Ken Ye
jy294

STA 325
HW 3

#1

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

$$\frac{\partial RSS(\beta_0, \beta_1)}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{\beta}_0 - \sum_{i=1}^{n} \hat{\beta}_1 x_i = 0$$

$$n\bar{y} - n\hat{\beta}_0 - n\hat{\beta}_1 \bar{x} = 0$$

$$\boxed{\hat{\beta}_0^{LS} = \bar{y} - \hat{\beta}_1^{LS} \bar{x}}$$

$$\frac{\partial RSS(\beta_0, \beta_1)}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \hat{\beta}_0 - \sum_{i=1}^{n} \hat{\beta}_1 x_i^2 = 0$$

$$\sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 n\bar{x} - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0 \quad (\text{sub } \hat{\beta}_0)$$

$$\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} + \hat{\beta}_1 n\bar{x}^2 - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0$$

$$\hat{\beta}_1 \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

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

$$\boxed{\hat{\beta}_1^{LS} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}$$

Scanned with CamScanner

#1

b) $\frac{\partial\left(\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2+\lambda\beta_1^2\right)}{\partial\beta_0} = -2\sum_{i=1}^{n}(y_i-\hat\beta_0-\hat\beta_1 x_i)=0$

$\sum_{i=1}^{n} y_i-\sum_{i=1}^{n}\hat\beta_0-\sum_{i=1}^{n}\hat\beta_1 x_i=0$

similar to part a) $\Rightarrow$

$n\bar y-n\hat\beta_0-n\hat\beta_1\bar x=0$

$$\boxed{\hat\beta_{0,\lambda}^{R}=\bar y-\hat\beta_{1,\lambda}^{R}\bar x}$$

$\frac{\partial\left(\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2+\lambda\beta_1^2\right)}{\partial\beta_1} = -2\sum_{i=1}^{n}x_i(y_i-\hat\beta_0-\hat\beta_1 x_i)+2\lambda\hat\beta_1=0$

$\sum_{i=1}^{n}x_i y_i-\hat\beta_0 n\bar x-\hat\beta_1\sum x_i^2-\lambda\hat\beta_1=0$ (sub $\hat\beta_0$)

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

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

$-n\bar x^2+\sum_{i=1}^{n}x_i^2=\sum_{i=1}^{n}(x_i-\bar x)^2=1$

$\hat\beta_1^{LS}=\sum x_i y_i-n\bar x\bar y$

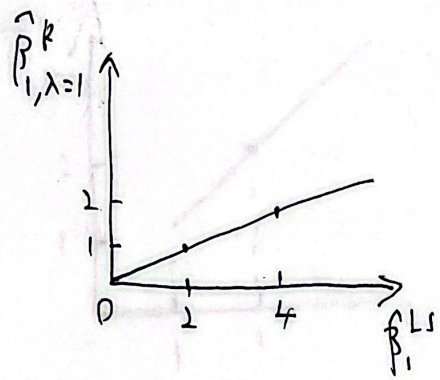
$(1+\lambda)\hat\beta_1=\hat\beta_1^{LS}$

$$\boxed{\hat\beta_{1,\lambda}^{R}=\frac{\hat\beta_1^{LS}}{1+\lambda}}$$

Scanned with CamScanner

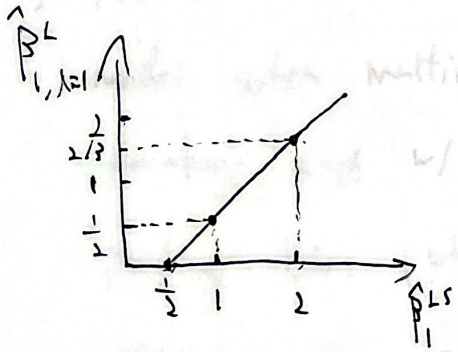
#4)

c) $\lambda = 1$, $\hat{\beta}^R_{1,\lambda} = \frac{\hat{\beta}^{LS}_1}{2}$



There is a linear relationship b/w $\hat{\beta}^R_{1,\lambda}$ and $\hat{\beta}^{LS}_1$. Since $\hat{\beta}^R_{1,\lambda=1}$ does not equal to 0 unless $\hat{\beta}^{LS}_1$ equal to 0, it doesn't have the ability to select important variables.

e) $\lambda = 1$, $\hat{\beta}^L_{1,\lambda} = (\hat{\beta}^{LS}_1 - \frac{1}{2})_+$



$\hat{\beta}^L_{1,\lambda=1}$ can perform variable selection. predictors with $\hat{\beta}^{LS}_1 < \frac{1}{2}$ will be considered inactive & shrunk to 0 when $\hat{\beta}^L_{1,\lambda}$ is used.

f) $\min_{\beta_0,\beta_1} \{RSS(\beta_0,\beta_1) + \lambda I(\beta_i \neq 0)\}$

⇓

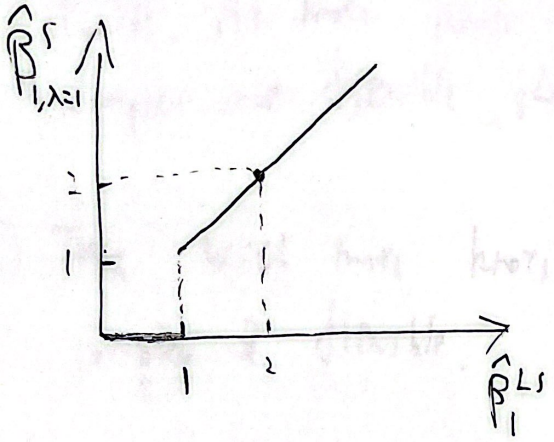
minimize $\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$, subject to $I(\beta_i \neq 0) \leq S$

When $S=1$, this is equivalent of minimizing RSS, subject to $\beta_i \neq 0$. Best subset selection finds model w/ smallest RSS, here we are doing sth. similar to best subset selection, while in most cases β_i may be 0.

