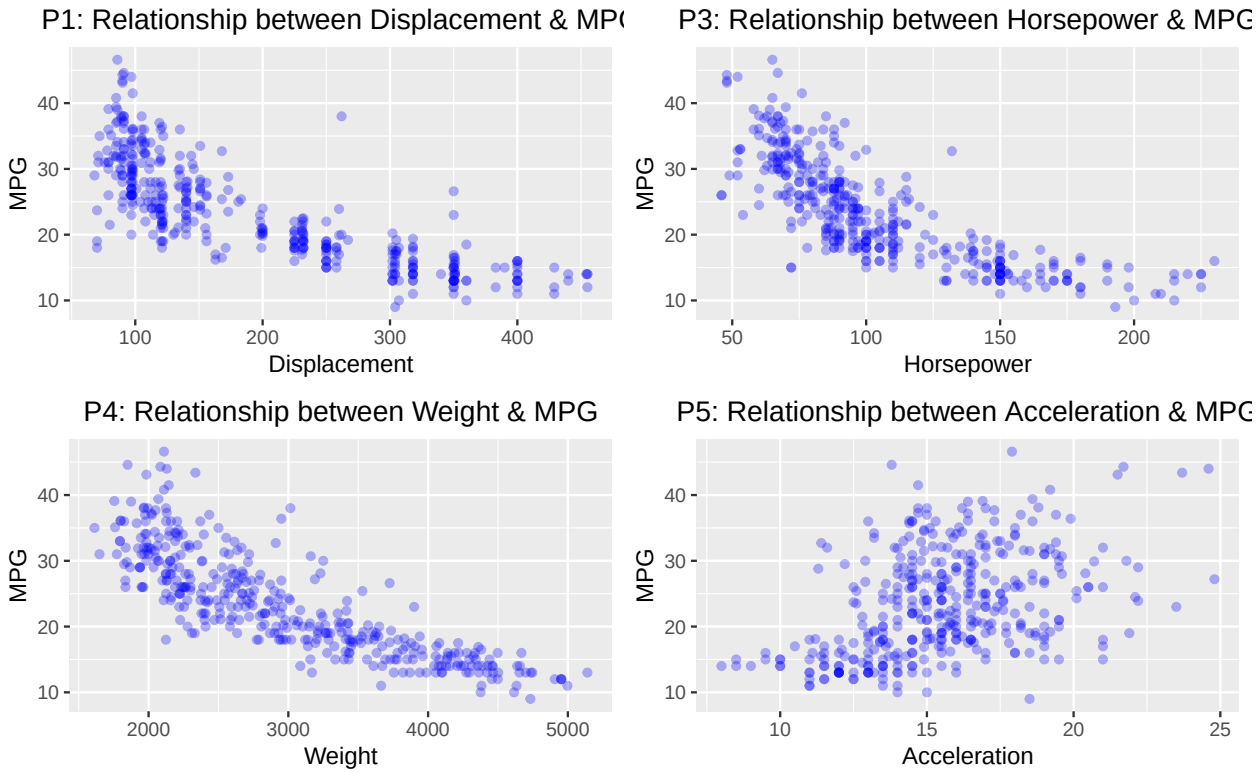
Qunatitative_predictor	Range	Mean	SD
mpg	11.0-46.0	24.40443	7.867283
cylinders	68-455	187.240506	99.678367
displacement	46-230	100.721519	35.708853
horsepower	1649-4997	2935.971519	811.300208
weight	8.5-24.8	15.726899	2.693721
acceleration	70-82	77.14557	3.106217
year	11.0-46.0	24.40443	7.867283

(e) Using the full data set, investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.

For this question, I choose to investigate the relationship between mpg and other quatitative variables.

```
#p1
p1 = ggplot(data = Auto,aes(x = displacement, y = mpg))+
  geom_point(color = 'blue', alpha = 0.3)+
  labs(x = 'Displacement', y = 'MPG', title = 'P1: Relationship between Displacement & MPG')+
  theme(plot.title = element_text(hjust = 0.5))
```

