**Chapter 6 Conceptual:**

1. **We perform best subset, forward stepwise, and backward stepwise selection on a single data set. For each approach, we obtain p + 1 models, containing 0, 1, 2..., p predictors. Explain your answers:**
2. **Which of the three models with k predictors has the smallest training RSS?**

The other two techniques determine models with a path dependency on the predictors they choose initially when they iterate to the k'th model, which results in best subset selection having the shortest training RSS.

1. **Which of the three models with k predictors has the smallest test RSS?**

Because it considers more models than the other techniques, best subset selection could have the smallest test RSS. The other models might have a higher chance of choosing a model that more closely matches the test data.

1. i) True, ii) True, iii) False, iv) False, iv) False.
2. **For parts (a) through (c), indicate which of i. through iv. is correct. Justify your answer.**
3. iii. Less flexible and better predictions because of less variance, more bias
4. iii. Same as Lasso
5. ii. More flexible, less bias, more variance

**Chapter 7 Conceptual:**

**4. Suppose we fit a curve with basis functions b1(X) = I (0 ≤ X ≤ 2) − (X −1)I(1 ≤ X ≤ 2), b2(X)=(X −3)I(3 ≤ X ≤ 4) +I(4 < X ≤ 5). We fit the linear regression model.**

**Y = β0 + β1b1(X) + β2b2(X) + ϵ,**

**and obtain coefficient estimates βˆ0 = 1, βˆ1 = 1, βˆ2 = 3. Sketch the estimated curve between X = −2 and X = 6. Note the intercepts, slopes, and other relevant information.**

We obtain the fitted model by plugging in these basis functions and coefficient estimates:

f^(X)=1+I(0≤X≤2)−(X−1)I(1≤X≤2)+3(X−3)I(3≤X≤4)+3I(4<X≤5)

We can simplify f^(X) throughout this range since we are only interested in the function over the range [2,2]:

f^(X)=1+I(0≤X≤2)−(X−1)I(1≤X≤2)f^(X)=1+I(0≤X≤2)−(X−1)I(1≤X≤2)

Graphical user interface, chart, line chart

Description automatically generated

For X<0X<0 none of the indicator variables are true, so f^(X)=1f^(X)=1 with a slope of 00.

For 0≤X<10≤X<1, the first indicator variable is true, so f^(X)=1+1=2f^(X)=1+1=2, so this is the intercept with a slope of 00.

For 1≤X≤21≤X≤2, both indicator variables are true, so f^(X)=1+1−(X−1)⋅1=3−Xf^(X)=1+1−(X−1)⋅1=3−X, so the slope is −1−1 here.

5.

a. Given that g^2 will be a higher order polynomial due to the order of the derivative penalty function, we would anticipate that g2 will have the smaller training RSS.

b. Because g^2 might be overfit with the additional degree of freedom, we would anticipate that g^1 would have the smaller test RSS.

c. when λ=0, g^1 = g^2.