STA 325: Homework 3

DUE: 11:59pm, October 29 (on Sakai)
COVERAGE: ISL Chapters 6.2, 7

1. **[35 points]** Let's dig deeper into the shrinkage behavior of ridge regression and Lasso. Consider the simple linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \cdots, n,$$

where β_0 and β_1 are model parameters. For simplicity, suppose the predictor x is standardized such that $\sum_{i=1}^n (x_i - \bar{x})^2 = 1$.

(a) **[5 points]** Recall the residual-sum-of-squares (RSS) criterion:

$$\text{RSS}(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 .$$

Show that the least-squares-estimators (LSE) for (β_0, β_1), which minimize $\text{RSS}(\beta_0, \beta_1)$, are given by:

$$\hat{\beta}_0^{\text{LS}} = \bar{y} - \hat{\beta}_1^{\text{LS}} \bar{x}, \quad \hat{\beta}_1^{\text{LS}} = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}.$$

Hint: Set the derivative of $\text{RSS}(\beta_0, \beta_1)$ *with respect to* β_0 *to zero, then solve for* β_0*. Plug this expression for* β_0 *into the derivative of* $\text{RSS}(\beta_0, \beta_1)$ *with respect to* β_1*, then set to zero and solve for* β_1*.*

Answer: Taking the derivative with respect to the parameters (β_0, β_1), we get:

$$-2\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) = 0, \quad -2\sum_{i=1}^n x_i (y_i - \beta_0 - \beta_1 x_i) = 0,$$

The LSE $(\hat{\beta}_0^{\text{LS}}, \hat{\beta}_1^{\text{LS}})$ should satisfy the above two equations.

For fixed β_1, solving the first equation for β_0 gives:

$$\beta_0 = \bar{y} - \beta_1 \bar{x}.$$

Plugging this expression for β_0 into the second equation, we get:

$$-2\sum_{i=1}^n x_i (y_i - (\bar{y} - \beta_1 \bar{x}) - \beta_1 x_i) = 0$$

$$\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y} - \beta_1 \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right) = 0$$

$$\beta_1 = \left(\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y} \right) / \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right).$$

But $\sum_{i=1}^n x_i^2 - n\bar{x}^2 = \sum_{i=1}^n (x_i - \bar{x})^2 = 1$, so $\beta_1 = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}$, which proves the claim.

(b) **[5 points]** Consider the ridge regression estimators $(\hat{\beta}^{\mathrm{R}}_{0,\lambda}, \hat{\beta}^{\mathrm{R}}_{1,\lambda})$, which minimize the following optimization problem:

$$\min_{\beta_0,\beta_1} \left\{ \mathrm{RSS}(\beta_0, \beta_1) + \lambda \beta_1^2 \right\}.$$

Show that:

$$\hat{\beta}^{\mathrm{R}}_{0,\lambda} = \bar{y} - \hat{\beta}^{\mathrm{R}}_{1,\lambda} \bar{x}, \quad \hat{\beta}^{\mathrm{R}}_{1,\lambda} = \frac{\hat{\beta}^{\mathrm{LS}}_1}{1+\lambda}.$$

Answer: Taking the derivative with respect to the parameters (β_0, β_1), we get:

$$-2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0, \quad -2\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) + 2\lambda\beta_1 = 0.$$

The ridge regression estimators $(\hat{\beta}^{\mathrm{R}}_{0,\lambda}, \hat{\beta}^{\mathrm{R}}_{1,\lambda})$ should satisfy the above two equations.

For fixed β_1, solving the first equation for β_0 gives:

$$\beta_0 = \bar{y} - \beta_1 \bar{x}.$$

Plugging this expression for β_0 into the second equation, we get:

$$-2\sum_{i=1}^{n} x_i(y_i - (\bar{y} - \beta_1\bar{x}) - \beta_1 x_i) + 2\lambda\beta_1 = 0$$

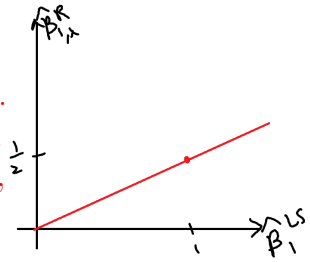
$$\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} - \beta_1 \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 + \lambda \right) = 0$$

$$\beta_1 = \left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} \right) / \left(\sum_{i=1}^{n} (x_i - \bar{x})^2 + \lambda \right) = \left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} \right) / (1+\lambda),$$

which proves the claim.

(c) **[5 points]** Suppose $\lambda = 1$. Plot the ridge regression estimator $\hat{\beta}^{\mathrm{R}}_{1,\lambda}$ (which is shrunk) as a function of the least-squares estimator $\hat{\beta}^{\mathrm{LS}}_1$, for $\hat{\beta}^{\mathrm{LS}}_1 \geq 0$. Comment on the shrinkage behavior of ridge regression. Does this plot give any insight on its ability to select important variables?

Answer: Here, the ridge regression estimator shrinks the least-squares estimator by a constant proportion of $1/(1+\lambda)$. Because of this proportional shrinkage, the ridge regression estimate can never shrink a coefficient estimate fully to zero, which explains its inability to perform variable selection.



(d) **[7 points]** Consider the Lasso estimators $(\hat{\beta}^{\mathrm{L}}_{0,\lambda}, \hat{\beta}^{\mathrm{L}}_{1,\lambda})$, which minimize the following

optimization problem:

$$\min_{\beta_0,\beta_1} \{\text{RSS}(\beta_0,\beta_1) + \lambda|\beta_1|\}\,.$$

Suppose $\hat{\beta}_1^{\text{LS}} \geq 0$. Show that:

$$\hat{\beta}_{0,\lambda}^{\text{L}} = \bar{y} - \hat{\beta}_{1,\lambda}^{\text{L}}\bar{x}, \quad \hat{\beta}_{1,\lambda}^{\text{L}} = (\hat{\beta}_1^{\text{LS}} - \lambda/2)_+ := \max\{\hat{\beta}_1^{\text{LS}} - \lambda/2, 0\}. \tag{1}$$

Hint: The key challenge here is that $|\beta_1|$ is not differentiable, so we need to generalize the notion of a derivative a bit. One can show that the Lasso estimators $(\hat{\beta}_{0,\lambda}^{\text{L}}, \hat{\beta}_{1,\lambda}^{\text{L}})$ solve the two equations:

$$\begin{aligned} -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) &= 0 \\ -2\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) + \lambda\,\partial\beta_1 &\ni 0, \end{aligned} \tag{2}$$

where $\partial\beta_1$ is the so-called subdifferential of $|\beta_1|$:

$$\partial\beta_1 = \begin{cases} -1, & \beta_1 < 0, \\ [-1,+1], & \beta_1 = 0, \\ +1, & \beta_1 > 0. \end{cases}$$

From (2), the Lasso estimator can be derived using the following two steps:

- *Suppose the least-squares estimate $\hat{\beta}_1^{\text{LS}} > \lambda/2$. What do the estimators in (1) simplify to? Do the simplified estimators solve (2)?*
- *Suppose the least-squares estimate $\hat{\beta}_1^{\text{LS}} \leq \lambda/2$. What do the estimators in (1) simplify to? Do the simplified estimators solve (2)?*

Answer: The first equation of (2) gives $\hat{\beta}_{0,\lambda}^{\text{L}} = \bar{y} - \hat{\beta}_{1,\lambda}^{\text{L}}\bar{x}$. So all that needs to be shown is that the Lasso slope estimate $\beta_1 = \hat{\beta}_{1,\lambda}^{\text{L}}$ satisfies the second equation of (2).