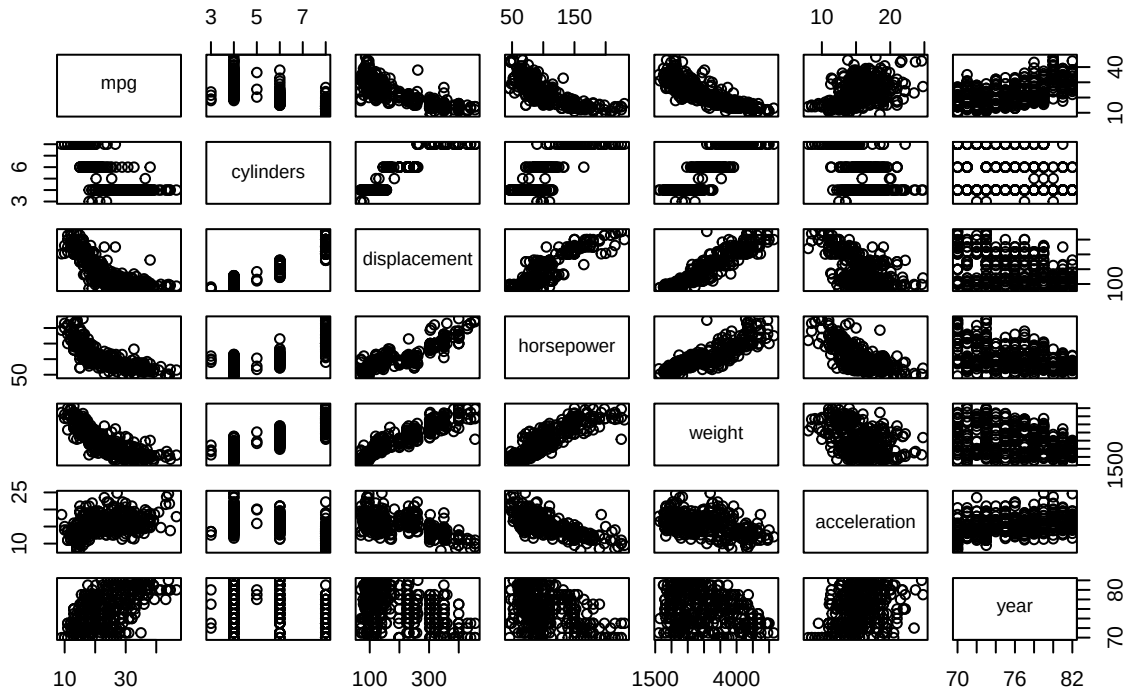
```r
#p2
p2 = ggplot(data = Auto,aes(x = horsepower, y = mpg))+
  geom_point(color = 'blue', alpha = 0.3)+
  labs(x = 'Horsepower', y = 'MPG', title = 'P3: Relationship between Horsepower & MPG')+
  theme(plot.title = element_text(hjust = 0.5))

#p3
p3 = ggplot(data = Auto,aes(x = weight, y = mpg))+
  geom_point(color = 'blue', alpha = 0.3)+
  labs(x = 'Weight', y = 'MPG', title = 'P4: Relationship between Weight & MPG')+
  theme(plot.title = element_text(hjust = 0.5))

#p4
p4 = ggplot(data = Auto,aes(x = acceleration, y = mpg))+
  geom_point(color = 'blue', alpha = 0.3)+
  labs(x = 'Acceleration', y = 'MPG', title = 'P5: Relationship between Acceleration & MPG')+
  theme(plot.title = element_text(hjust = 0.5))

grid.arrange(p1, p2, p3, p4, ncol=2)
```



Answer: For the predictors `displacement`, `horsepower` and `weight`, as these three quantities gradually increase, the value of `mpg` gradually decreases, suggesting that there may be a negative correlation between these three predictors and mpg. And it is not possible to see a significant linear relationship between `mpg` and `acceleration` from Picture 4, which may indicate that they have a low or no correlation.

(f) Suppose that we wish to predict gas mileage (mpg) on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting mpg? Justify your answer.

```
pairs(~mpg + cylinders + displacement + horsepower + weight + acceleration + year, Auto,main =
```




```
#pairs(Auto)
```

From the graphs I drew above, we can consider using the `displacement`, `horsepower` and `weight` as predictors to predict `mpg` because there is a clear correlation between them. However, for `acceleration` might not fit for prediction and we also need to explore other perspectives to prove our guess, like maybe predictors would influence each other and we need to eliminate this when building the model.

The correlation plot for other predictors also has been shown (only focus on the first row). I added `cylinders` and `year` into consideration and found there is no significant correlation between `mpg` and `cylinders`. And for another variable `year`, we cannot make any decisions solely based on this correlation plot. Although there are many data gather in the center in the plot, we cannot general any conclusion before deep dive into it.