TEXAS A&M UNIVERSITY

Department of Statistics

STAT 608 - Regression Analysis
Homework IV

Salih Kilicli

June 29, 2019

Question 1: **Suppose we have a linear model**

$$y_i = \alpha_1 x_{i1} + \alpha_2 x_{i2} + e_i, \quad i = 1, 2, ..., n$$

with two "dummy" predictor variables

$$x_{i1} = \begin{cases} 1, & i = 1, 2, ..., m \\ 0, & i = m+1, ..., n \end{cases} \quad ; \quad x_{i2} = \begin{cases} 0, & i = 1, 2, ..., m \\ 1, & i = m+1, ..., n \end{cases}$$

There are m people in the first group, and $n - m$ people in the second group.

1.1 Interpret the parameters α_1 and α_2 in the context of the problem.

1.2 Use the formula $\hat{\alpha} = (X'X)^{-1}X'y$ to obtain explicit expressions for α_1 and α_2 in terms of m, n and $y_1, ..., y_n$.

Solution: 1.1 α_1 and α_2 measures the additive change in Y_i due to dummy variables x_{i1}, x_{i2}, respectively. For example,

$$Y_i = \alpha_1 + e_i \quad \text{for} \quad i = 1, 2, \ldots, m$$

$$Y_i = \alpha_2 + e_i \quad \text{for} \quad i = m+1, \ldots, n \quad .$$

Therefore, the mean difference between dummy variables is $\alpha_1 - \alpha_2$.

1.2

$$\begin{bmatrix} \hat{\alpha_1} \\ \hat{\alpha_2} \end{bmatrix} = \hat{\alpha} = (X'X)^{-1}X'y$$

where X is an (nx2) matrix whose 1^{st} column consists of 1's for first m row, whereas 2^{nd} column consists of 1's for last $(n-m)$ rows, and zero elsewhere. Then,

$$X = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix} ; \quad (X'X) = \begin{bmatrix} m & 0 \\ 0 & n-m \end{bmatrix} ; \quad (X'X)^{-1} = \begin{bmatrix} \frac{1}{m} & 0 \\ 0 & \frac{1}{n-m} \end{bmatrix}$$

Then, multiplying $(X'X)^{-1}$ by $X'y$ yields the $\hat{\alpha}$ matrix as;

$$\begin{bmatrix} \hat{\alpha_1} \\ \hat{\alpha_2} \end{bmatrix} = \begin{bmatrix} \frac{1}{m} \sum_{i=1}^{m} y_i \\ \frac{1}{(n-m)} \sum_{i=m+1}^{n} y_i \end{bmatrix} = \begin{bmatrix} \frac{1}{m}(y_1 + y_2 + \cdots + y_{m-1} + y_m) \\ \frac{1}{(n-m)}(y_{m+1} + y_{m+2} + \cdots + y_n) \end{bmatrix}$$

Question 2: **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 the sum of the true weight and 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 with mean 0 and variance σ^2.

2.1 Write a linear model in matrix form and find explicit expressions in terms of $y_1, ..., y_4$ for the least-squares estimates of the coin weights.

2.2 Explain in words why these estimates make intuitive sense.

Solution :

2.1 A linear model can be written in the form

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + e_i, \quad i = 1, 2, ..., n$$

with two "dummy" predictor variables

$$x_{i1} = \begin{cases} 1, & i = 1, 3, 4 \\ 0, & i = 2 \end{cases} \quad ; \quad x_{i2} = \begin{cases} 0, & i = 1 \\ 1, & i = 2, 3, 4 \end{cases}$$

. Then the model can be written in matrix form $Y = X\beta + E$ where;

$$X = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} ; \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} ; \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} ; \quad E = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{bmatrix}$$

Then, the coefficients $\hat{\beta}_1$, and $\hat{\beta}_2$ can be estimated using $\hat{\beta} = (X'X)^{-1}X'y$. Calculating $(X'X)$ and multiplying it by $X'y$ yields;

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

Therefore, the estimates for the coefficients are;