Consider the first case of $\hat{\beta}_1^{\text{LS}} > \lambda/2$. Here, the Lasso slope estimator simplifies to $\hat{\beta}_{1,\lambda}^{\text{L}} = (\hat{\beta}_1^{\text{LS}} - \lambda/2)_+ = \hat{\beta}_1^{\text{LS}} - \lambda/2$. Let's check that this choice of $\beta_1 = \hat{\beta}_{1,\lambda}^{\text{L}}$ indeed satisfies the second equation in (2):

$$-2\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) + \lambda\,\partial\beta_1 \ni 0$$

$$-2\sum_{i=1}^{n} x_i(y_i - \beta_0 - \hat{\beta}_1^{\text{L}} x_i) + \lambda \cdot 1 = 0 \qquad \text{(since } \beta_1 > 0\text{)}$$

$$-2\sum_{i=1}^{n} x_i(y_i - (\bar{y} - \hat{\beta}_1^{\text{L}}\bar{x}) - \hat{\beta}_1^{\text{L}} x_i) + \lambda \cdot 1 = 0 \qquad \text{(first equation of (2))}$$

$$\hat{\beta}_1^{\text{L}} = \hat{\beta}_1^{\text{LS}} - \lambda/2. \qquad \text{(solve for } \hat{\beta}_1^{\text{L}}\text{)}$$

Since the last statement is true, the second equation in (2) is satisfied.

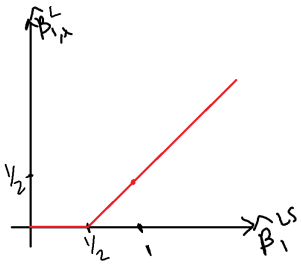
Consider the second case of $\hat{\beta}_1^{\text{LS}} \leq \lambda/2$. Then the Lasso slope estimator simplifies to $\hat{\beta}_{1,\lambda}^{\text{L}} = (\hat{\beta}_1^{\text{LS}} - \lambda/2)_+ = 0$. Let's check that this choice of $\beta_1 = \hat{\beta}_{1,\lambda}^{\text{L}}$ indeed satisfies the second equation in (2):

$$
\begin{aligned}
-2\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) + \lambda\, \partial\beta_1 \ni 0 & \\
-2\sum_{i=1}^{n} x_i(y_i - \beta_0 - 0 \cdot x_i) + \lambda \cdot [-1, +1] \ni 0 & \qquad \text{(since } \beta_1 = 0\text{)} \\
-2\sum_{i=1}^{n} x_i(y_i - (\bar{y} - 0 \cdot \bar{x})) + [-\lambda, \lambda] \ni 0 & \qquad \text{(first equation of (2))} \\
2\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right) \in [-\lambda, \lambda] & \qquad \text{(algebra)} \\
\hat{\beta}_1^{\text{LS}} \in [-\lambda/2, \lambda/2]. & \qquad \text{(algebra)}
\end{aligned}
$$

The last equation is true, since we assume $\hat{\beta}_1^{\text{LS}} \leq \lambda/2$. Hence, the second equation in (2) is satisfied. The claim is proven by combining both cases.

(e) **[5 points]** Suppose $\lambda = 1$. Plot the lasso estimator $\hat{\beta}_{1,\lambda}^{\text{L}}$ (which is shrunk) as a function of the least-squares estimator $\hat{\beta}_1^{\text{LS}}$, for $\hat{\beta}_1^{\text{LS}} \geq 0$. Comment on the shrinkage behavior of Lasso. Does this plot give any insight on its ability to select important variables?

Answer: Here, the Lasso estimator shrinks the least-squares estimator to zero if it's below the threshold of 1/2, and reduces the least-squares estimator by 1/2 if it exceeds the threshold of 1/2. You can think of this as shrinking the effect of a variable to zero if it's below a certain threshold, and retaining only a portion of its signal if it's above the threshold. In doing so, the Lasso estimate can shrink a coefficient estimate to zero (i.e., when it falls below the threshold), and can therefore perform variable selection.



(f) **[3 points]** Having used the squared-l_2 norm (part (b)) and the l_1-norm (part (d)), let's now try the so-called l_0-norm on β_1: $I(\beta_1 \neq 0)$. This new "norm" gives a value of 1 whenever β_1 is non-zero (i.e., the variable is active), and a value of 0 whenever β_1 equals zero (i.e., the variable is inert). Using this, the penalized regression problem becomes:

$$\min_{\beta_0, \beta_1} \left\{ \text{RSS}(\beta_0, \beta_1) + \lambda I(\beta_1 \neq 0) \right\}.$$

Reformulate this penalized problem into its constrained form with radius s (see Equations (6.8) or (6.9) in ISL). We've seen this constrained problem before for

variable selection. What is it? Explain.

Answer: Constrained problem reformulation:

$$\min_{\beta_0,\beta_1} \text{RSS}(\beta_0,\beta_1) \quad \text{s.t.} \quad I(\beta_1 \neq 0) \leq s.$$

This optimization is a special case of the best subset selection method in Chapter 5. With $s = 1$, we are solving for the model with lowest RSS, given that this model has at most one effect. With $s = 0$, we are solving for the model with lowest RSS, given that this model has no effects (this is just the null model).

(g) **[BONUS 5 points]:** Show that the estimators $(\hat{\beta}^{\text{S}}_{1,\lambda}, \hat{\beta}^{\text{S}}_{0,\lambda})$ which minimize the constrained problem in part (f) are given by:

$$\hat{\beta}^{\text{S}}_{0,\lambda} = \bar{y} - \hat{\beta}^{\text{S}}_{1,\lambda}\bar{x}, \quad \hat{\beta}^{\text{S}}_{1,\lambda} = \hat{\beta}^{\text{LS}}_1 \cdot I(\hat{\beta}^{\text{LS}}_1 > \sqrt{\lambda}).$$

The optimization problem here is:

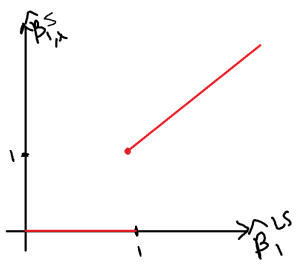
$$\min_{\beta_0,\beta_1} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 + \lambda I(\beta_1 \neq 0) \right\}$$

For any choice of β_1, the optimal choice of intercept β_0 is $\beta_0 = \bar{y} - \hat{\beta}_1\bar{x}$. Consider now the cost (in terms of the optimization criterion) of adding variable x to the model:

- If x *is* in the model, then the optimal RSS is $\sum_{i=1}^{n} \left(y_i - \hat{\beta}^{\text{LS}}_0 - \hat{\beta}^{\text{LS}}_1 x_i\right)^2$, with an additional penalty of λ.
- If x is *not* in the model, then the optimal RSS is $\sum_{i=1}^{n} (y_i - \bar{y})^2$ (which is higher than when x is in the model), but there is no additional penalty.
- Hence, the RSS reduction from adding x is $\sum_{i=1}^{n} (y_i - \bar{y})^2 - \sum_{i=1}^{n} \left(y_i - \hat{\beta}^{\text{LS}}_0 - \hat{\beta}^{\text{LS}}_1 x_i\right)^2 = (\hat{\beta}^{\text{LS}}_1)^2$, and the penalty increase in adding x is λ. From this, it follows that whenever $(\hat{\beta}^{\text{LS}}_1)^2 > \lambda$, or equivalently $\hat{\beta}^{\text{LS}}_1 > \sqrt{\lambda}$, it is more beneficial to add x to the model. Similarly, when $\hat{\beta}^{\text{LS}}_1 \leq \sqrt{\lambda}$, the null model (without x) yields a lower optimization criterion. This proves the claim.

(h) **[5 points]** Suppose $\lambda = 1$. Plot the estimator $\hat{\beta}^{\text{S}}_{1,\lambda}$ as a function of the least-squares estimator $\hat{\beta}^{\text{LS}}_1$, for $\hat{\beta}^{\text{LS}}_1 \geq 0$. Comment on the shrinkage behavior of this method. Does this plot give any insight on its ability to select important variables?

Answer: Here, the "best-subset" estimator shrinks the least-squares estimator to zero if it's below the threshold of 1, but keeps the least-squares estimator as-is if it exceeds the threshold of 1/2. You can think of this as shrinking the effect of a variable to zero if it's below a certain threshold, but keeping its full signal if it's above the threshold. In doing so, the "best-subset" estimate can shrink a coefficient estimate to zero (i.e., when it falls below the threshold), and can therefore perform variable selection.



2. [**21 points**] State whether each of the following statements are TRUE or FALSE. Briefly justify why in a couple of sentences.

