Homework 3

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 ($n = 10$), $\sum_{i=1}^{5} \hat{e}_i = 0$.

2. (From Meyer, Practical Statistical Models, 2002.) A botanist is interested in the efficacy on predator bugs in reducing pests on garden plants. In particular, two species (A and B) of praying mantis are to be compared to see which devours potato beetles at a higher rate. One hundred grams of potato beetles are released into each of four cages containing potato foliage, and at the end of a week, the reduction (in grams) of potato beetles is measured.

   - The first cage contains one mantis of each species.
   - The second cage contains two of each species.
   - For the third cage, there are two of Species A and one of Species B,
   - In the fourth cage, there are two of Species B and one of Species A.

   Let β_A be the average grams of potato beetles eaten per week per praying mantis for Species A, and let β_B be the average for Species B. Write down a model using a matrix equation to estimate β_A and β_B, giving your design matrix $\mathbf{X}$. Assume that the consumption of each praying mantis is independent of others in the cage. Don't worry about trying to solve for estimates of β_A and β_B; just write down the model.

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

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

   (a) Define the parameters α_1 and α_2 in the context of the problem.

   (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.)

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 a 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.

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

(b) Explain in words why at least parts of the estimates of the coins make sense. (The denominator 5 may not be intuitive.)

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 $t = (\hat{\beta}_1 - \beta_1)/se(\hat{\beta}_1)$.

   (a) Which of the usual assumptions for 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? (Hint: I'm not 100% convinced the 408 students will be able to answer this question on their own; ask me about it in class.)

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