

# Midterm Exam Solutions

**Statistical Methods(통계적방법론)**

**2025 2nd semester**

Section(교번): <u>A1</u> Cadet Number(교번): _____ Name(성명): _____ Score(점수): _____
---

- All solutions must include a detailed step-by-step explanation.
- If an answer has more than four decimal places, round to the **fourth decimal place**.

1. Consider the following linear regression model.[25 points]

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

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

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

Solution: Differentiate and set to zero:

$$\begin{aligned} \frac{dS}{d\beta} &= -2 \sum_{i=1}^n (1 + x_i) (y_i - \hat{\beta} - \hat{\beta} x_i) = 0 & (1) \\ \Rightarrow \hat{\beta} &= \frac{\sum_{i=1}^n (1 + x_i) y_i}{\sum_{i=1}^n (1 + x_i)^2}. \end{aligned}$$

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

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

Solution:

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

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

Solution: By linearity of expectation and  $E(\varepsilon_i) = 0$ ,

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

(4) Show that  $Var(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^n (1 + x_i)^2}$ .

Solution: From the independence of  $y_i$  and  $Var(y_i) = \sigma^2$ ,

$$Var(\hat{\beta}) = \frac{\sum_{i=1}^n (1 + x_i)^2 Var(y_i)}{(\sum_{i=1}^n (1 + x_i)^2)^2} = \frac{\sum_{i=1}^n (1 + x_i)^2 \sigma^2}{(\sum_{i=1}^n (1 + x_i)^2)^2} = \boxed{\frac{\sigma^2}{\sum_{i=1}^n (1 + x_i)^2}}.$$

# Midterm Exam Solutions

**Statistical Methods(통계적방법론)**

**2025 2nd semester**

Section(교반): A1 Cadet Number(교번): \_\_\_\_\_ Name(성명): \_\_\_\_\_ Score(점수): \_\_\_\_\_

2. Suppose we collect data for a group of Major League Baseball players with variables  $X_1$ : number of hits,  $X_2$ : division indicator ( $X_2 = 1$  for West,  $X_2 = 0$  for East), and  $Y$ : salary. We fit a linear regression model in R. [25 points]

$$\text{Model A: } y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i,$$

```
> lm.fit = lm(Salary ~ Hits + Division, data = Hitters)
> coef(summary(lm.fit))
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   149.231    70.8751   2.106 3.620e-02
Hits           4.255     0.5505   7.728 2.383e-13
DivisionWest -141.476    49.6015  -2.852 4.690e-03
```

(1) Predict the average salary of a player with 140 hits who plays in the East division.

Solution: Substituting  $x_1 = 140$ ,  $x_2 = 0$ ,

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 = 149.231 + 4.255 \times 140 - 141.476 \times 0 = \boxed{744.931}.$$

(2) Estimate the mean difference in salary between West and East (West - East).

Solution: The conditional means for each division are:

$$\mathbb{E}[y \mid x_1, \text{West}] = \mathbb{E}[y \mid x_1, x_2 = 1] = \beta_0 + \beta_1 x_1 + \beta_2, \quad \mathbb{E}[y \mid x_1, \text{East}] = \mathbb{E}[y \mid x_1, x_2 = 0] = \beta_0 + \beta_1 x_1.$$

Therefore, the mean difference of *salary* (west minus east) is

$$\mathbb{E}[y \mid x_1, \text{West}] - \mathbb{E}[y \mid x_1, \text{East}] = (\beta_0 + \beta_1 x_1 + \beta_2) - (\beta_0 + \beta_1 x_1) = \beta_2 \quad \therefore \hat{\beta}_2 = \boxed{-141.476}.$$

(3) Is the number of hits ( $X_1$ ) a significant predictor of salary ( $Y$ )? Provide the associated  $p$ -value.

Solution: We test the hypotheses  $H_0 : \beta_1 = 0$  versus  $H_1 : \beta_1 \neq 0$ . From the table, the  $p$ -value is  $2.383 \times 10^{-13}$ . Since the  $p$ -value is small, we reject  $H_0$ . The number of hits is a statistically significant predictor of salary.

(4) Consider a linear model with an interaction term:  $x_{i3} = x_{i1}x_{i2}$ . Answer the following questions.

$$\text{Model B: } y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i,$$

```
> lm.fit2 = lm(Salary ~ Hits + Division + Hits * Division, data = Hitters)
> coef(summary(lm.fit2))
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    35.927    89.9013   0.3996 6.898e-01
Hits            5.269     0.7414   7.1075 1.146e-11
DivisionWest   98.370   128.0819   0.7680 4.432e-01
Hits:DivisionWest -1.454     1.7890  -0.8125 0.4184937
```

(4-A) Compare the residual sum of squares(RSS) of Model A and the RSS of Model B.

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.

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

(4-B) Is the interaction term ( $X_3$ ) a significant predictor of salary ( $Y$ )? Provide the associated  $p$ -value. Which model would you prefer based on this  $p$ -value: Model A or Model B?