   (a) Least-squares estimation should be used over ridge regression when there is high multi-collinearity in the data.
   Answer: FALSE. When there's multicollinearity, least-squares estimation can yield very high variance in terms of model fit. Ridge regression can stabilize this variance inflation by shrinking the estimates of $\boldsymbol{\beta}$.

   (b) Lasso should be used over ridge regression when we know a priori that only a small handful of predictors are active.
   Answer: TRUE. Given that the true model is sparse, the bias-variance trade-off for Lasso would likely yield a better predictive model (since it selects sparse models), whereas the bias-variance trade-off for ridge regression would likely yield a poorer predictive model (since it doesn't perform any selection). For inference, Lasso can also select the underlying active predictors, whereas ridge regression cannot.

   (c) Piecewise polynomial models can be discontinuous without constraints.
   Answer: TRUE. See top-left plot of Figure 7.3 in ISL.

   (d) For cubic splines, the variance of the fitted model decreases as more knots are added.
   Answer: FALSE. A cubic spline with more knots has higher degrees-of-freedom, which then results in higher flexibility (and higher variance) for its model fit.

   (e) Splines provide greater model flexibility in regions with many knots.
   Answer: TRUE. Since different polynomials are fit between two neighboring knots, regions with many knots allow for greater model flexibility.

   (f) A model with high degrees-of-freedom implies a greater bias in its fit.
   Answer: FALSE. A model with high degrees-of-freedom has greater flexibility and thereby lower bias in its model fit, at the cost of high variance.

   (g) Generalized additive models can be much more computationally expensive to fit compared to multiple linear regression.
   Answer: FALSE. A generalized additive model can be massaged into a linear modeling framework (with appropriately chosen basis functions), so it requires the same computation time as multiple linear regression (with the same number of d.f.s).

3. **[22 points]** Consider a *quartic spline* model with distinct knots ξ_k, $k = 1, \cdots, K$. A quartic spline satisfies two properties: (i) it is a quartic (i.e., degree-4) polynomial between any two neighboring knots, and (ii) it has continuous derivatives of up to order 3 at each knot. Note that property (ii) includes derivatives of order 0, meaning the quartic spline should be continuous at knots.

   (a) **[5 points]** Write out the full model specification for the quartic spline, including model parameters and basis functions (see Equation (7.9) in ISL). How many degrees-of-freedom (d.f.s) are in your model?
   Answer: The model can be written as:

   $$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 + \beta_5 b_5(x_i) + \cdots + \beta_{K+4} b_{K+4}(x_i) + \epsilon_i, \quad i = 1, \cdots, n,$$

   where the basis functions $b_5, \cdots, b_{K+4}$ are defined as:

   $$b_{k+4}(x) = (x - \xi_k)_+^4, \quad k = 1, \cdots, K.$$

   There are $K + 5$ d.f.s in the model.

   (b) **[7 points]** Prove that properties (i) and (ii) hold for your model in (a).

   Answer: Let's show (i) and (ii) separately:

   - Consider first property (i). Before the first knot, the model is $\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4$, which is clearly quartic. Before the second knot, the model also must be quartic, since it's the sum of two quartic functions. Extending this logic over all knots, one can then show that property (i) holds.
   - Consider next property (ii). Let $f(\cdot)$ be the mean function in part (a). We wish to show that $f(\cdot)$, $f'(\cdot)$, $f''(\cdot)$ and $f'''(\cdot)$ are continuous at every knot ξ. Without loss of generality, let's consider just the first knot ξ_1.
     - *Order 0*:

     $$\begin{aligned} \lim_{x \to \xi_1^+} f(x) &= \beta_0 + \beta_1 \xi_1 + \beta_2 \xi_1^2 + \beta_3 \xi_1^3 + \beta_4 \xi_1^4 + \lim_{x \to \xi_1^+} (x - \xi_1)^4 \\ &= \beta_0 + \beta_1 \xi_1 + \beta_2 \xi_1^2 + \beta_3 \xi_1^3 + \beta_4 \xi_1^4 \\ &= \lim_{x \to \xi_1^-} f(x) \end{aligned}$$

     - *Order 1*:

     $$\begin{aligned} \lim_{x \to \xi_1^+} f'(x) &= \beta_1 + 2\beta_2 \xi_1 + 3\beta_3 \xi_1^2 + 4\beta_4 \xi_1^3 + \lim_{x \to \xi_1^+} 4(x - \xi_1)^3 \\ &= \beta_1 + 2\beta_2 \xi_1 + 3\beta_3 \xi_1^2 + 4\beta_4 \xi_1^3 \\ &= \lim_{x \to \xi_1^-} f'(x) \end{aligned}$$

– *Order 2*:

$$\begin{aligned}\lim_{x\to\xi_1^+} f''(x) &= 2\beta_2 + 6\beta_3\xi_1 + 12\beta_4\xi_1^2 + \lim_{x\to\xi_1^+} 12(x-\xi_1)^2 \\ &= 2\beta_2 + 6\beta_3\xi_1 + 12\beta_4\xi_1^2 \\ &= \lim_{x\to\xi_1^-} f''(x)\end{aligned}$$

– *Order 3*:

$$\begin{aligned}\lim_{x\to\xi_1^+} f'''(x) &= 6\beta_3 + 24\beta_4\xi_1 + \lim_{x\to\xi_1^+} 24(x-\xi_1) \\ &= 6\beta_3 + 24\beta_4\xi_1 \\ &= \lim_{x\to\xi_1^-} f'''(x)\end{aligned}$$

I wrote the above in gory detail, but full marks as long as you show that the right-limits of the derivatives of $(x-\xi_1)_+^4$ zeros out at $x=\xi_1$.

(c) **[5 points]** Suppose you present this model to your boss. Her initial reaction was that, while she likes the flexibility of your model, she is afraid that this comes at a huge computational cost. She is worried that model fitting (e.g., estimation, prediction, computing confidence intervals) will be too time-consuming for large datasets. Because of this, she suggests you try a simpler linear model instead, which can be fit efficiently. Should you agree with her? Explain why or why not.
Answer: No, you should not. From part (a), a quartic spline can be equivalently represented as a linear model with an appropriate choice of basis functions. Because of this, model estimation, prediction, and confidence intervals can be easily computed by applying the same techniques as in linear regression.

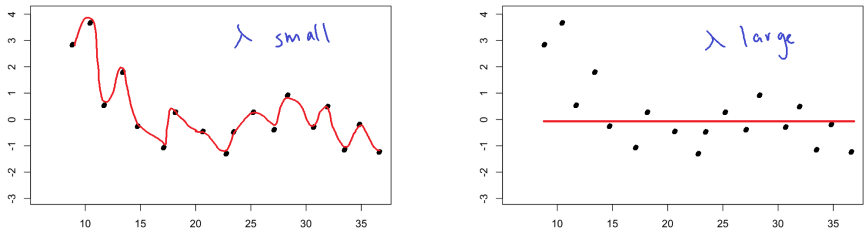
(d) **[5 points]** Suppose, after some discussion, she begrudgingly adopts your quartic spline model. After seeing your R output, however, she complains that your fit requires 16 d.f.s, which she believes to be too many. She claims that, with that many d.f.s, a degree-15 polynomial model can be fit, which can capture higher order effects than your quartic model. Should you agree with her? Explain why or why not.
Answer: No, you should not. Even though a degree-15 polynomial model accounts for higher-order polynomial effects, your quartic spline model would be a more flexible model, since it allows different (lower-order) polynomials to be fit over the prediction domain. The true regression function f is often better approximated by lower-order polynomials locally, rather than a single high-order polynomial over the full domain. The latter typically results in very poor prediction near the boundaries.

4. **[30 points]** Consider the following generalization of the smoothing spline:

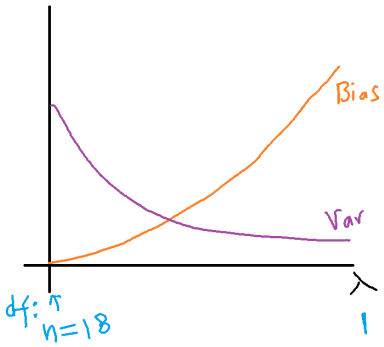
$$\hat{g} = \arg\min_g \left\{ \sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int [g^{(m)}(x)]^2 dx \right\},$$

where $\hat{g}$ is the fitted predictor, and $g^{(m)}$ is the m-th derivative of g.

