Jack Rodan
STAT 6078
Assignment 2
2/9/22

→ see Chp 2 Notes (slides 77-83)

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

- $\sum_{i=1}^{5} \hat{e}_i = \underline{\hat{e}} \cdot \underline{1}$. The vector $\underline{1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \in \text{span}(X)$ (where X is our design matrix $X = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_{10} \end{bmatrix}$ in the case given).
  The vector $\underline{\hat{e}}$ is the distance btwn the vector y and its projection $\hat{y}$ onto X. Thus $\underline{\hat{e}} \perp \text{span}(X)$ ∴ the dot product of any two orthogonal vectors is 0.

ask if this is the answer professor wants

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. (Q. for me: is it the case that if there are no interactions btwn variables x_i, x_j in our model then we are assuming these two variables are independent?) 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 evenly first and then eats as much regular rat food as desired. The response variable measured is the amount of weight gained by the rats (w/ weight loss measured as negative values) after two weeks on the diet.

R_1:	1 mg fat
R_2:	1 mg sugar
R_3:	1 mg fat + 1 mg sugar
R_4:	2 mg fat + 2 mg sugar
R_5:	2 mg fat + 1 mg sugar
R_6:	1 mg fat + 2 mg sugar

Let β_1 be the average weight gain due to consuming fat & β_2 be the average weight gain due to consuming sugar. Write down a model using a matrix equation to estimate β_1, β_2, giving your design matrix X.

model: $\underline{y} = X\underline{\beta} + \underline{\epsilon} \iff y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \epsilon_i$

$$X = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix} \qquad \underline{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_6 \end{bmatrix}$$

$$\hat{\beta} = (X'X)^{-1}X'\underline{y}$$

3.) Instead of using the y-intercept as in the notes & textbook, suppose we wanted to create a linear model using two dummy variables like this one:

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

You might think of the calcium supplement – blood pressure problem from class (see chp 2 notes pg 68 - 76)*, but this time, there are m people in the 1st group and $n-m$ people in the 2nd group. Our dummy variables are then defined as follows:

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

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

- In the context of the example problem from chp 2 (see slides 68 - 76),
  x_1 is an indicator variable that equals 1 if the person received a placebo;
  x_2 is an indicator variable that equals 1 if the person received a calcium supplement.
- Now our model for blood pressure would look like:

$$y_i = \alpha_1 x_{1i} + \alpha_2 x_{2i} + e_i \quad , \quad i = 1, 2, \ldots, n$$

- For placebo group:

$$y_i = \alpha_1 x_1 + e_i$$

- For calcium group:

$$y_i = \alpha_2 x_2 + e_i.$$

- α_1 is the average blood pressure for an individual receiving the placebo
- α_2 is the average blood pressure for an individual receiving the calcium supplement.

(b) Use the formula $\hat{\alpha} = (X'X)^{-1} X'y$ to solve for the parameter estimates of α_1 & α_2.

- $(X'X) = \begin{bmatrix} 1 & 1 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} m & 0 \\ 0 & n-m \end{bmatrix}$
- $(X'X)^{-1} = \frac{1}{mn - m^2} \begin{bmatrix} n-m & 0 \\ 0 & m \end{bmatrix} = \begin{bmatrix} \frac{1}{m} & 0 \\ 0 & \frac{1}{n-m} \end{bmatrix}$
- $X'y = \begin{bmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{m} y_i \\ \sum_{i=m+1}^{n} y_i \end{bmatrix}$

$$\Rightarrow (X'X)^{-1} X'y = \begin{bmatrix} \frac{1}{m} & 0 \\ 0 & \frac{1}{n-m} \end{bmatrix} \begin{bmatrix} \sum_{i=1}^{m} y_i \\ \sum_{i=m+1}^{n} y_i \end{bmatrix}$$

$$\begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{m} \sum_{i=1}^{m} y_i \\ \frac{1}{n-m} \sum_{i=m+1}^{n} y_i \end{bmatrix} = \begin{bmatrix} \bar{y}_{placebo} \\ \bar{y}_{calcium} \end{bmatrix}$$

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 & β_2. To estimate the weight 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 & observe y_4.

