BIOS663 Homework 1 Solution

1.

$$|C - \lambda I| = \begin{vmatrix} 2-\lambda & 3 \\ 3 & 5-\lambda \end{vmatrix} = \lambda^2 - 7\lambda + 1 = 0$$

Therefore, eigenvalue

$$\lambda_{1,2} = \frac{7 \pm 3\sqrt{5}}{2}$$

2. (a)

$$E[x_1 + 2x_2 + 4x_3] = E[x_1] + 2E[x_2] + 4E[x_3] = 0$$

$$var(x_1 + 2x_2 + 4x_3) = \begin{pmatrix} 1 & 2 & 4 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0.5 \\ 0 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 4 \end{pmatrix} = 33$$

Then the distribution of $x_1 + 2x_2 + 4x_3$ is $N(0, 33)$.

(b)

$$E\left[\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \middle| x_3 = 2 \right] = \begin{pmatrix} 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix} 1^{-1}(2 - 0) = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$var\left\{ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \middle| x_3 = 2 \right\} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix} 1^{-1} \begin{pmatrix} 0.5 & 0.5 \end{pmatrix} = \begin{pmatrix} 0.75 & -0.25 \\ -0.25 & 0.75 \end{pmatrix}$$

Then the distribution for $(x_1, x_2 | x_3 = 2)$ is $N\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 0.75 & -0.25 \\ -0.25 & 0.75 \end{pmatrix} \right)$.

3. (a) The ridge regression has solution satisfies

$$\frac{\partial SSE_{ridge}}{\beta} = \frac{\partial(\boldsymbol{y'y - 2y'X\beta + \beta'X'X\beta} + \lambda\boldsymbol{\beta'\beta})}{\boldsymbol{\beta}} = -2\boldsymbol{X'y} + 2\boldsymbol{X'X\beta} + 2\lambda\boldsymbol{\beta} = 0$$

$$\Rightarrow \hat{\boldsymbol{\beta}}_R = (\boldsymbol{X'X} + \lambda \boldsymbol{I_{p \times p}})^{-1} \boldsymbol{X'y}$$

(b)

$$E\left[\hat{\boldsymbol{\beta}}_R\right] = (\boldsymbol{X'X} + \lambda \boldsymbol{I_{p\times p}})^{-1}\boldsymbol{X'}E(\boldsymbol{y}) = (\boldsymbol{X'X} + \lambda \boldsymbol{I_{p\times p}})^{-1}\boldsymbol{X'X}\boldsymbol{\beta}$$

If λ is not 0, $\hat{\boldsymbol{\beta}}_R$ is not an unbiased estimator of β.

(c)

$$\begin{aligned} Cov\left(\hat{\boldsymbol{\beta}}_R\right) =& (\boldsymbol{X'X} + \lambda \boldsymbol{I_{p\times p}})^{-1}\boldsymbol{X'}Cov(\boldsymbol{y})\boldsymbol{X}(\boldsymbol{X'X} + \lambda \boldsymbol{I_{p\times p}})^{-1} \\ =& \sigma^2(\boldsymbol{X'X} + \lambda \boldsymbol{I_{p\times p}})^{-1}\boldsymbol{X'X}(\boldsymbol{X'X} + \lambda \boldsymbol{I_{p\times p}})^{-1} \end{aligned}$$

(d) Ridge regression has the effect of shrinking the estimates towards zero. It introduces bias to the parameter estimates but reduces the variance. If $\boldsymbol{X}$ is not full rank, $\boldsymbol{X'X}$ is not invertible and there is no unique solution for the least squares regression. However there's always unique solution for ridge regression.

4 To minimize the sum of square errors, we have

$$\frac{\partial SSE}{\boldsymbol{\beta}} = \frac{\sum_{i=1}^n \partial(y_i - \beta_0 - \beta_1 x_{i1})^2}{\partial \boldsymbol{\beta}} = \begin{pmatrix} -2\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1}) \\ -2\sum_{i=1}^n x_{i1}(y_i - \beta_0 - \beta_1 x_{i1}) \end{pmatrix} = 0.$$

From the first equation we have

$$\hat{\beta}_0 = \frac{1}{n}\left(\sum_{i=1}^n y_i - \hat{\beta}_1 \sum_{i=1}^n x_{i1}\right)$$

Plug it in the second equation, we have

$$\begin{aligned} &\sum_{i=1}^n x_{i1}y_i - \frac{1}{n}\left(\sum_{i=1}^n y_i - \hat{\beta}_1 \sum_{i=1}^n x_{i1}\right)\sum_{i=1}^n x_{i1} - \hat{\beta}_1 \sum_{i=1}^n x_{i1}^2 = 0 \\ &\Rightarrow \hat{\beta}_1 \left(\sum_{i=1}^n x_{i1}^2 - \left(\sum_{i=1}^n x_{i1}\right)^2\right) = \sum_{i=1}^n x_{i1}y_i - \frac{1}{n}\sum_{i=1}^n y_i \sum_{i=1}^n x_{i1} \\ &\Rightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_{i1} - \bar{x}_1)}{\sum_{i=1}^n)(x_{i1} - \bar{x}_1)^2} \end{aligned}$$

