Midterm Exam

Statistical Methods(통계적방법론) **2025 2nd semester**

Section(교반): _A1_ 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]

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

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

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

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

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

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

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

Midterm Exam

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.

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

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

(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.

(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?

Midterm Exam

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

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

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

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

Midterm Exam

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
Intercept	-5.0000	1.4000	-3.57	0.0004
training hours (X_1)	0.1000	0.0250	4.00	6.34×10^{-5}
supplement use (X_2)	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.

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

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

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.