

## STAT 608, Spring 2022 - Assignment 2

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:  $\sum_{i=1}^5 \hat{e}_i = 0$ .
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  $\beta_1$  be the average weight gain due to consuming fat and  $\beta_2$  be the average weight gain due to consuming sugar. Write down a model using a matrix equation to estimate  $\beta_1$  and  $\beta_2$ , giving your design matrix  $\mathbf{X}$ . Don't worry about trying to solve for estimates of  $\beta_1$  and  $\beta_2$ ; just write down the model and the design matrix.

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:  $y_i = \alpha_1 x_1 + \alpha_2 x_2 + e_i$ ,  $i = 1, 2, \dots, n$ . 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:

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

- (a) Define the parameters  $\alpha_1$  and  $\alpha_2$  in the context of the problem.
  - (b) Use the usual formula  $\hat{\alpha} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$  to solve for the parameter estimates of  $\alpha_1$  and  $\alpha_2$ . (Double-check: the estimates should be consistent with your definitions above.)
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  $\beta_1$  and  $\beta_2$ . 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  $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ .

5. Question 2.5, Textbook (pp. 41-42)
6. Question 2.6, Textbook (p. 42)
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  $t = (\hat{\beta}_1 - 0) / \text{se}(\hat{\beta}_1)$ .
  - (a) Which of the usual assumptions of the model must be met in order for the  $t$ -statistic to have the  $t$  distribution? Why?
  - (b) Does having a larger sample size change your answer? Why or why not?
8. Suppose  $\mathbf{x}$  is a random  $n$ -dimensional vector and that  $E(\mathbf{x}) = \boldsymbol{\mu}$ . Show that the covariance matrix  $\boldsymbol{\Sigma} = E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})']$  is equal to  $E(\mathbf{x}\mathbf{x}') - \boldsymbol{\mu}\boldsymbol{\mu}'$ .