(a) **[8 points]** Suppose $m = 1$. For the data below, draw out what the fitted predictor $\hat{g}$ looks like for both λ small and λ large. For both plots, identify and justify important features of the predictor. For the same data, draw out what the bias and variance curves may look like as a function of λ, and label the d.f.s corresponding to the left and right endpoints of the x-axis. Identify and justify important features of these curves.

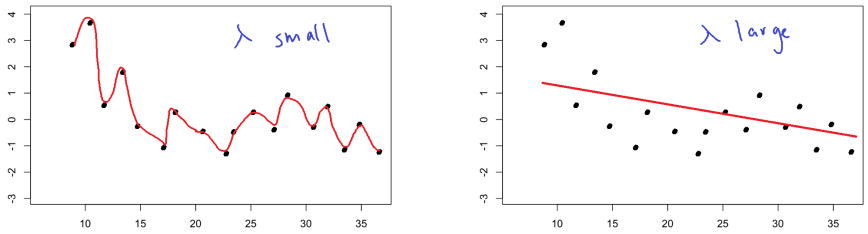


Answer: The fitted curve $\hat{g}$ with λ small should connect the dots, whereas the fitted curve with λ large should be the constant function equaling the sample mean $\bar{y}$. This is because, with λ large, the only functions g with $g'(x) = 0$ everywhere are constant functions.

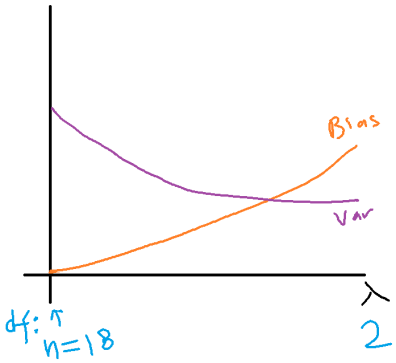


For $\lambda = 0$, the bias should be very close to zero, and the variance should be very large, since the fitted function $\hat{g}$ interpolates the data points. For λ large, the bias should be very large, since (from the data) the true regression function appears to be very different from a constant function. The variance, however, should be very small, since the intercept model only has one model parameter to be estimated.

(b) **[10 points]** Suppose $m = 2$. Perform the same analysis as in part (a), and comment on any differences between the bias and variance curves here from the curves in part (a).

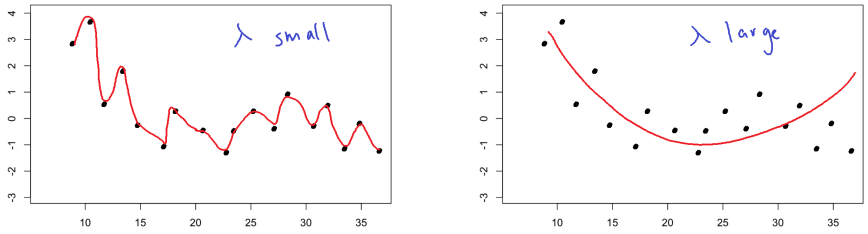


Answer: The fitted curve $\hat{g}$ with λ small should connect the dots, whereas the fitted curve with λ large should be the line-of-best-fit. This is because, with λ large, the only functions g with $g''(x) = 0$ everywhere are linear functions.



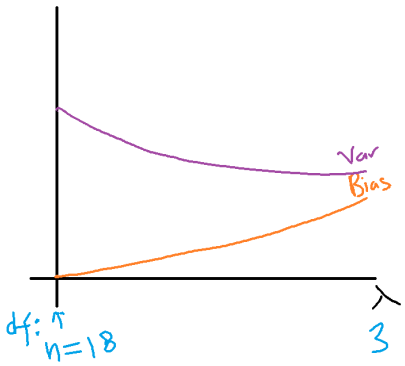
For $\lambda = 0$, the bias should be very close to zero, and the variance should be very large, since the fitted function $\hat{g}$ interpolates the data points. For λ large, the bias should be large (but smaller than that in part (a)), since the true regression function appears to be different from a linear function. The variance, however, should be small (but larger than that in part (a)), since a line only has two model parameters (slope and intercept) to be estimated.

(c) **[10 points]** Suppose $m = 3$. Perform the same analysis as in part (a), and comment on any differences between the bias and variance curves here from the curves in parts (a) and (b).



Answer: The fitted curve $\hat{g}$ with λ small should connect the dots, whereas the fitted curve with λ large should be the best-fitting quadratic curve. This is because, with λ large, the only functions g with $g'''(x) = 0$ everywhere are quadratic functions.

For $\lambda = 0$, the bias should be very close to zero, and the variance should be very large, since the fitted function $\hat{g}$ interpolates the data points. For λ large, the bias



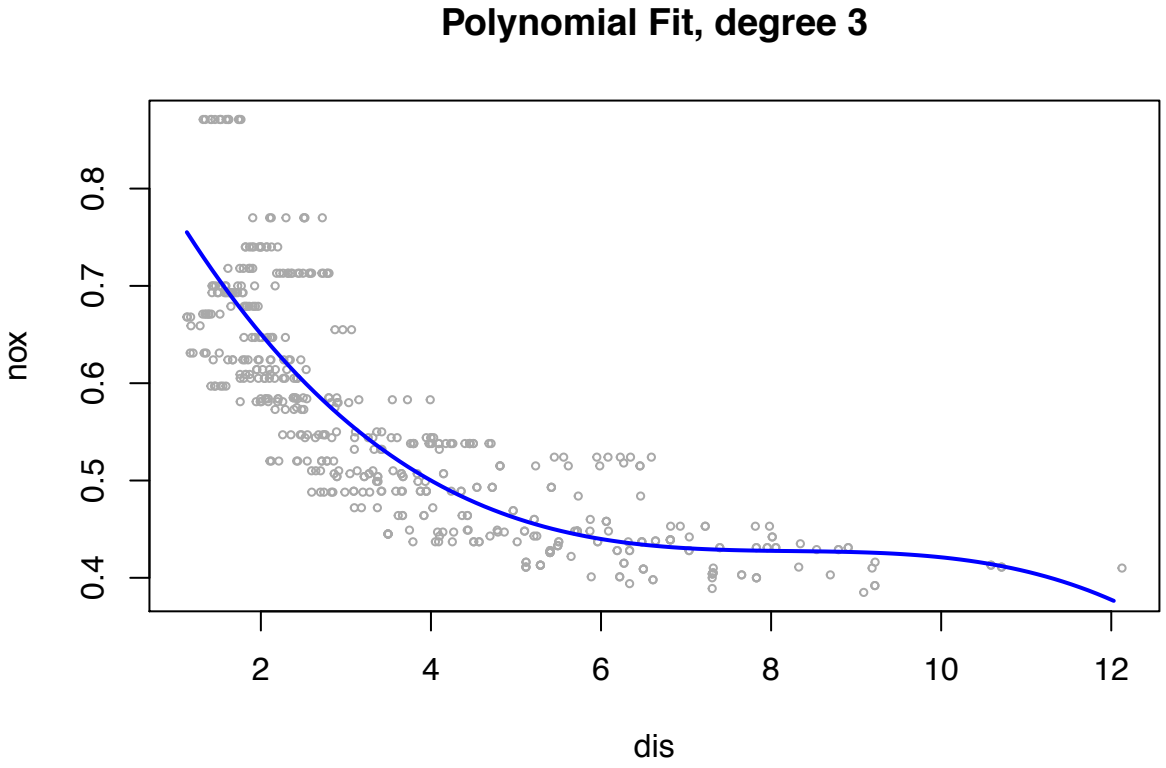
should be moderate (and smaller than that in parts (a) and (b)), since the true regression function appears to be slightly different from a quadratic function. The variance should also be moderate (and larger than that in parts (a) and (b)), since a quadratic function has three model parameters to be estimated.