Suppose the random errors are i.i.d.

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

$$\underline{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_4 \end{bmatrix} \qquad X = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \qquad \underline{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$

$$(X'X) = \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 3 & 2 \\ 2 & 3 \end{bmatrix} \Rightarrow (X'X)^{-1} = \frac{1}{5}\begin{bmatrix} 3 & -2 \\ -2 & 3 \end{bmatrix}$$

$$X'\underline{y} = \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_4 \end{bmatrix} = \begin{bmatrix} y_1 + y_3 + y_4 \\ y_2 + y_3 + y_4 \end{bmatrix}$$

$$(X'X)^{-1}X'\underline{y} = \frac{1}{5}\begin{bmatrix} 3 & -2 \\ -2 & 3 \end{bmatrix} \begin{bmatrix} y_1 + y_3 + y_4 \\ y_2 + y_3 + y_4 \end{bmatrix} = \frac{1}{5}\begin{bmatrix} 3(y_1 + y_3 + y_4) - 2(y_2 + y_3 + y_4) \\ 3(y_2 + y_3 + y_4) - 2(y_1 + y_3 + y_4) \end{bmatrix}$$

$$\hat{\underline{\beta}} = \begin{bmatrix} \frac{1}{5}(3y_1 - 2y_2 + y_3 + y_4) \\ \frac{1}{5}(3y_2 - 2y_1 + y_3 + y_4) \end{bmatrix} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}$$

Q: Why do these answers make sense?

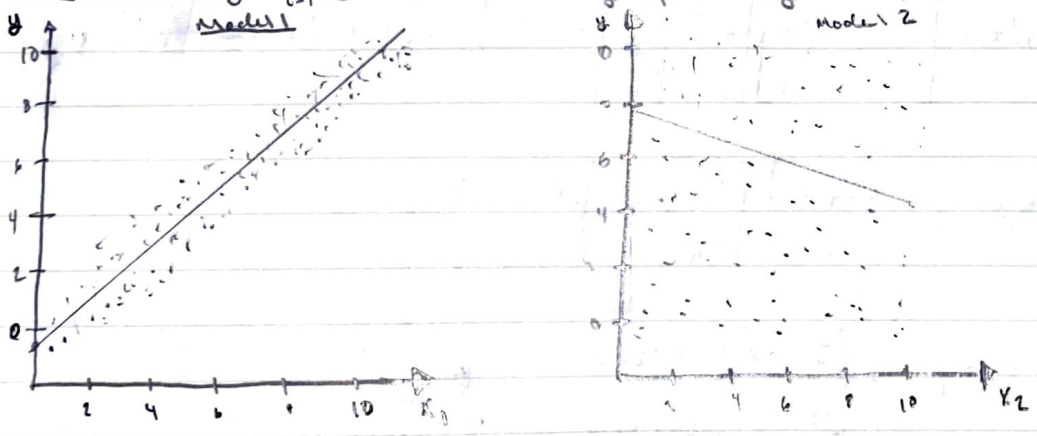
- $3y_1$ - we have 3 obs involving coin 1. y_1 is an estimate of just coin 1's weight.
- $(y_3 + y_4) - 2y_2$ - y_3 & y_4 are both estimates of the combined weight of coin 1 & coin 2, two to get the estimated contribution of coin 1 to both total weights, we subtract off y_2 from each, b/c y_2 is an estimate of just coin 2's weight.
- $1/5$ - we use 5 weights to estimate the weight of coin 1. y_1 involves 1 weight, y_3 & y_4 involve 2 weights each.
- Analogous reasoning for $\hat{\beta}_2$.

5.) (Chp 2 Question 5.)

