

# **STAT151A Quiz 2**

Please write your full name and email address here:

Also, please put your initials on each page in case the pages get separated.

**You have 30 minutes for this quiz.**

**There are three questions, (1), (2), and (3), each weighted equally..**

**There are extra pages at the end if you need more space for solutions.**

## Question 1

Consider the regression  $y_n \sim \beta_1 + \beta_2 z_n + \beta_3 r_n$ , where  $n = 1, \dots, N$ . Let  $\beta = (\beta_1, \beta_2, \beta_3)^\top$ , and define the following quantities:

(a) Write the regression in the form  $\mathbf{Y} \sim \mathbf{X}\beta$ . Be precise about the dimensions and entries of the matrix  $\mathbf{X}$ .

(b) In terms of  $N$  and the quantities defined in Equation 1, write expressions for  $\mathbf{X}^\top \mathbf{X}$  and  $\mathbf{X}^\top \mathbf{Y}$ .

$$\begin{aligned} \bar{y} &= \frac{1}{N} \sum_{n=1}^N y_n & \overline{yz} &= \frac{1}{N} \sum_{n=1}^N y_n z_n & \overline{yr} &:= \frac{1}{N} \sum_{n=1}^N y_n r_n \\ \bar{z} &= \frac{1}{N} \sum_{n=1}^N z_n & \overline{zr} &= \frac{1}{N} \sum_{n=1}^N z_n r_n & \bar{r} &:= \frac{1}{N} \sum_{n=1}^N r_n \end{aligned} \tag{1}$$

(c) Suppose now that:

- $z_n$  and  $r_n$  are both random and IID. So  $z_n$  is independent of  $r_n$ , and both  $z_n$  and  $r_n$  are independent of  $z_m$  and  $r_m$  for  $m \neq n$ ,
- $\mathbb{E}[z_n] = \mathbb{E}[r_n] = 0$ , and
- $\text{Var}(z_n) = \sigma_z^2$ , and  $\text{Var}(r_n) = \sigma_r^2$ .

What is  $\lim_{N \rightarrow \infty} \frac{1}{N} \mathbf{X}^\top \mathbf{X}$ ?

## Question 2

Consider two categorical variables,  $z_{n1}$  and  $z_{n2}$ , where  $z_{n1}$  is either “good” or “bad”, and  $z_{n2}$  is either “red” or “yellow”. Note, for example, that an observation can be “good” and “red” at the same time.

However, an observation cannot be “good” and “bad” at the same time, nor can it be “red” and “yellow” at the same time.

Define the one-hot encodings

$x_{ng} = 1$  when  $z_{n1}$  is “good” and 0 otherwise

$x_{nb} = 1$  when  $z_{n1}$  is “bad” and 0 otherwise

and

$x_{nr} = 1$  when  $z_{n2}$  is “red” and 0 otherwise

$x_{ny} = 1$  when  $z_{n2}$  is “yellow” and 0 otherwise.

Consider the regression  $y_n \sim \beta_1 x_{ng} + \beta_2 x_{nr}$ , where  $n = 1, \dots, N$ . Let  $\beta = (\beta_1, \beta_2)^\top$ . That is, we are regressing only on the one-hot encodings for “good” and for “red”.

Let  $N_g$  denote the number of rows with  $z_{n1} = \text{“good”}$ ,  $N_r$  denote the number of rows with  $z_{n2} = \text{“red”}$ , and so on. Similarly, let  $N_{gr}$  denote the number of rows with both  $z_{n1} = \text{“good”}$  and  $z_{n2} = \text{“red”}$ .

(a) Write  $\mathbf{X}^\top \mathbf{X}$  in terms of  $N_g$ ,  $N_r$ , and  $N_{gr}$ .

(b) Write a formula in terms of  $N_g$ ,  $N_r$ , and  $N_{gr}$  that tells when  $\mathbf{X}^\top \mathbf{X}$  is invertible.

Hint: recall that the determinant of the  $2 \times 2$  matrix  $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$  is given by  $ad - bc$ .

(c) Suppose that every “good” row is also “red”, and every “red” row is “good”. Is  $\mathbf{X}^\top \mathbf{X}$  invertible? Justify your answer.

### Question 3

In the setting of **Question 2**, consider the regression  $y_n \sim \beta_0 + \beta_1 x_{ng} + \beta_2 x_{nb}$ . Note that a row is either “good” or “bad”, so that exactly one of  $x_{ng}$  or  $x_{nb}$  is equal to 1 for any particular observation  $n$ . Let  $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^\top$ , and  $\mathbf{X}$  the corresponding regressor matrix.

(a) Let  $N_g$  denote the number of rows with  $z_{n1} = \text{“good”}$  and  $N_b$  denote the number of rows with  $z_{n1} = \text{“bad”}$ . In terms of  $N$ ,  $N_g$ , and  $N_b$ , write an expression for  $\mathbf{X}^\top \mathbf{X}$ .

(b) Suppose that  $\bar{y}_g = \frac{1}{N_g} \sum_{\text{good } n} y_n$  and  $\bar{y}_b = \frac{1}{N_b} \sum_{\text{bad } n} y_n$  denote the average of  $y_n$  in “good” and “bad” rows, respectively. Find at least one  $\hat{\boldsymbol{\beta}}$  that satisfies  $\mathbf{X}^\top \mathbf{X} \hat{\boldsymbol{\beta}} = \mathbf{X}^\top \mathbf{Y}$ .

(c) Find another value  $\hat{\boldsymbol{\beta}}'$ , different than the answer you gave in (b), such that  $\hat{\boldsymbol{\beta}}'$  also satisfies  $\mathbf{X}^\top \mathbf{X} \hat{\boldsymbol{\beta}}' = \mathbf{X}^\top \mathbf{Y}$ .

Extra space for answers (indicate clearly which problem you are working on)

Extra space for answers (indicate clearly which problem you are working on)