(d) **[2 points]** Given data, how should the derivative order m be chosen in practice? **Answer**: Perform cross-validation on several choices of order m, and choose the one yielding the lowest estimated test error.

```
library(MASS)
library(boot)
library(splines)
data(Boston)
set.seed(1)
```

5(a)

```
model.fit <- lm(nox ~ poly(dis, 3), data=Boston)
dislims <- range(Boston$dis)
grid <- seq(dislims[1], dislims[2], 0.1)
preds <- predict(model.fit, newdata=list(dis=grid), se=TRUE)
plot(Boston$dis, Boston$nox, xlim=dislims, cex=0.5, col="darkgrey", xlab="dis", ylab="nox")
title("Polynomial Fit, degree 3")
lines(grid, preds$fit, lwd=2, col="blue")
```




```
summary(model.fit)
```

```
##
## Call:
## lm(formula = nox ~ poly(dis, 3), data = Boston)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.121130 -0.040619 -0.009738  0.023385  0.194904
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
```

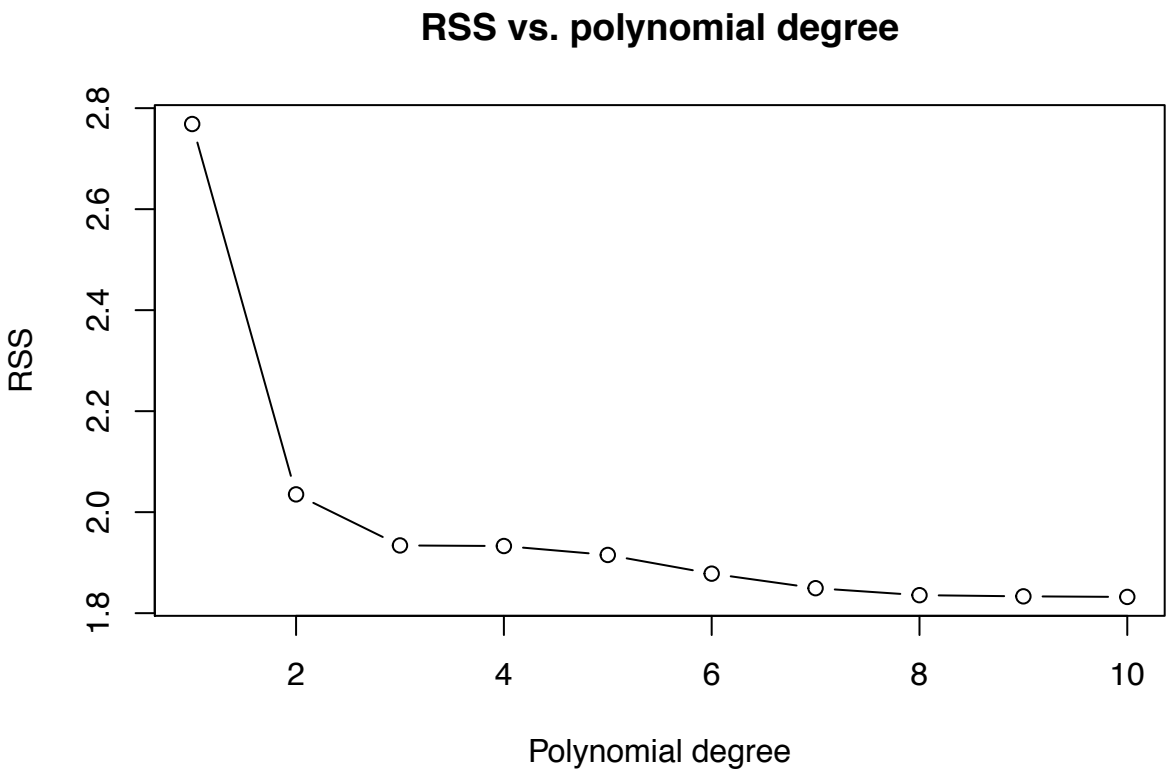
```
## (Intercept)    0.554695   0.002759 201.021  < 2e-16 ***
## poly(dis, 3)1 -2.003096   0.062071 -32.271  < 2e-16 ***
## poly(dis, 3)2  0.856330   0.062071  13.796  < 2e-16 ***
## poly(dis, 3)3 -0.318049   0.062071  -5.124 4.27e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06207 on 502 degrees of freedom
## Multiple R-squared:  0.7148, Adjusted R-squared:  0.7131
## F-statistic: 419.3 on 3 and 502 DF,  p-value: < 2.2e-16
```

5(b)

```
RSS <- rep(0,10)
for (i in 1:10) {
  model.fit <- lm(nox ~ poly(dis,i), data=Boston)
  RSS[i] <- sum(model.fit$residuals^2)
}
RSS
```

```
##  [1] 2.768563 2.035262 1.934107 1.932981 1.915290 1.878257 1.849484
##  [8] 1.835630 1.833331 1.832171
```

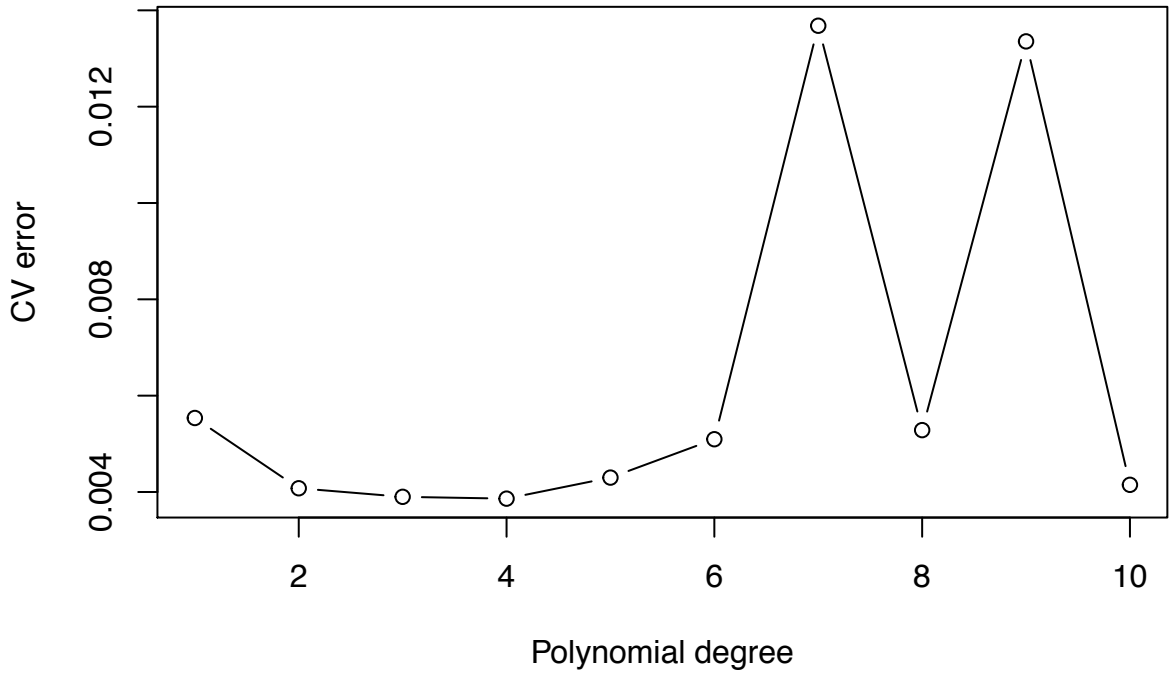
```
plot(RSS, type='b', xlab='Polynomial degree')
title("RSS vs. polynomial degree")
```



RSS decreases as degree of polynomial increases.