Solution: We test the hypotheses  $H_0 : \beta_3 = 0$  versus  $H_1 : \beta_3 \neq 0$ . From the table, the  $p$ -value is 0.4184937. Since this is larger than 0.05, the interaction is not statistically significant. Therefore, we prefer the simpler Model A.

# Midterm Exam Solutions

Statistical Methods(통계적방법론)

2025 2nd semester

Section(교반): A1 Cadet Number(교번): \_\_\_\_\_ Name(성명): \_\_\_\_\_ Score(점수): \_\_\_\_\_

3. Suppose that

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X.$$

Answer the following questions. [15 points]

(1) Show that

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}.$$

Solution: If  $\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$ , then  $\frac{p(X)}{1-p(X)} = e^{\beta_0 + \beta_1 X}$  and

$$\frac{1-p(X)}{p(X)} = \frac{1}{p(X)} - 1 = e^{-(\beta_0 + \beta_1 X)}.$$

It follows that

$$p(X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}} = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}.$$

(2) Show that the odds ratio(OR) satisfies

$$\frac{p(1)/(1-p(1))}{p(0)/(1-p(0))} = e^{\beta_1}.$$

Solution: Since  $\frac{p(X)}{1-p(X)} = e^{\beta_0 + \beta_1 X}$ ,

$$\frac{p(1)}{1-p(1)} = e^{\beta_0 + \beta_1}, \quad \frac{p(0)}{1-p(0)} = e^{\beta_0}.$$

It follows that

$$\frac{p(1)/(1-p(1))}{p(0)/(1-p(0))} = \frac{e^{\beta_0 + \beta_1}}{e^{\beta_0}} = e^{\beta_1}.$$

# Midterm Exam Solutions

**Statistical Methods(통계적방법론)**

**2025 2nd semester**

Section(교반): A1 Cadet Number(교번): \_\_\_\_\_ Name(성명): \_\_\_\_\_ Score(점수): \_\_\_\_\_

4. Suppose we collect data for a group of kinesiology students with variables  $X_1$  = weekly training hours in the gym,  $X_2$  = protein supplement use ( $X_2 = 1$  if yes,  $X_2 = 0$  if no), and  $Y$  = pass the push-up test standard. We fit a logistic regression and produce the following table. [25 points]

Coefficient	Estimate	Std. Err.	Z-statistic	P-value
<b>Intercept</b>	-5.0000	1.4000	-3.57	0.0004
<b>training hours (<math>X_1</math>)</b>	0.1000	0.0250	4.00	$6.34 \times 10^{-5}$
<b>supplement use (<math>X_2</math>)</b>	0.7000	0.3000	2.33	0.0199

(1) Estimate the probability that a student who trains for 25 hours per week and uses protein supplements passes the push-up test.

Solution: The estimated probability from the logistic regression model is

$$p(X_1, X_2) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2}}$$

Substituting  $X_1 = 25$ ,  $X_2 = 1$ ,

$$p = \frac{e^{-5 + 0.10 \cdot 25 + 0.7 \cdot 1}}{1 + e^{-5 + 0.10 \cdot 25 + 0.7 \cdot 1}} = \frac{e^{-5 + 2.5 + 0.7}}{1 + e^{-5 + 2.5 + 0.7}} = \frac{e^{-1.8}}{1 + e^{-1.8}} \approx \boxed{0.1419}$$

(2) How many training hours would the student in part (1) need to have a **60%** chance of passing the push-up test?

Solution: For a 60% chance, we need  $p = 0.6$ . This occurs when

$$\frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2}} = 0.6$$

Taking the logit of 0.6 gives

$$\hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 = \log\left(\frac{0.6}{0.4}\right) \approx 0.40547$$

With  $X_2 = 1$ , this gives

$$-5 + 0.10X_1 + 0.7 = 0.40547 \Rightarrow 0.10X_1 - 4.3 = 0.40547$$

Solving yields  $X_1 = \boxed{47.0547}$  hours.

(3) Is weekly training hours ( $X_1$ ) a significant predictor of passing the push-up test ( $Y$ )? Provide the associated  $p$ -value.

Solution: We test the hypotheses  $H_0 : \beta_1 = 0$  versus  $H_1 : \beta_1 \neq 0$ . From the table, the  $p$ -value is  $6.34 \times 10^{-5}$ .

Since the  $p$ -value is very small, we reject  $H_0$  and conclude that training hours ( $X_1$ ) is a significant predictor of passing the push-up test.

5. Read the following statements and choose the most appropriate word to complete the sentence. [10 points]

- (1) Cohort studies and case-control studies are examples of (experimental study / observational study).
- (2) (Smaller / Larger) trees are more interpretable and reduce variance, but may increase bias.
- (3) Classification trees split the data so that the resulting leaf nodes are (less / more) pure.
- (4) (Bagging / Boosting) reduces the variance of a method by bootstrapping samples.
- (5) (Decision tree / Random forest) improves on bagging by choosing a random subset of predictors at each split.

Solution: observational study / Smaller / more / Bagging / Random forest