STAT 608, Spring 2021 - Assignment 2
SOLUTIONS

1. State the geometric reason that for a dummy variable model with a single dummy variable (i.e., $y_i = \alpha_0 + \alpha_1 x_i + e_i$, where $x_i = 1$ if success, 0 if failure) such that the first 5 observations are successes and the last 5 are failures ($n = 10$), the sum of the first five residuals equals zero: $\sum_{i=1}^{5} \hat{e}_i = 0$.

   THE DESIGN MATRIX HERE IS

   $$\mathbf{X} = \begin{bmatrix} \mathbf{1}_5 & \mathbf{1}_5 \\ \mathbf{1}_5 & \mathbf{0}_5 \end{bmatrix}$$

   WHERE $\mathbf{1}_5 = (1,1,1,1,1)'$ AND $\mathbf{0}_5 = (0,0,0,0,0)'$. SINCE THE RESIDUAL VECTOR $\hat{\mathbf{e}}$ IS ORTHOGONAL TO THE SPACE SPANNED BY THE COLUMNS OF $\mathbf{X}$ AND $\mathbf{v} = (\mathbf{1}_5', \mathbf{0}_5')'$ IS IN THAT SPACE (BECAUSE $\mathbf{v}$ IS JUST THE SECOND COLUMN OF $\mathbf{X}$), WE HAVE $\hat{\mathbf{e}}'\mathbf{v} = 0$, WHICH IMPLIES $\sum_{i=1}^{5} \hat{e}_i = 0$.

2. A researcher is interested in how consumption of fat and sugar affects weight gain in rats. Assume for a moment that fat and sugar do not interact; that is, the effect of sugar on weight gain is the same whether or not a rat is consuming a high-fat diet, and vice versa. In an experiment, each of 6 rats are fed high-fat and/or high-sugar diets as follows: each rat takes their fat or sugar serving first, and then eats as much regular rat chow as desired. The response variable measured is the amount of weight gain of the rats (with weight loss measured as negative values) after two weeks on the diet.

   - The first rat is fed one mg of fat.
   - The second rat is fed one mg of sugar.
   - The third rat is fed one mg of fat and one mg of sugar.
   - The fourth rat is fed two mg of fat and two mg of sugar.
   - The fifth rat is fed two mg of fat and one mg of sugar.
   - The sixth rat is fed one mg of fat and two mg of sugar.

   Let β_1 be the average weight gain due to consuming fat and β_2 be the average weight gain due to consuming sugar. Write down a model using a matrix equation to estimate β_1 and β_2, giving your design matrix $\mathbf{X}$. Don't worry about trying to solve for estimates of β_1 and β_2; just write down the model and the design matrix.

   LET $\mathbf{y}$ BE THE VECTOR OF THE WEIGHT GAINS OF THE RATS. THEN WE'LL HAVE:

   $$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

   WHERE $\boldsymbol{\beta} = (\beta_1, \beta_2)'$, AND THE DESIGN MATRIX

   $$\mathbf{X} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \\ 2 & 2 \\ 2 & 1 \\ 1 & 2 \end{bmatrix}$$

3. Instead of using the y-intercept as in the notes and textbook, suppose we wanted to create a linear model using two dummy variables like this one: $y_i = \alpha_1 x_1 + \alpha_2 x_2 + e_i$, $i = 1, 2, \ldots, n$. You might think of the calcium supplement - blood pressure problem from class, but this time, in general there are m people in the first group and $n - m$ people in the second group. Our dummy variables are then defined as follows:

$$x_1 = \begin{cases} 1, & i = 1, 2, \ldots, m \\ 0, & i = m+1, m+2, \ldots, n \end{cases} \qquad x_2 = \begin{cases} 0, & i = 1, 2, \ldots, m \\ 1, & i = m+1, m+2, \ldots, n \end{cases}$$

   (a) Define the parameters α_1 and α_2 in the context of the problem.
   Since $\mu_1 = \mathrm{E}(y_i | x_1 = 1, x_2 = 0) = \mathrm{E}(\alpha_1 + e_i) = \alpha_1$ and $\mu_2 = \mathrm{E}(y_i | x_1 = 0, x_2 = 1) = \mathrm{E}(\alpha_2 + e_i) = \alpha_2$, we have that α_1 is the mean of the response from the first group, and α_2 is the mean of the response from the second group.

