BIOS663 Homework 2 Solution

1. (a) The normal equation could be written as

$$\begin{pmatrix} \boldsymbol{X}_1'\boldsymbol{X}_1 & \boldsymbol{X}_1'\boldsymbol{X}_2 \\ \boldsymbol{X}_2'\boldsymbol{X}_1 & \boldsymbol{X}_2'\boldsymbol{X}_2 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_1'\boldsymbol{y} \\ \boldsymbol{X}_2'\boldsymbol{y} \end{pmatrix}.$$

Solve the first equation, we have

$$\beta_1 = \left(\boldsymbol{X}_1'\boldsymbol{X}_1\right)^{-1}\left(\boldsymbol{X}_1'\boldsymbol{y} - \boldsymbol{X}_1'\boldsymbol{X}_2\beta_2\right).$$

Then the solution $\widehat{\beta}_1 = \left(\boldsymbol{X}_1'\boldsymbol{X}_1\right)^{-1}\left(\boldsymbol{X}_1'\boldsymbol{y} - \boldsymbol{X}_1'\boldsymbol{X}_2\beta_2\right)$.

(b) If we ignore $\boldsymbol{X}_2$ in the regression, we have $\widehat{\beta}_1^M = \left(\boldsymbol{X}_1'\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1'\boldsymbol{y}$. Then $\widehat{\beta}_1 = \widehat{\beta}_1^M$ if and only if $\left(\boldsymbol{X}_1'\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_2\beta_2 = 0$, which is equivalent as $\boldsymbol{X}_1'\boldsymbol{X}_2 = 0$ or $\beta_2 = 0$.

(c) Solve the second normal equation, and plug in the estimate for β_1, we have

$$\begin{aligned} \boldsymbol{X}_2'\boldsymbol{X}_1\left(\boldsymbol{X}_1'\boldsymbol{X}_1\right)^{-1}\left(\boldsymbol{X}_1'\boldsymbol{y} - \boldsymbol{X}_1'\boldsymbol{X}_2\beta_2\right) + \boldsymbol{X}_2'\boldsymbol{X}_2\beta_2 =& \boldsymbol{X}_2'\boldsymbol{y} \\ \boldsymbol{X}_2'\boldsymbol{H}_1\boldsymbol{y} - \boldsymbol{X}_2'\boldsymbol{H}_1\boldsymbol{X}_2\beta_2 + \boldsymbol{X}_2'\boldsymbol{X}_2\beta_2 =& \boldsymbol{X}_2'\boldsymbol{y} \\ \beta_2 = \left(\boldsymbol{X}_2'(\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{X}_2\right)^{-1}& \boldsymbol{X}_2'(\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{y} \end{aligned}$$

(d) We have

$$\boldsymbol{\epsilon}_{\boldsymbol{X}_2|\boldsymbol{X}_1} = \left(\boldsymbol{I} - \boldsymbol{H}_1\right)\boldsymbol{X}_2, \quad \boldsymbol{\epsilon}_{\boldsymbol{y}|\boldsymbol{X}_1} = \left(\boldsymbol{I} - \boldsymbol{H}_1\right)\boldsymbol{y}.$$

Then $\widehat{\beta}_2$ could be written as

$$\begin{aligned} \widehat{\beta}_2 =& \left(\boldsymbol{X}_2'(\boldsymbol{I} - \boldsymbol{H}_1)(\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{X}_2\right)^{-1}\boldsymbol{X}_2'(\boldsymbol{I} - \boldsymbol{H}_1)(\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{y} \\ =& \left(\boldsymbol{\epsilon}_{\boldsymbol{X}_2|\boldsymbol{X}_1}'\boldsymbol{\epsilon}_{\boldsymbol{X}_2|\boldsymbol{X}_1}\right)^{-1}\boldsymbol{\epsilon}_{\boldsymbol{X}_2|\boldsymbol{X}_1}'\boldsymbol{\epsilon}_{\boldsymbol{y}|\boldsymbol{X}_1} \end{aligned}$$

2. (a) The table could be completed as follows,

```
Dependent Variable:  WGHT
                Sum of
Source   DF  Squares      Mean Square  F Value  Pr > F
Model    1   2624.670184  2624.67      137.91   <0.0001
Error    96  1827.099916  19.03
```

(b) The model assumptions for the ANOVA table are: homogenerity, independence, linearity, existence and Gaussian assumption.

(c) For the model $y_i = \beta_0 + \beta_1 \text{TIME}_i + \epsilon_i$, we test H_0: $\beta_1 = 0$ versus H_A: $\beta_1 \neq 0$. Thus, with a p-value < 0.0001, we reject the null hypothesis and conclude that average daily exercise time is a significant predictor of weight loss.

(d) Considering average daily exercise time and average daily running mileage, the results do indicate that neither variable is significant. This occurs in this added-last test because, after adjusting for running mileage, exercise time does not provide any additional useful information. Clearly, running mileage and exercise time are highly correlated.