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

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

2.2 Plugging y_i values into the estimates we get,

$$\hat{\beta}_1 = \frac{1}{5}(3(\beta_1+e_1)-2(\beta_2+e_2)+(\beta_1+\beta_2+e_3)+(\beta_1+\beta_2+e_4)) = \beta_1+\frac{1}{5}(3e_1-2e_2+e_3+e_4)$$

$$\hat{\beta}_2 = \frac{1}{5}(-2(\beta_1+e_1)+3(\beta_2+e_2)+(\beta_1+\beta_2+e_3)+(\beta_1+\beta_2+e_4)) = \beta_2+\frac{1}{5}(3e_2-2e_1+e_3+e_4)$$

Moreover, $E[\hat{\beta}_1] = \beta_1$ and $E[\hat{\beta}_2] = \beta_2$ since $E[e_i] = 0$ for every i. Intuitively it makes sense to me because $\hat{\beta}_1$ and $\hat{\beta}_2$ are unbiased estimates for β_1 and β_2, respectively.

(This part of the solution is inspired from TA's help on the discussion board.)

Additionally, we can represent these estimators as a weighted mean of the two unbiased estimators $\hat{\alpha}_1, \hat{\alpha}_2$ and $\hat{\gamma}_1, \hat{\gamma}_2$, that is,

$$\hat{\beta}_1 = (3/5)\hat{\alpha}_1 + (2/5)\hat{\alpha}_2; \quad \hat{\beta}_2 = (3/5)\hat{\gamma}_1 + (2/5)\hat{\gamma}_2$$

where

$$\hat{\alpha_1} = y_1 = \beta_1 + e_1, \quad \hat{\alpha_2} = (y_3 + y_4 - 2y_2)/2 = \beta_1 + (e_3 + e_4 - 2e_2)/2$$

and

$$\hat{\gamma}_1 = y_2 = \beta_2 + e_2, \quad \hat{\gamma}_2 = (y_3 + y_4 - 2y_1)/2 = \beta_2 + (e_3 + e_4 - 2e_1)/2.$$

Clearly, $\hat{\alpha_1}$ and $\hat{\alpha_2}$ are unbiased estimates for β_1 and, similarly, $\hat{\gamma}_1$ and $\hat{\gamma}_2$ are unbiased estimates for β_2 since $E[e_i] = 0$ for every i. In general, the weighted mean of $\hat{\alpha}_i^1$ and $\hat{\alpha}_i^2$ can be written as

$$w_i\hat{\alpha}_i^2 + (1 - w_i)\hat{\alpha}_i^2$$

where weights are picked to be inversely proportional to variance of each error term e_i in order to fix a non-constant error variance issue (the new error terms will be represented by $\epsilon_i = \sqrt{w_i}e_i$ with constant variance σ^2, where var(e$_i$) $= \frac{\sigma^2}{w_i}$).

Question 3: **Consider the linear model**

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$$

in which the column vectors x_1 and x_2 of the design matrix have mean 0 and length 1. Let ρ be the Pearson correlation coefficient between x_1 and x_2.

3.1 Show that

$$X'X = \begin{bmatrix} n & 0 & 0 \\ 0 & 1 & \rho \\ 0 & \rho & 1 \end{bmatrix}$$

and verify that

$$(X'X)^{-1} = \begin{bmatrix} \frac{1}{n} & 0 & 0 \\ 0 & \frac{1}{1-\rho^2} & \frac{-\rho}{1-\rho^2} \\ 0 & \frac{-\rho}{1-\rho^2} & \frac{1}{1-\rho^2} \end{bmatrix}.$$

3.2 Determine what values of ρ will make the variance of $\hat{\beta}_1$ and $\hat{\beta}_2$ larger than $5\sigma^2$.

Solution: 3.1 Let's assume

$$x_1 = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}; \quad x_2 = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

where $\bar{x_1} = \frac{1}{n}\sum_{i=1}^{n} a_i = 0 = \frac{1}{n}\sum_{i=1}^{n} b_i = \bar{x_2}$, and $\sum_{i=1}^{n} a_i^2 = \sum_{i=1}^{n} b_i^2 = 1$. Notice that,

$$\rho = \frac{\sum_{i=1}^{n}(x_{1i} - \bar{x_1})(x_{2i} - \bar{x_2})}{\sqrt{\sum_{i=1}^{n}(x_{1i} - \bar{x_1})^2 \sum_{i=1}^{n}(x_{2i} - \bar{x_2})^2}} = \frac{\sum_{i=1}^{n}(a_i b_i)}{\sqrt{\sum_{i=1}^{n} a_i^2 \sum_{i=1}^{n} b_i^2}} = \sum_{i=1}^{n} a_i b_i. \text{ Now,}$$