5(c)

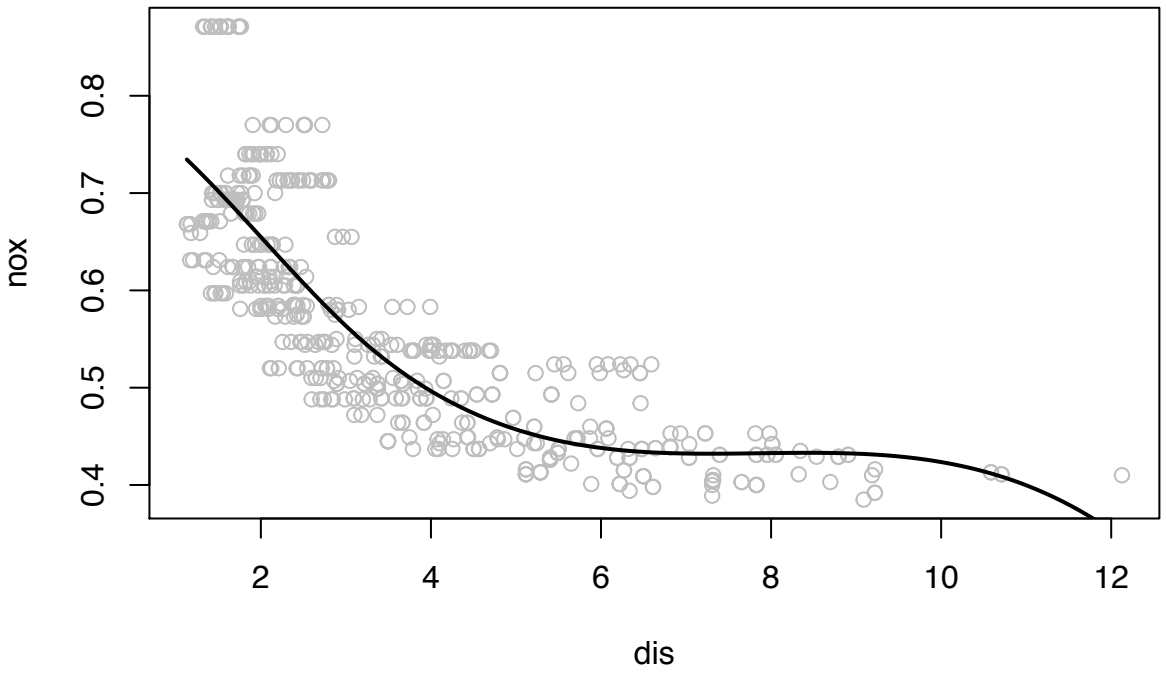
```
cv.error <- rep(0,10)
for (i in 1:10) {
  model.fit <- glm(nox ~ poly(dis,i), data=Boston)
  cv.error[i] <- cv.glm(Boston, model.fit, K=10)$delta[1]
}

plot(cv.error, type="b", ylab="CV error", xlab="Polynomial degree")
```



The CV results indicate that the optimal degree is around 4, since that produces the lowest CV error.

5(d)

```
model.fit <- lm(nox ~ bs(dis, df=4), data=Boston)
pred <- predict(model.fit, newdata=list(dis=grid), se=T)

plot(Boston$dis, Boston$nox, col="gray", ylab="nox", xlab="dis")
lines(grid, pred$fit, lwd=2)
```



Note that with a cubic spline, there is only 1 knot, which is automatically placed at the relevant quantiles (which in this case is at the 50th percentile).

```
attr(bs(Boston$dis,df=4),"knots")
```

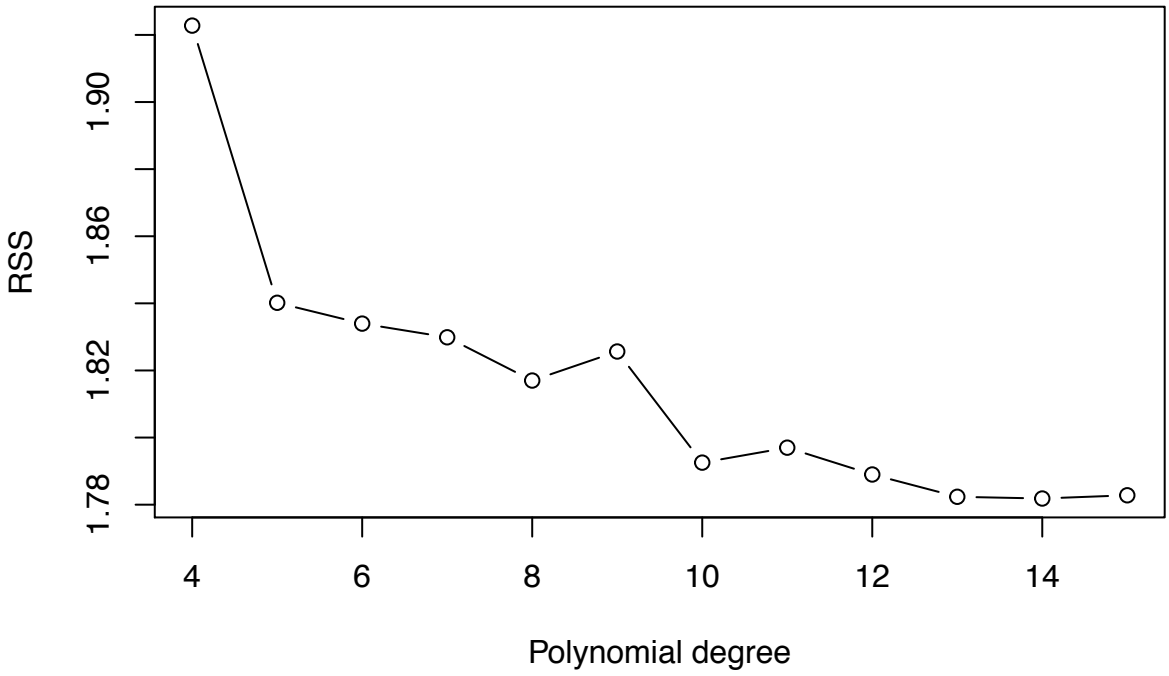
```
##     50%
## 3.20745
```

```
summary(model.fit)
```

```
##
## Call:
## lm(formula = nox ~ bs(dis, df = 4), data = Boston)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.124622 -0.039259 -0.008514  0.020850  0.193891
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        0.73447    0.01460  50.306  < 2e-16 ***
## bs(dis, df = 4)1 -0.05810    0.02186  -2.658  0.00812 **
## bs(dis, df = 4)2 -0.46356    0.02366 -19.596  < 2e-16 ***
## bs(dis, df = 4)3 -0.19979    0.04311  -4.634 4.58e-06 ***
## bs(dis, df = 4)4 -0.38881    0.04551  -8.544  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06195 on 501 degrees of freedom
## Multiple R-squared:  0.7164, Adjusted R-squared:  0.7142
## F-statistic: 316.5 on 4 and 501 DF,  p-value: < 2.2e-16
```

5(e)

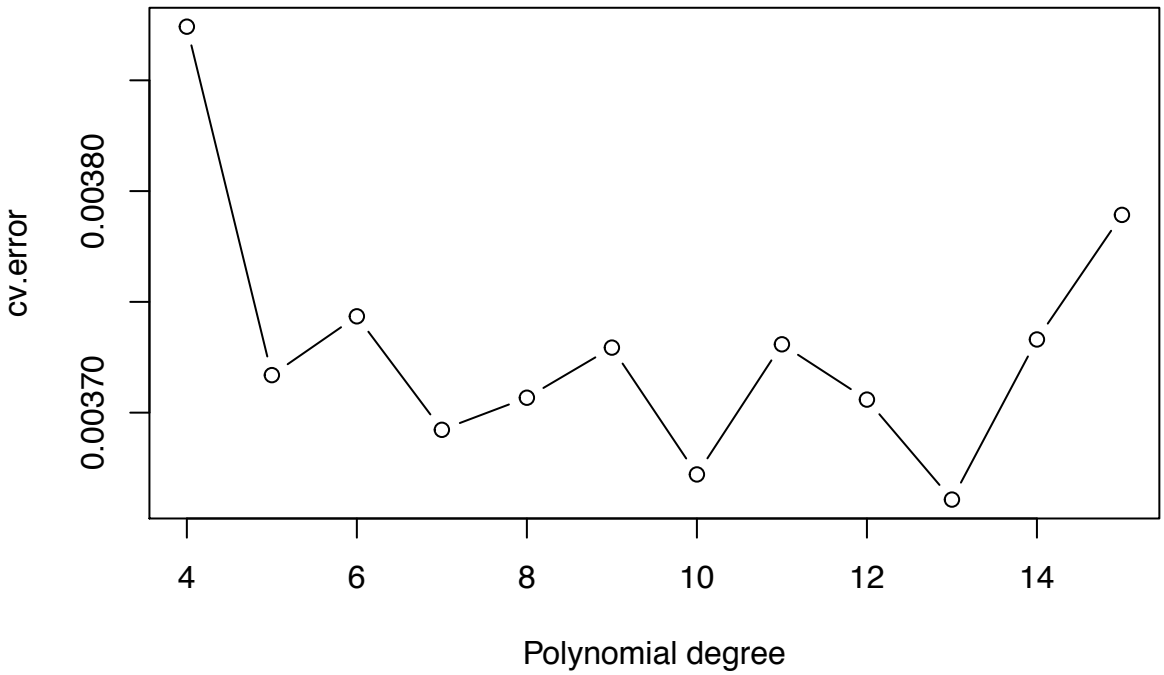
```
RSS <- rep(0,12)
for (i in 4:15) {
  model.fit <- lm(nox ~ bs(dis, df=i), data=Boston)
  RSS[i-3] <- sum(model.fit$residuals^2)
}


plot(4:15, RSS, type='b', xlab="Polynomial degree")
```



The RSS decreases until it flattens out around degree 13.