   (b) Use the usual formula $\hat{\boldsymbol{\alpha}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ to solve for the parameter estimates of α_1 and α_2. (Double-check: the estimates should be consistent with your definitions above.)
   Here the design matrix is

$$\mathbf{X} = \begin{bmatrix} \mathbf{1}_m & \mathbf{0}_m \\ \mathbf{0}_{n-m} & \mathbf{1}_{n-m} \end{bmatrix}$$

   Then

$$\begin{aligned} \hat{\boldsymbol{\alpha}} &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \\ &= \left(\begin{bmatrix} \mathbf{1}'_m & \mathbf{0}'_{n-m} \\ \mathbf{0}'_m & \mathbf{1}'_{n-m} \end{bmatrix} \begin{bmatrix} \mathbf{1}_m & \mathbf{0}_m \\ \mathbf{0}_{n-m} & \mathbf{1}_{n-m} \end{bmatrix} \right)^{-1} \begin{bmatrix} \mathbf{1}'_m & \mathbf{0}'_{n-m} \\ \mathbf{0}'_m & \mathbf{1}'_{n-m} \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_m \\ y_{m+1} \\ \vdots \\ y_n \end{bmatrix} \\ &= \begin{bmatrix} m & 0 \\ 0 & n-m \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^{m} y_i \\ \sum_{i=m+1}^{n} y_i \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{m}\sum_{i=1}^{m} y_i \\ \frac{1}{n-m}\sum_{i=m+1}^{n} y_i \end{bmatrix} \end{aligned}$$

   So $\hat{\alpha}_1 = \frac{1}{m}\sum_{i=1}^{m} y_i$, which is the sample mean of the response from the first group, and $\hat{\alpha}_2 = \frac{1}{n-m}\sum_{i=m+1}^{n} y_i$, which is the sample mean of the response from the second group. These estimates are consistent with the definitions of parameters in part (a).

4. (From Stapleton, 1995.) Suppose we have an ordinary household scale such as might be used in a kitchen. When an object is placed on the scale, the reading is a combination of the true weight plus random error. You have two coins of unknown weights β_1 and β_2. To estimate the weights of the coins, you take four observations:

   - Put coin 1 on the scale and observe y_1.

- Put coin 2 on the scale and observe y_2.
- Put both coins on the scale and observe y_3.
- Put both coins on the scale again and observe y_4.

Suppose the random errors are independent and identically distributed.

Write a linear model in matrix form and find the least-squares estimates of the coins' weights using the usual formula $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

Let x_1 be a dummy variable that takes the value 1 when coin 1 is on the scale and 0 otherwise. Similarly, let x_2 be a dummy variable that takes value 1 when coin 2 is on the scale and 0 otherwise. Then we have $y_i = \beta_1 x_1 + \beta_2 x_2 + e_i$. In matrix form, it is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

where $\mathbf{y} = (y_1, y_2, \ldots, y_4)'$, $\boldsymbol{\beta} = (\beta_1, \beta_2)'$, $\mathbf{e} = (e_1, e_2, \ldots, e_4)'$, and design matrix

$$\mathbf{X} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}$$

Then

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \\
&= \left(\begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} \\
&= \begin{bmatrix} 3 & 2 \\ 2 & 3 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} \\
&= \frac{1}{5} \begin{bmatrix} 3 & -2 & 1 & 1 \\ -2 & 3 & 1 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} \\
&= \frac{1}{5} \begin{bmatrix} 3y_1 - 2y_2 + y_3 + y_4 \\ -2y_1 + 3y_2 + y_3 + y_4 \end{bmatrix}
\end{aligned}$$

5. Question 2.5, Textbook (pp. 41-42)

D. From the two plots, we find the data points are much closer to the regression line of model 1 compared to that of model 2. So model 1 will have smaller sum of squared residuals $RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$. Because we are using the same responses, we have the same SST for both models. And we know $SST = SS_{\text{REG}} + RSS$, so model 1 will have greater SS_{REG}.