3. (a) (i) $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4$

$$\theta = \begin{pmatrix} \beta_1 - \beta_2 \\ \beta_1 - \beta_3 \\ \beta_1 - \beta_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \Rightarrow \theta_0 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad C = \begin{pmatrix} 0 & 1 & -1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & -1 \end{pmatrix}$$

(ii) $H_0 : \begin{pmatrix} \beta_1 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} \beta_2 \\ \beta_4 \end{pmatrix}$

$$\begin{pmatrix} \beta_1 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} \beta_2 \\ \beta_4 \end{pmatrix} \Rightarrow \theta = \begin{pmatrix} \beta_1 - \beta_2 \\ \beta_3 - \beta_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} = \theta_0, \quad C = \begin{pmatrix} 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{pmatrix}$$

(iii) $H_0 : \begin{pmatrix} \beta_1 - 2\beta_2 \\ \beta_1 + 2\beta_2 \end{pmatrix} = \begin{pmatrix} 4\beta_3 \\ 0 \end{pmatrix}$

$$\theta = \begin{pmatrix} \beta_1 - 2\beta_2 - 4\beta_3 \\ \beta_1 - 2\beta_2 - 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} = \theta_0, \quad C = \begin{pmatrix} 0 & 1 & -2 & -4 & 0 \\ 0 & 1 & 2 & 0 & 0 \end{pmatrix}$$

(b) This is an added-last test, testing the effect of $\beta_2 = 0$ on the full model. Thus, we define:

$$C = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \end{pmatrix} \text{ and } \theta_0 = 0$$

and compare the full model:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$$

to the reduced model:

$$y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4 + \epsilon.$$

4 Consider the full model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \epsilon,$$

where the covariates $(X_1, \ldots, X_9)$ are height, weight, BMI, age, average treadmill elevation, average treadmill speed, temperature, barometric pressure, and humidity.

(a) To report the test of whether the group of predictors is important, we will test H_0: $\beta_1 = \beta_2 = ...\beta_8 = \beta_9 = 0$ versus H_A: at least one β_i, i = 1,...,9, not equal to zero. The F statistic takes the value of 13.90, with a p-value $<$ 0.0001, then we reject the null hypothesis and conclude that at least one of the predictors is a significant predictor of FVC.

(b) The corrected R^2 for the model using all the predictors is 0.4389.

(c) (i) It compares the models:

$$M_{10} : Y = \beta_0 + \epsilon$$
$$M_{11} : Y = \beta_0 + \sum_{j=1}^{9} \beta_j X_j + \epsilon.$$

$F_1 = 13.90 \sim F(9, 160)$ under null hypothesis. The p-value < 0.0001. Thus, we reject the null hypothesis and conclude that at least one of the predictors provides useful information about FVC.

(ii) It compares the models:

$$M_{20} : Y = \beta_0 + \epsilon$$
$$M_{21} : Y = \beta_0 + \beta_1 X_1 + \epsilon.$$

For the denominator of F statistic, we could plug in the MSE estimated from the full model or from M_{21}. So we have $F_{21} = 95.15 \sim F(1, 160)$ under null hypothesis with MSE from full model or $F_{22} = 85.05 \sim F(1, 168)$ under null hypothesis with MSE from M_{21}. The corresponding p-values are both < 0.0001. Thus, we reject the null hypothesis and conclude that height is a significant predictor for FVC.

(iii) It compares the models:

$$M_{30} : Y = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$
$$M_{31} : Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon.$$

For the denominator of F statistic, we could plug in the MSE estimated from the full model or from M_{31}. So we have $F_{31} = 0.050 \sim F(1, 160)$ under null hypothesis with MSE from full model or $F_{32} = 0.048 \sim F(1, 166)$ under null hypothesis with MSE from M_{31}. The corresponding p-values are 0.8233 and 0.8277. Thus, we do not reject the null hypothesis and conclude that height provides no information about FVC, when adjusting for weight and BMI.

(iv) It compares the models:

$$M_{40} : Y = \beta_0 + \sum_{j=2}^{9} \beta_j X_j + \epsilon$$
$$M_{41} : Y = \beta_0 + \sum_{j=1}^{9} \beta_j X_j + \epsilon.$$

$F_4 = 0.157 \sim F(1, 160)$ under null hypothesis. The p-value is 0.6926. Thus, we do not reject the null hypothesis and conclude that height does not provide any additional information about FVC when adjusting for all other predictors in the model.

(v) It compares the models:

$$M_{50} : Y = \beta_0 + \epsilon$$
$$M_{51} : Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon.$$

For the denominator of F statistic, we could plug in the MSE estimated from the full model or from M_{51}. So we have $F_{51} = 37.32 \sim F(3, 160)$ under null hypothesis with MSE from full model or $F_{52} = 35.77 \sim F(3, 166)$ under null hypothesis with MSE from model M_{51}. The corresponding p-values are both < 0.0001. Thus, we reject the null hypothesis and conclude that the group of body size variables (height, weight, BMI) is a significant predictor for the mean level of FVC.

(d) Here we test $H_0 : \beta_9 = 0$ versus $H_A : \beta_9 \neq 0$. This is an added-in-order test (or added last test). With an observed F-statistic of 0.28 and a p-value = 0.5982, we do not reject the null hypothesis and conclude that humidity has no affect on FVC after adjusting for all other variables in the model.

(e) Considering the parameter estimates for the full model, we see that one unit increase in height is associated with a 30.32 unit decrease in FVC. Additionally, one unit increase in weight is associated with a 116.07 unit increase in FVC. Finally, one unit increase in BMI is associated with a 275.62 unit decrease in FVC. However, because these three variable are highly correlated, any single variable is not significantly correlated with FVC given the other variables.

(f) The parameter estimates from the original model indicate that the largest FVC tends to be associated with individuals who are shorter, heavier, older, and have lower BMI, larger average treadmill spead, larger average treadmill elevation, running under conditions of lower temperature, higher barometer pressure and higher humidity.