$$
\begin{aligned}
E[\hat{\beta}_1] &= E\left[\frac{\sum_{i=1}^{n}(y_i-\bar{y})(x_{i1}-\bar{x}_1)}{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2}\right] \\
&= \sum_{i=1}^{n}\frac{E\left[y_i-\bar{y}\right](x_{i1}-\bar{x}_1)}{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2} \\
&= \sum_{i=1}^{n}\frac{[(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2})-(\beta_0+\beta_1\bar{x}_1+\beta_2\bar{x}_2)](x_{i1}-\bar{x}_1)}{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2} \\
&= \sum_{i=1}^{n}\frac{[(\beta_1(x_{i1}-\bar{x}_1)+\beta_2(x_{i2}-\bar{x}_2)](x_{i1}-\bar{x}_1)}{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2} \\
&= \beta_1 + \frac{\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)(x_{i1}-\bar{x}_1)}{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2}\beta_2 \\
\Rightarrow E[\hat{\beta}_1]-\beta_1 &= \frac{\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)(x_{i1}-\bar{x}_1)}{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2}\beta_2
\end{aligned}
$$

5. (a)

$$
\theta_1 = \begin{pmatrix}\beta_0+\beta_1\\ \beta_0-\beta_2\end{pmatrix} = \begin{pmatrix}1 & 1 & 0\\ 1 & 0 & -1\end{pmatrix}\begin{pmatrix}\beta_0\\ \beta_1\\ \beta_2\end{pmatrix} = \boldsymbol{C}_1\boldsymbol{\beta}
$$

Assume we could find $\boldsymbol{T} = \begin{pmatrix}t_{11} & \dots & t_{15}\\ t_{21} & \dots & t_{25}\end{pmatrix}$ that satisfies

$$
\boldsymbol{C}_1 = \begin{pmatrix}1 & 1 & 0\\ 1 & 0 & -1\end{pmatrix} = \boldsymbol{T}\boldsymbol{X} = \begin{pmatrix}\sum_{j=1}^{5}t_{1j} & t_{11}+t_{12} & -\sum_{j=3}^{5}t_{1j}\\ \sum_{j=1}^{5}t_{2j} & t_{21}+t_{22} & -\sum_{j=3}^{5}t_{2j}\end{pmatrix}
$$

We could come up a solution (there are infinite number of solutions) that $\boldsymbol{T} = \begin{pmatrix}1 & 0 & 0 & 0 & 0\\ 0 & 0 & 1 & 0 & 0\end{pmatrix}$, so $\boldsymbol{\theta}_1$ is estimable.

(b)

$$
\theta_2 = \beta_2 = \begin{pmatrix}0 & 0 & 1\end{pmatrix}\begin{pmatrix}\beta_0\\ \beta_1\\ \beta_2\end{pmatrix} = \boldsymbol{C}_2\boldsymbol{\beta}
$$

Assume we could find $\boldsymbol{T} = \begin{pmatrix}t_{11} & \dots & t_{15}\end{pmatrix}$ that satisfies

$$
\boldsymbol{C}_2 = \begin{pmatrix}0 & 0 & 1\end{pmatrix} = \boldsymbol{T}\boldsymbol{X} = \begin{pmatrix}\sum_{j=1}^{5}t_{1j} & t_{11}+t_{12} & -\sum_{j=3}^{5}t_{1j}\end{pmatrix}
$$

We have $\sum_{j=1}^{2} t_{1j} = 0$ and $\sum_{j=3}^{5} t_{1j} = -1$, then consequently we have $\sum_{j=1}^{5} t_{1j} = -1$, which contradict the equation that $\sum_{j=1}^{5} t_{1j} = 0$. Thus we cannot find a $\boldsymbol{T}$ such that $\boldsymbol{C}_2 = \boldsymbol{T}\boldsymbol{X}$, so $\boldsymbol{\theta}_2$ is not estimable.

6. (a) The regression model could be specified as

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, \ldots, n,$$

where Y_i is the number of high ozone days and X_i is the meteorological index for observation i. ϵ_i's are the error terms, and are identically and independentally distributed. Using SAS IML, we can fit a linear regression model. The intercept β_0 has parameter estimate -193.0, with standard error 163.5, indicating that the average number of high ozone days when the meterological index is zero is -193.0. This result seems unreasonable since the number of days can never be negative. To properly interpret the intercept, instead we fit a linear regression model with centered seasonal average temperature. The intercept estimate for the new centered temperature model is 72.3 with standard error 5.95, indicating the average number of high ozone day with seasonal meteorological index 17.34 (the mean of sampled meteorological indices) is 72.3.

β_1 has parameter estimate 15.3, with standard error 9.42, which are the same in the two models, indicating that with one degree increase in seasonal average temperature, the number of high ozone days will averagely increase 15.3.

(b) All of the βs are estimable, because $\boldsymbol{X}$ is full rank.

(c) The hypothesis is $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$. We could form a F test, and the test statistic $F = 2.64 \sim F_{1,14}$ under null hypothesis. The p-value is 0.1267, so we accept the null hypothesis that number of high ozone days may not be associated with seasonal meteorological index.

(d) The hypothesis is $H_0 : \beta_1 = 12$ vs. $H_1 : \beta_1 \neq 12$. We could form a F test, and the test statistic $F = 0.1224 \sim F_{1,14}$ under null hypothesis. The p-value is $0.7316 > \alpha = 0.05$. So we accept the null hypothesis that 1 degree increase in average temperature is associated with a 12 day increase in the number of days the ozone levels exceed 0.20 ppm.