Scanned with CamScanner

#1

h) $\lambda=1,\ \hat{\beta}^{S}_{1,\lambda} = \hat{\beta}^{LS}_{1} \cdot I\left(\hat{\beta}^{LS}_{1} \geq 1\right)$



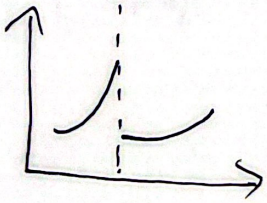
$\hat{\beta}^{S}_{1,\lambda=1}$ can perform variable selection as it's shrunk to 0 when $\hat{\beta}^{LS}_{1} < 1$. Variables w/ $\hat{\beta}^{LS}_{1} < 1$ is considered inactive.

#2

a) False. LS estimation produces a high-variance and unbiased model when multicollinearity exists. Ridge regression performs shrinkage, and w/ an optimal λ value chosen by CV, trading bias (which becomes higher) for variance (which becomes lower), a smaller MSPE could be achieved.

b) True. Lasso can perform variable selection, but not Ridge. Therefore, Lasso would result in a smaller model with fewer active predictors.

c) True. Without constraints, piecewise polynomial may not be continuous at breaking points (knots), like the graph below. Constraints can be imposed to ensure continuity if desired.



Scanned with CamScanner

#2

d) False. As more knots are added, the model becomes more complex and flexible, thus has higher variance.

e) True. With more knots, spline models can be more complex & flexible.

f) False. A model with more degrees of freedom has more effective variables that can be adjusted to fit the data, so the model can be more complex & flexible. This leads to higher variance and lower bias.

#3

a) A quartic spline with k knots can be modeled as

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \beta_3 b_3(x_i) + \dots + \beta_{k+4} b_{4}(x_i) + \varepsilon_i$$

where $b_1(x_i) = x_i$

$b_2(x_i) = x_i^2$

$b_3(x_i) = x_i^3$

$b_4(x_i) = x_i^4$

$b_{k+4}(x_i) = (x_i - \xi)_+^4 = \begin{cases} (x-\xi)^4 & \text{if } x > \xi \\ 0 & \text{otherwise} \end{cases}$ for $k = 1, \dots, K$

1 intercept + 4 basis functions + k knots

$\Downarrow$

k+5 d.f.s

Scanned with CamScanner

#3

b) property (i): it is quartic polynomial b/w any two neighboring knots.

① Before first knot: $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 + \varepsilon_i$

② At m^{th} knot: $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 + \ldots + \beta_{m+4}(x_i - \xi_m)^4 + \varepsilon_i$
$1 < m < k$

③ After last knot: $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 + \ldots + \beta_{k+4}(x_i - \xi_k)^4 + \varepsilon_i$
(k^{th})

After the intercept and the basis functions, all polynomials being added with the addition of knot $k = 1 \ldots k$ are polynomial of degree 4.

property (ii): continuous derivatives of up to order 3 at each knot.

Suppose $x_i > \xi_{k=0}$

Left

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4$$

$$\frac{dy_i}{dx_i} = \beta_1 + 2\beta_2 x_i + 3\beta_3 x_i^2 + 4\beta_4 x_i^3$$

$$\frac{d^2 y_i}{dx_i^2} = 2\beta_2 + 6\beta_3 x_i + 12\beta_4 x_i^2$$

$$\frac{d^3 y_i}{dx_i^3} = 6\beta_3 + 24\beta_4 x_i$$

$$\frac{d^4 y_i}{dx_i^4} = 24\beta_4$$

Right

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 + \underline{\beta_5 (x_i - \xi_1)^4}$$

$$\frac{dy_i}{dx_i} = \beta_1 + 2\beta_2 x_i + 3\beta_3 x_i^2 + 4\beta_4 x_i^3 + \underline{4\beta_5 (x_i - \xi_1)^3}$$

$$\frac{d^2 y_i}{dx_i^2} = 2\beta_2 + 6\beta_3 x_i + 12\beta_4 x_i^2 + \underline{12\beta_5 (x_i - \xi_1)^2}$$

$$\frac{d^3 y_i}{dx_i^3} = 6\beta_3 + 24\beta_4 x_i + \underline{24\beta_5 (x_i - \xi_1)}$$

$$\frac{d^4 y_i}{dx_i^4} = 24\beta_4 + 24\beta_5$$

✓ These become 0 when approach from right.

Therefore, the 1st, 2nd, and 3rd derivative from left & right side equal to each other, thus continuous. However, the 4th derivatives are different.

Scanned with CamScanner

#3

c) I disagree. Quartic splines, as formulated in part a), can be efficiently computed just like linear regression. We can also easily compute coefficient estimation, standard errors, confidence intervals, hypothesis testings. Fundamentally, it can be represented as a linear model. Therefore, quartic splines can be fit as efficiently as a single linear model.

d) I disagree.

First, higher-order polynomial (such as degree 15) model is unstable and can perform poorly at the boundaries (Runge's phenomenon), whereas quartic spline is more stable.

Second, splines often give a better predictive performance over polynomial models because the true regression function is typically never exactly a k^{th} order polynomial over the whole domain. And splines give a more flexible model that allow for rapid changes in certain regions, but not in others. In fact, true regression function f can often be well-approximated by low-order polynomials in local regions.

Scanned with CamScanner