$$X = \begin{bmatrix} J & x_1 & x_2 \end{bmatrix} = \begin{bmatrix} 1 & a_1 & b_1 \\ \vdots & \vdots & \vdots \\ 1 & a_n & b_n \end{bmatrix}; and \quad X' = \begin{bmatrix} 1 & \dots & 1 \\ a_1 & \dots & a_n \\ b_1 & \dots & b_n \end{bmatrix} Then,$$

$$(X'X) = \begin{bmatrix} \sum_{i=1}^{n} 1 & \sum_{i=1}^{n} a_i & \sum_{i=1}^{n} b_i \\ \sum_{i=1}^{n} a_i & \sum_{i=1}^{n} a_i^2 & \sum_{i=1}^{n} a_i b_i \\ \sum_{i=1}^{n} b_i & \sum_{i=1}^{n} b_i a_i & \sum_{i=1}^{n} b_i^2 \end{bmatrix} = \begin{bmatrix} n & 0 & 0 \\ 0 & 1 & \rho \\ 0 & \rho & 1 \end{bmatrix}$$

Also, $det(X'X) = n(-1)^{1+1}(1-\rho^2) + 0 + 0 = n(1-\rho^2)$. Now, let us calculate the inverse matrix $(X'X)^{-1}$.

$$(X'X)^{-1} = \frac{1}{n(1-\rho^2)} \begin{bmatrix} (-1)^{1+1}\begin{vmatrix} 1 & \rho \\ \rho & 1 \end{vmatrix} & 0 & 0 \\ 0 & (-1)^{2+2}\begin{vmatrix} n & 0 \\ 0 & 1 \end{vmatrix} & (-1)^{2+3}\begin{vmatrix} n & 0 \\ 0 & \rho \end{vmatrix} \\ 0 & (-1)^{2+3}\begin{vmatrix} n & 0 \\ 0 & \rho \end{vmatrix} & (-1)^{3+3}\begin{vmatrix} n & 0 \\ 0 & 1 \end{vmatrix} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{n} & 0 & 0 \\ 0 & \frac{1}{1-\rho^2} & \frac{-\rho}{1-\rho^2} \\ 0 & \frac{-\rho}{1-\rho^2} & \frac{1}{1-\rho^2} \end{bmatrix}$$

3.2 Notice, the $Var(\hat{\beta}_i) = \sigma^2 t_{ii}$ where t_{ii} is the diagonal element of $(X'X)^{-1}$ matrix. Therefore $Var(\hat{\beta}_1) = Var(\hat{\beta}_2) = \frac{\cancel{\sigma^2}}{1-\rho^2} > 5\cancel{\sigma^2} \Rightarrow (1-\rho^2) < \frac{1}{5} \Rightarrow \rho^2 > \frac{4}{5} \Rightarrow |\rho| > \frac{2}{\sqrt{5}}$ is the solution.

Question 4: **In a study on weight gain in rabbits, researchers randomly assigned 6 rabbits to $1,\ 2$ or 3 mg of one of dietary supplement A or B (one rabbit to each level of each supplement). Consider the linear model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$, where x_1 is the dosage level of the supplement, and x_2 is a dummy variable indicating the type of supplement used.**

4.1 Compute the variance inflation factor for the covariate x_1.

4.2 Now suppose the researcher used instead $1,\ 2$ and 3 mg for supplement A, and $2,\ 3$ and 4 mg for supplement B. What is the variance inflation factor for the covariate x_1 in this case? Explain why it is larger or smaller than in 4.1 above.