5(f)

```
cv.error <- rep(0,12)
set.seed(1)
for (i in 4:15) {
  model.fit <- glm(nox ~ bs(dis, df=i), data=Boston)
  cv.error[i-3] <- cv.glm(Boston, model.fit, K=10)$delta[1]
}


plot(4:15, cv.error, type='b', xlab='Polynomial degree')
```



The results are a bit erratic/jumpy, but it seems like choosing a degree of 10 achieves a pretty low CV error.

6(a)

```
library(leaps)
library(ISLR)
library(gam)
```

```
## Loading required package: foreach
```

```
## Loaded gam 1.16.1
```

```
data(College)
set.seed(1)
```

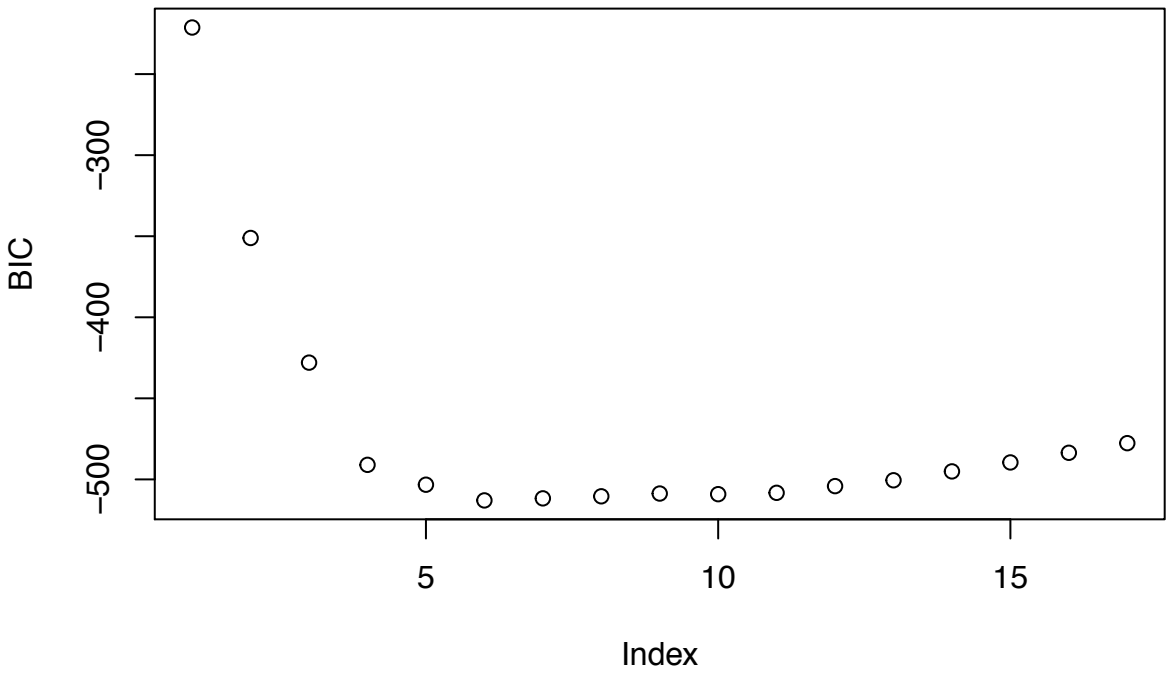
```
# Make training and test sets
train.idx <- sample(1:nrow(College), nrow(College)/2)
train <- College[train.idx,]
test <- College[-train.idx,]


# forward selection
model.fwd <- regsubsets(Outstate ~ ., data=train, nvmax=ncol(College)-1, method = "forward")
summary(model.fwd)
```

```
## Subset selection object
## Call: regsubsets.formula(Outstate ~ ., data = train, nvmax = ncol(College) -
##     1, method = "forward")
## 17 Variables  (and intercept)
##             Forced in Forced out
## PrivateYes      FALSE      FALSE
## Apps            FALSE      FALSE
```

```
## Accept          FALSE      FALSE
## Enroll          FALSE      FALSE
## Top10perc       FALSE      FALSE
## Top25perc       FALSE      FALSE
## F.Undergrad     FALSE      FALSE
## P.Undergrad     FALSE      FALSE
## Room.Board      FALSE      FALSE
## Books           FALSE      FALSE
## Personal        FALSE      FALSE
## PhD             FALSE      FALSE
## Terminal        FALSE      FALSE
## S.F.Ratio       FALSE      FALSE
## perc.alumni     FALSE      FALSE
## Expend          FALSE      FALSE
## Grad.Rate       FALSE      FALSE
## 1 subsets of each size up to 17
## Selection Algorithm: forward
##           PrivateYes Apps Accept Enroll Top10perc Top25perc F.Undergrad
## 1  ( 1 )  " "        " "  " "    " "    " "        " "        " "
## 2  ( 1 )  "*"        " "  " "    " "    " "        " "        " "
## 3  ( 1 )  "*"        " "  " "    " "    " "        " "        " "
## 4  ( 1 )  "*"        " "  " "    " "    " "        " "        " "
## 5  ( 1 )  "*"        " "  " "    " "    " "        " "        " "
## 6  ( 1 )  "*"        " "  " "    " "    " "        " "        " "
## 7  ( 1 )  "*"        " "  " "    " "    " "        " "        " "
## 8  ( 1 )  "*"        " "  " "    " "    " "        " "        " "
## 9  ( 1 )  "*"        " "  "*"    " "    " "        " "        " "
## 10  ( 1 ) "*"        "*"  "*"    " "    " "        " "        " "
## 11  ( 1 ) "*"        "*"  "*"    " "    " "        " "        "*"
## 12  ( 1 ) "*"        "*"  "*"    " "    " "        " "        "*"
## 13  ( 1 ) "*"        "*"  "*"    " "    "*"        " "        "*"
## 14  ( 1 ) "*"        "*"  "*"    " "    "*"        " "        "*"
## 15  ( 1 ) "*"        "*"  "*"    " "    "*"        "*"        "*"
## 16  ( 1 ) "*"        "*"  "*"    "*"    "*"        "*"        "*"
## 17  ( 1 ) "*"        "*"  "*"    "*"    "*"        "*"        "*"
##           P.Undergrad Room.Board Books Personal PhD Terminal S.F.Ratio
## 1  ( 1 )  " "          " "        " "   " "      " " " "      " "
## 2  ( 1 )  " "          " "        " "   " "      " " " "      " "
## 3  ( 1 )  " "          "*"        " "   " "      " " " "      " "
## 4  ( 1 )  " "          "*"        " "   " "      " " " "      " "
## 5  ( 1 )  " "          "*"        " "   " "      " " "*"      " "
## 6  ( 1 )  " "          "*"        " "   " "      " " "*"      " "
## 7  ( 1 )  " "          "*"        " "   "*"      " " "*"      " "
## 8  ( 1 )  " "          "*"        " "   "*"      " " "*"      "*"
## 9  ( 1 )  " "          "*"        " "   "*"      " " "*"      "*"
## 10  ( 1 ) " "          "*"        " "   "*"      " " "*"      "*"
## 11  ( 1 ) " "          "*"        " "   "*"      " " "*"      "*"
## 12  ( 1 ) "*"          "*"        " "   "*"      " " "*"      "*"
## 13  ( 1 ) "*"          "*"        " "   "*"      " " "*"      "*"
## 14  ( 1 ) "*"          "*"        " "   "*"      "*" "*"      "*"
## 15  ( 1 ) "*"          "*"        " "   "*"      "*" "*"      "*"
## 16  ( 1 ) "*"          "*"        " "   "*"      "*" "*"      "*"
## 17  ( 1 ) "*"          "*"        "*"   "*"      "*" "*"      "*"
##           perc.alumni Expend Grad.Rate
```

```
## 1  ( 1 )  " "         "*"    " "
## 2  ( 1 )  " "         "*"    " "
## 3  ( 1 )  " "         "*"    " "
## 4  ( 1 )  "*"         "*"    " "
## 5  ( 1 )  "*"         "*"    " "
## 6  ( 1 )  "*"         "*"    "*"
## 7  ( 1 )  "*"         "*"    "*"
## 8  ( 1 )  "*"         "*"    "*"
## 9  ( 1 )  "*"         "*"    "*"
## 10  ( 1 ) "*"         "*"    "*"
## 11  ( 1 ) "*"         "*"    "*"
## 12  ( 1 ) "*"         "*"    "*"
## 13  ( 1 ) "*"         "*"    "*"
## 14  ( 1 ) "*"         "*"    "*"
## 15  ( 1 ) "*"         "*"    "*"
## 16  ( 1 ) "*"         "*"    "*"
## 17  ( 1 ) "*"         "*"    "*"
```

```r
# Plot BIC results
plot(summary(model.fwd)$bic, ylab="BIC")
```