(5) Two alternative straight line regression models have been proposed for Y. In the first model, Y is a linear function of X_1, while in the second model Y is a linear function of X_2. The plot in the first column of figure 2.8 is that of Y against X_1 while the second plot is that of Y against X_2. Those plots also show the least squares regression lines. In the following statements RSS stands for residual sum of squares while SSreg stands for regression sum of squares. Which of the following is true?

NOTE (see book pg. 28)*

- $SST = SYY = \sum_{i=1}^{n}(y_i - \bar{y})^2 = RSS + SSreg$ (Total sample variability)
- $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ (variability of errors)
- $SSreg = \sum_{i=1}^{\hat{y}_i}(\hat{y}_i - \bar{y})^2$ (variability explained by model)



- We can see clearly from the above plots that

(i) $RSS(M_1) < RSS(M_2)$. M_1 seems to be a much better fit than M_2. We can see that the y values are much closer to the fitted line of M_1 than they are for M_2. Thus, the $RSS(M_1) < RSS(M_2)$.

(ii) $SSreg(M_1) > SSreg(M_2)$. The line fitted by M_2 is almost a straight line at $y = \bar{y}$, thus the $SSreg(M_2) = \sum(\hat{y}_i - \bar{y})^2$ will be relatively small. This while y & X_2 seem to be uncorrelated, thus. $E[Y|X] \approx E[Y] = \bar{y}$.

$\Rightarrow$ **D**

6.) (Chp 2, Question 6)

Show that $SST = SSreg + RSS$. To do this, show that

$$\sum (y_i - \hat{y}_i)(y_i - \bar{y}) = 0$$

(a) Show that $(y_i - \hat{y}_i) = (y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x})$

$$\begin{aligned} y_i - \hat{y}_i &= y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \\ &= y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i \\ &= (y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x}) \end{aligned}$$

* we know from previous derivations of $\hat{\beta}_0$ & $\hat{\beta}_1$ that

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ & $\hat{\beta}_1 = \frac{SXY}{SXX} = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$

(b) Show that $(\hat{y}_i - \bar{y}) = \hat{\beta}_1 (x_i - \bar{x})$

$$\hat{y}_i - \bar{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y} = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i - \bar{y} = \hat{\beta}_1 (x_i - \bar{x})$$

(c) Utilizing the fact that $\hat{\beta}_1 = \frac{SXY}{SXX}$, show that $\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$

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

see Chp3 notes: pg 9, 35

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

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

- Normality of errors. If the ϵ_i are not $\sim N$, then $\hat{\beta}_1$ is not a linear combination of normally distributed random variables $\Rightarrow$ β_1 is not normally distributed $\Rightarrow$ $\hat{\beta}_1$ doesn't have a t-distribution.

the post about this on moodle.

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

Yes, if we have a larger sample size then the CLT applies.

8.) Suppose $\underline{X}$ is a random n dimensional vector and that $E[\underline{X}] = \underline{\mu}$. Show that the covariance matrix $\Sigma = E[(\underline{X}-\underline{\mu})(\underline{X}-\underline{\mu}')] = E[XX'] - \underline{\mu}\,\underline{\mu}'$

- $\Sigma = E[(\underline{X}-\underline{\mu})(\underline{X}-\underline{\mu}')] = E[\underline{X}^2 - 2\underline{\mu}'\underline{X} - \underline{\mu}'\underline{\mu}]$
  $= E[X'X - 2\underline{\mu}'\underline{X} + \underline{\mu}'\underline{\mu}]$
  $= E[\underline{X}'\underline{X}] - 2E[\underline{\mu}'\underline{X}] + E[\underline{\mu}'\underline{\mu}]$
  $= E[\underline{X}'X] - 2\underline{\mu}'\underline{\mu} + \underline{\mu}'\underline{\mu}$

$\Sigma = E[(\underline{X}-\underline{\mu})(\underline{X}-\underline{\mu}')] = E[X'X] - \underline{\mu}'\underline{\mu}$