Solution: 4.1 First of all, let

$$x_1 = \begin{bmatrix}1\\2\\3\\1\\2\\3\end{bmatrix};\quad x_2 = \begin{bmatrix}1\\1\\1\\0\\0\\0\end{bmatrix}\text{; and}\quad X = \begin{bmatrix}1&1&1\\1&2&1\\1&3&1\\1&1&0\\1&2&0\\1&3&0\end{bmatrix}$$

Therefore; $\bar{x_1} = \frac{1}{6}2(1+2+3) = 2$, and $\bar{x_2} = \frac{1}{6}(1+1+1) = 0.5$. Moreover,

$$\rho(x_1, x_2) = \frac{\sum_{i=1}^{6}(x_{1i} - \bar{x_1})(x_{2i} - \bar{x_2})}{\sqrt{\sum_{i=1}^{6}(x_{1i} - \bar{x_1})^2 \sum_{i=1}^{6}(x_{2i} - \bar{x_2})^2}}$$
$$= \frac{0.5[-1+0+1] - 0.5[-1+0+1]}{\sqrt{6}} = 0$$

Therefore the variance inflation factor for the covariate x_1,

$$VIF(x_1) = \frac{1}{1-\rho^2(x_1, x_2)} = \frac{1}{1-0} = 1$$

4.2 Now, let

$$x_1 = \begin{bmatrix}1\\2\\3\\2\\3\\4\end{bmatrix};\quad x_2 = \begin{bmatrix}1\\1\\1\\0\\0\\0\end{bmatrix}\text{; and}\quad X = \begin{bmatrix}1&1&1\\1&2&1\\1&3&1\\1&2&0\\1&3&0\\1&4&0\end{bmatrix}$$

Therefore; $\bar{x_1} = \frac{1}{6}(1+2(2+3)+4) = 2.5$, and $\bar{x_2} = \frac{1}{6}(1+1+1) = 0.5$. Then,

$$\rho(x_1, x_2) = \frac{\sum_{i=1}^{6}(x_{1i} - \bar{x_1})(x_{2i} - \bar{x_2})}{\sqrt{\sum_{i=1}^{6}(x_{1i} - \bar{x_1})^2 \sum_{i=1}^{6}(x_{2i} - \bar{x_2})^2}}$$
$$= \frac{[-1.5-0.5+0.5]0.5 - [-0.5+0.5+1.5]0.5}{\sqrt{(5.5)(1.5)}} \approx -0.522$$

Therefore, in this case $VIF(x_1) = \dfrac{1}{1-\rho^2(x_1,x_2)} = \dfrac{1}{1-(-0.522)^2} \approx 1.375$ which is bigger than first value. In the first case, VIF 1 implies that the predictor x_1 is not correlated with x_2 (since correlation=0), and x_1 values are independent of supplement type A or B. In the second case, VIF is higher because there is a negative correlation between variable x_1 with x_2 since x_1 values increase as x_2 values decrease.

Question 5: **(Kernel density estimation.** $Appendix A.1$**) Suppose the random variable V has a $N(0, h^2)$ distribution and that the random variable U is uniformly distributed on the set of numbers $x_1, ..., x_n$, that is,**

$$Pr[U = x_i] = \frac{1}{n} \quad \textbf{for} \quad i = 1, ..., n$$

Suppose also that V and U are independently distributed. Show that $Z = V + U$ has density function

$$f(x) = \frac{1}{nh} \sum_{i=1}^{n} K\big(\frac{z - x_i}{h}\big), \quad -\infty < z < \infty$$

where

$$K(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}$$

denotes the standard normal (0, 1) density function. [Hint: the density function of V at a point y is $K(y/h)/h$.]

Solution: First, the probability density function of U and V are, respectively, given by;

$$F_U(x) = \begin{cases} \frac{1}{n}, & x \in \{x_1, x_2, \ldots, x_n\} \\ 0, & otherwise \end{cases} \quad ; \quad F_V(x) = \frac{K(x/h)}{h}$$

where $K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ pdf of the standart normal N(0,1) distribution. Density function of sum of two independent random variables is found by the convolution of their density functions,i.e.,

$$\begin{aligned} f(z) &= \int_{-\infty}^{\infty} F_V(z - x) F_U(x) dx \\ &= \int_{\{x_1, x_2, \ldots, x_n\}} \frac{K(\frac{z-x}{h})}{h} \frac{1}{n} dx \quad \text{(since } F_U(x) = 0 \text{ otherwise)} \\ &= \frac{1}{nh} \sum_{i=1}^{n} K\Big(\frac{z - x}{h}\Big) \quad \text{(since integral is taken over a discrete set)} \end{aligned}$$

for $-\infty < z < \infty$.