(c) Repeat (a) for non-linear methods relative to least squares.

3. Suppose we estimate the regression coefficients in a linear regression model by minimizing

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p}|\beta_j| \leq s$$

for a particular value of s. For parts (a) through (e), indicate which of i. through v. is correct. Justify your answer.

(a) As we increase s from 0, the training RSS will:

i. Increase initially, and then eventually start decreasing in an inverted U shape.

ii. Decrease initially, and then eventually start increasing in a U shape.

iii. Steadily increase.

iv. Steadily decrease.

v. Remain constant.

(b) Repeat (a) for test RSS.

(c) Repeat (a) for variance.

(d) Repeat (a) for (squared) bias.

(e) Repeat (a) for the irreducible error.

Answer

a) The training RSS will steadily decrease.
As s increases, the constraints on βs will decrease, as a result of which the training RSS will keep decreasing (leading to an overfit on the training data).

b) The test RSS will decrease initially, and then eventually start increasing in a U shape. This is because the test RSS decreases till we reach an optimal s value, and then starts increasing if s is increased further since we overfit to the training data.

c) The variance steadily increases due to the bias-variance tradeoff. As s increases, the variance of the model also increases because the β values are less-constrained.

d) The squared bias steadily decreases due to the bias-variance tradeoff. As s increases, the bias of the model decreases since the variance increases.

e) The irreducible error remains constant since it is the characteristic of the data that cannot be explained by the 'best estimated model'.