Statistical Methods - Quiz #1(45 minutes)

September 15, 2025 (Monday)

Section(교반): __A1__ Cadet Number(교번): _____ Name(성명): _____ Score: _____

- All solutions must include a detailed step-by-step explanation.

1. Consider the following linear regression model.[50 points]

$$y_i = 1 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2) \ \text{ independently for } i = 1, \dots, n.$$

(1) Find the least squares estimate of β_1, denoted $\hat{\beta}_1$, that minimizes

$$S(\beta_1) = \sum_{i=1}^{n} (y_i - 1 - \beta_1 x_i)^2.$$

Solution:

Differentiate and set to zero:

$$\frac{dS}{d\beta_1} = -2 \sum_{i=1}^{n} x_i \left(y_i - 1 - \hat{\beta}_1 x_i \right) = 0 \tag{1}$$

$$\Rightarrow \boxed{\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i (y_i - 1)}{\sum_{i=1}^{n} x_i^2}}.$$

(2) Let the fitted values be $\hat{y}_i = 1 + \hat{\beta}_1 x_i$. Using the result in (A), show that

$$\sum_{i=1}^{n} (y_i - 1)^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - 1)^2.$$

Solution:

$$\begin{aligned}
\sum_{i=1}^{n} (y_i - 1)^2 &= \sum_{i=1}^{n} (y_i - \hat{y}_i + \hat{y}_i - 1)^2 \\
&= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - 1)^2 + 2 \sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - 1) \\
&= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} \hat{y}_i^2 + 2\hat{\beta}_1 \underbrace{\sum_{i=1}^{n} (y_i - 1 - \hat{\beta}_1 x_i) x_i}_{=0},
\end{aligned}$$

because of (1).

(3) Find $\mathrm{E}(\hat{\beta}_1)$.

Solution:

Starting from the estimator in (A),

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i (y_i - 1)}{\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} x_i (\beta_1 x_i + \varepsilon_i)}{\sum_{i=1}^{n} x_i^2} = \beta_1 + \frac{\sum_{i=1}^{n} x_i \varepsilon_i}{\sum_{i=1}^{n} x_i^2}.$$

By linearity of expectation and $\mathrm{E}(\varepsilon_i) = 0$,

$$\mathrm{E}(\hat{\beta}_1) = \beta_1 + \frac{\sum_{i=1}^{n} x_i \, \mathrm{E}(\varepsilon_i)}{\sum_{i=1}^{n} x_i^2} = \boxed{\beta_1}. \tag{2}$$

2. Suppose that we have the two multiple linear regression models.

$$\begin{aligned}&\text{Model A: } y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i \\ &\text{Model B: } y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i\end{aligned}$$

for $i = 1, \ldots, n$, where $\varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$. Let $\hat{y}_i^{(\text{A})}$ and $\hat{y}_i^{(\text{B})}$ be the fitted values for Model A and Model B, respectively. Compare each of the following quantities.(Hint: use projection.)[30 points]

(1) $\sum_{i=1}^n (y_i - \hat{y}_i^{(\text{A})})^2$ and $\sum_{i=1}^n (y_i - \hat{y}_i^{(\text{B})})^2$

Solution: Model B projects $\mathbf{Y}$ onto a larger subspace, so the projection is at least as close to $\mathbf{Y}$ as in Model A.

$$\sum_{i=1}^n (y_i - \hat{y}_i^{(\text{A})})^2 \;\; \geq \;\; \sum_{i=1}^n (y_i - \hat{y}_i^{(\text{B})})^2.$$

(2) $\sum_{i=1}^n (\hat{y}_i^{(\text{A})} - \bar{y})^2$ and $\sum_{i=1}^n (\hat{y}_i^{(\text{B})} - \bar{y})^2$

Solution: From the projection property we also have

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i^{(\text{A})})^2 + \sum_{i=1}^n (\hat{y}_i^{(\text{A})} - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i^{(\text{B})})^2 + \sum_{i=1}^n (\hat{y}_i^{(\text{B})} - \bar{y})^2.$$

Therefore, the explained sum of squares (ESS) satisfies $\sum_{i=1}^n (\hat{y}_i^{(\text{A})} - \bar{y})^2 \;\; \leq \;\; \sum_{i=1}^n (\hat{y}_i^{(\text{B})} - \bar{y})^2$.

(3) R^2 for Model A($= R^2_{\text{A}}$) and R^2 for Model B($= R^2_{\text{B}}$)

Solution: Finally, note that $R^2 = \text{ESS}/\text{TSS}$, where TSS is the same for both models. Thus,

$$R^2_{\text{A}} \;\; \leq \;\; R^2_{\text{B}}.$$

3. Consider the regression of *heart rate after exercise*(y_i) on categorical *exercise type* (with three levels: Running, Cycling, Swimming) using two indicator variables x_{i1} and x_{i2}. [20 points]

$$y_i \;\; = \;\; \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where $\varepsilon_i \sim N(0, \sigma^2)$ independently. x_{i1} and x_{i2} are defined as:

$$x_{i1} = \begin{cases} 1 & \text{if participant } i \text{ did Running,} \\ 0 & \text{otherwise,} \end{cases} \qquad x_{i2} = \begin{cases} 1 & \text{if participant } i \text{ did Cycling,} \\ 0 & \text{otherwise.} \end{cases}$$

Subject No. i	1	2	3	4	5	$\cdots$	n
Exercise type	Swimming	Swimming	Cycling	Cycling	Running	$\cdots$	Running
Heart rate y_i **(bpm)**	118	125	131	136	145	$\cdots$	138

(1) Express the mean of *heart rate* after cycling in terms of $\beta_0, \beta_1, \beta_2$.

Solution: The conditional means for each exercise type are:

$$\mathbb{E}[y \mid \text{Cycling}] = \mathbb{E}[y \mid x_1 = 0, x_2 = 1] = \boxed{\beta_0 + \beta_2}$$

(2) Express the mean difference in *heart rate* between Running and Swimming *(heart rate after Running - heart rate after Swimming)* in terms of $\beta_0, \beta_1, \beta_2$.

Solution: The conditional means for each exercise type are:

$$\mathbb{E}[y \mid \text{Running}] = \mathbb{E}[y \mid x_1 = 1, x_2 = 0] = \beta_0 + \beta_1, \quad \mathbb{E}[y \mid \text{Swimming}] = \mathbb{E}[y \mid x_1 = 0, x_2 = 0] = \beta_0.$$

Therefore, the mean difference in *heart rate* (Running minus Swimming) is

$$\mathbb{E}[y \mid \text{Running}] - \mathbb{E}[y \mid \text{Swimming}] \;\; = \;\; (\beta_0 + \beta_1) - (\beta_0) \;\; = \;\; \boxed{\beta_1}.$$