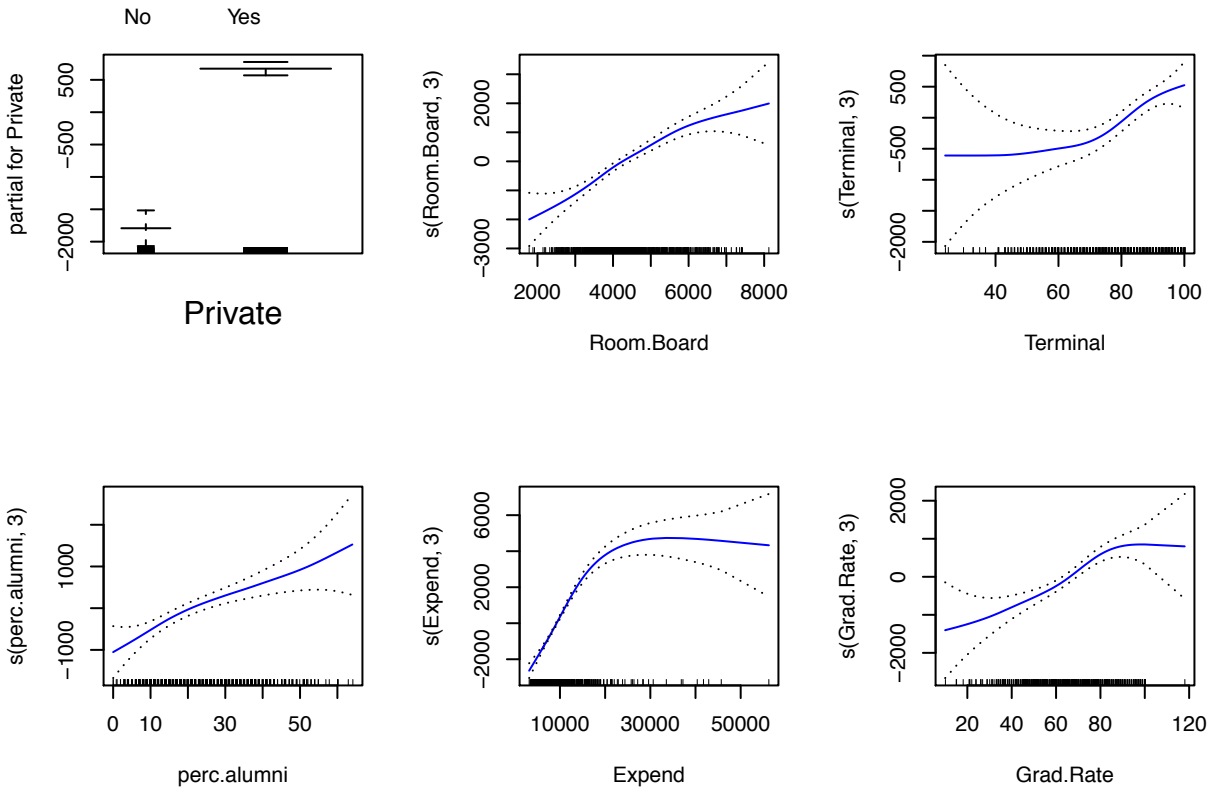
Based on this plot, it seems like performance is best with around 6 predictors:

```r
coef(model.fwd, 6)
```

```
##   (Intercept)    PrivateYes    Room.Board      Terminal   perc.alumni
## -4241.4402916  2790.4303173     0.9629335    37.8412517    60.6406044
##        Expend     Grad.Rate
##     0.2149396    30.3831268
```

6(b)

```
model.gam <- gam(Outstate ~
                   Private + # don't fit spline on categorical variables
                   s(Room.Board,3) +
                   s(Terminal,3) +
                   s(perc.alumni,3) +
                   s(Expend,3) +
                   s(Grad.Rate,3),
                 data=College)
par(mfrow=c(2,3))
plot(model.gam, se=TRUE, col="blue")
```



We see that some of the predictors have more/less linear relationships with the outcome variable. One of the benefits of GAMs is that we are able to make these type of plots, and interpret each variable independently of the others.

6(c)

```
pred <- predict(model.gam, test) # predict on test set

MSE <- mean((test$Outstate - pred)^2)
MSE
```

```
## [1] 3587099
```

We can also examine the RMSE, which is on the same scale as the original outcome variable.

```r
sqrt(MSE)
```

```
## [1] 1893.964
```

6(d)

```r
summary(model.gam)
```

```
##
## Call: gam(formula = Outstate ~ Private + s(Room.Board, 3) + s(Terminal,
##     3) + s(perc.alumni, 3) + s(Expend, 3) + s(Grad.Rate, 3),
##     data = College)
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -7110.16 -1137.02    50.44  1285.38  8278.86
##
## (Dispersion Parameter for gaussian family taken to be 3520187)
##
##     Null Deviance: 12559297426 on 776 degrees of freedom
## Residual Deviance: 2675342725 on 760.0001 degrees of freedom
## AIC: 13936.36
##
## Number of Local Scoring Iterations: 2
##
## Anova for Parametric Effects
##                    Df     Sum Sq    Mean Sq F value    Pr(>F)
## Private             1 3366732308 3366732308 956.407 < 2.2e-16 ***
## s(Room.Board, 3)    1 2549088628 2549088628 724.134 < 2.2e-16 ***
## s(Terminal, 3)      1  802254341  802254341 227.901 < 2.2e-16 ***
## s(perc.alumni, 3)   1  525154274  525154274 149.184 < 2.2e-16 ***
## s(Expend, 3)        1 1022010841 1022010841 290.329 < 2.2e-16 ***
## s(Grad.Rate, 3)     1  151344060  151344060  42.993 1.014e-10 ***
## Residuals         760 2675342725    3520187
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Anova for Nonparametric Effects
##                   Npar Df Npar F   Pr(F)
## (Intercept)
## Private
## s(Room.Board, 3)        2  2.591 0.07557 .
## s(Terminal, 3)          2  2.558 0.07815 .
## s(perc.alumni, 3)       2  0.835 0.43446
## s(Expend, 3)            2 56.179 < 2e-16 ***
## s(Grad.Rate, 3)         2  3.363 0.03515 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

According to the p-values of the F-statistic, there is strong evidence for non-linear effects for Expend and Grad.Rate (at the 0.05 level), and also some evidence for non-linear effects for Room.Board and Terminal (at the 0.1 level).