6. Question 2.6, Textbook (p. 42)

   (a) $y_i - \hat{y}_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right) = y_i - \left(\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i\right) = (y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x})$;

   (b) $\hat{y}_1 - \bar{y} = \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right) - \left(\hat{\beta}_0 + \hat{\beta}_1 \bar{x}\right) = \hat{\beta}_1 (x_i - \bar{x})$;

   (c) Based on the results from (a) and (b),

$$\begin{aligned}
\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) &= \sum_{i=1}^{n} \left[(y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x})\right] \hat{\beta}_1 (x_i - \bar{x}) \\
&= \hat{\beta}_1 \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) - \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \\
&= \frac{SXY}{SXX} \cdot SXY - \frac{SXY^2}{SXX^2} \cdot SXX \\
&= \frac{SXY^2}{SXX} - \frac{SXY^2}{SXX} \\
&= 0
\end{aligned}$$

7. For the simple linear regression model $y_i = \beta_0 + \beta_1 x_i + e_i$, we find the t-statistic for testing $H_0 : \beta_1 = 0$ to be $t = \left(\hat{\beta}_1 - 0\right) / \text{se}\left(\hat{\beta}_1\right)$.

   (a) Which of the usual assumptions of the model must be met in order for the t-statistic to have the t distribution? Why?

   The errors are independently and normally distributed with mean 0 and equal variance σ^2. Based on these conditions, y_i's are also independently and normally distributed with equal variance σ^2 given $\mathbf{X}$. So, from $\hat{\boldsymbol{\beta}} = (\mathbf{X}\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, we know that $\hat{\beta}_1$ is a linear combination of y_i's, then $\hat{\beta}_1$ is normally distributed. We also know that $\text{E}\left(\hat{\beta}_1 | \mathbf{X}\right) = \beta_1$, so $t = \left(\hat{\beta}_1 - 0\right) / \text{se}\left(\hat{\beta}_1\right)$ will have a t distribution.

   (b) Does having a larger sample size change your answer? Why or why not?

   If we have a larger sample size, we will have $\hat{\beta}_1$ approximately normally distributed without the condition of normality on the errors by CLT (Central Limit Theorem). Then $t = \left(\hat{\beta}_1 - 0\right) / \text{se}\left(\hat{\beta}_1\right)$ will still have a t distribution approximately. So we can drop the condition of normality on the errors.

8. Suppose $\mathbf{x}$ is a random n-dimensional vector and that $\text{E}(\mathbf{x}) = \boldsymbol{\mu}$. Show that the covariance matrix $\boldsymbol{\Sigma} = \text{E}\left[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})'\right]$ is equal to $\text{E}(\mathbf{x}\mathbf{x}') - \boldsymbol{\mu}\boldsymbol{\mu}'$.

$$\begin{aligned}
\boldsymbol{\Sigma} &= \text{E}\left[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})'\right] = \text{E}\left(\mathbf{x}\mathbf{x}' - \mathbf{x}\boldsymbol{\mu}' - \boldsymbol{\mu}\mathbf{x}' + \boldsymbol{\mu}\boldsymbol{\mu}'\right) \\
&= \text{E}\left(\mathbf{x}\mathbf{x}'\right) - \text{E}(\mathbf{x})\boldsymbol{\mu}' - \boldsymbol{\mu}\text{E}(\mathbf{x}') + \boldsymbol{\mu}\boldsymbol{\mu}' \\
&= \text{E}\left(\mathbf{x}\mathbf{x}'\right) - \boldsymbol{\mu}\boldsymbol{\mu}'
\end{aligned}$$