# STA 35C Statistical Data Science III

## Midterm exam 1

Instructor: Dogyoon Song

**Name:** _____ **Student ID:** _____

**Instructions:** This midterm exam is **closed-book**, except for the permitted note sheet described below. You may bring a pen or pencil, one letter-sized sheet of *hand-written* notes (both sides), and a *non-graphing* calculator. No other materials (e.g., textbooks) are allowed. You have 50 minutes to complete all problems. The **total score is 120 points**, with *up to 10 bonus points available*. Once you receive this exam, please confirm that you have all 8 pages.

- Make sure to clearly write your name and student ID above.
- Present answers succinctly, but include all relevant steps for full credit. Partial credit is possible only if your reasoning is clearly shown and traceable by the grader.
- If necessary, round all numerical answers to three decimal places; you may leave fractions such as 8/55 or logarithms such as log(3/4) unevaluated.
- Bonus problems can be more challenging; consider attempting them after you finish the main problems.

| **Problem** | **Score** |
|---|---|
| Problem 1 | |
| Problem 2 | |
| Problem 3 | |
| Problem 4 | |
| **Total** | |

## Problem 1 (30 points). Warm-up

Answer each part briefly and clearly.

**(a) (12 points)** For each scenario, identify the response $Y$, the predictor(s) $X$, whether the problem is regression or classification, and whether the primary goal is prediction or inference.

(i) A bank wants to predict whether a new loan applicant will default or not, using the applicant's income, balance, and credit history.

(ii) A researcher wants to understand how study hours and lecture attendance are associated with students' exam scores.

**(b) (6 points)** Consider two fair coin tosses with

$$\Omega = \{HH, HT, TH, TT\}, \qquad \text{where} \qquad H = \text{Head and } T = \text{Tail}.$$

Let $X$ = number of Heads. List all the outcomes in the event $\{X = 1\}$ and compute $P(X = 1)$.

**(c) (6 points)** Suppose that we have a fitted regression model for exam score as

$$\widehat{Y} = 62 + 4D + 3X,$$

where $D = 1$ if the student attended more than half of the lectures and $D = 0$ otherwise, and $X$ denotes study hours. Interpret the coefficient 4 for the variable $D$ in context.

**(d) (6 points)** Suppose a logistic regression classifier gives $\hat{p}(x_{\text{new}}) = 0.32$, where $\hat{p}(x)$ is the estimated probability of class 1 at $x$. Using the rule "predict class 1 if $\hat{p}(x) \geq p^*$," what class is predicted if $p^* = 0.5$? What class is predicted if $p^* = 0.2$?

## Problem 2 (30 points). Probability, random variables, and Bayes' rule

A data science team is auditing predictions made by two candidate classification models, Model A and Model B. For each audited batch, the team checks 3 randomly selected predictions and records

$$X = \text{number of incorrect predictions among the 3 audited predictions.}$$

**(a) (10 points)** First suppose the audited batch was produced by Model A. Each prediction is incorrect independently with probability 1/3, so

$$X \mid A \sim \text{Binomial}(3, \tfrac{1}{3}).$$

Compute

$$P(X = 0 \mid A), \qquad P(X \geq 1 \mid A), \qquad \mathbb{E}[X \mid A], \qquad \text{Var}(X \mid A).$$

*Hint:* A Binomial(3, $\frac{1}{3}$) random variable is the sum of three independent Bernoulli($\frac{1}{3}$) random variables.

**(b) (10 points)** For this subproblem only, suppose

$$\mathbb{E}[X] = \frac{2}{3}, \qquad \text{Var}(X) = \frac{5}{9}, \qquad \mathbb{E}[Y] = 4, \qquad \text{Var}(Y) = 5, \qquad \text{Corr}(X, Y) = 0.3,$$

where $Y$ is the processing time, in seconds, for the audited batch. We define a deployment loss score as

$$W = 9X + 2Y + 5.$$

Compute $\mathbb{E}[W]$ and $\text{Var}(W)$.

*Hint:* $\text{Cov}(X, Y) = \text{Corr}(X, Y)\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}$.

**(c) (10 points)** Now suppose that the team does not know which model produced the audited batch, and believes that Model A and Model B are equally likely to have produced the audited batch. If the batch was produced by Model B, each prediction is incorrect independently with probability 1/6, so

$$X \mid B \sim \text{Binomial}(3, \tfrac{1}{6}).$$

Compute

$$P(B \cap \{X = 0\}), \qquad P(X = 0), \qquad P(B \mid X = 0).$$

Interpret $P(B \mid X = 0)$ in one sentence, including why it is larger or smaller than $P(B) = 1/2$.

## Problem 3 (30 points + 5 bonus). Regression

A study examines how students' exam scores are related to study hours and lecture attendance. Let

$$Y = \text{exam score}, \qquad X = \text{study hours}, \qquad D = \begin{cases} 1, & \text{attended lectures,} \\ 0, & \text{did not attend lectures.} \end{cases}$$

**(a) (10 points)** In a simple linear regression of $Y$ on $X$, suppose

$$\bar{x} = 5, \qquad \bar{y} = 76, \qquad \sum_i (x_i - \bar{x})^2 = 40, \qquad \sum_i (x_i - \bar{x})(y_i - \bar{y}) = 120.$$

Compute $\hat{\beta}_1$ and $\hat{\beta}_0$ in the model

$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

Then predict the exam score for a student who studied 8 hours.

**(b) (10 points)** Now suppose we fit the multiple regression model

$$Y = \beta_0 + \beta_1 X + \beta_2 D + \varepsilon,$$

and obtained a fitted coefficient table as follows:

| **Coefficient** | **Estimate** | **Std. Error** | ***t*-statistic** |
|---|---|---|---|
| $X$ | 2.4 | 0.6 | ? |
| $D$ | 5.0 | 3.5 | ? |

State which coefficients are statistically significant at the 0.05 level, using the table below.

| $t$ | 0.5 | 1.0 | 1.5 | 2.0 | 2.5 | 3.0 | 3.5 |
|---|---|---|---|---|---|---|---|
| Approx. $p$-value | 0.6171 | 0.3173 | 0.1336 | 0.0455 | 0.0124 | 0.0027 | 0.000465 |

Thereafter, interpret the estimates 2.4 and 5.0 in context, including what the significance results suggest.

**(c) (10 points)** Three candidate models are fit to the same training dataset and evaluated on the same separate test dataset:

| **Model** | **Predictors** | **Training $R^2$** | **Test MSE** |
|---|---|---|---|
| $A$ | $X$ | 0.58 | 3.7 |
| $B$ | $X, D$ | 0.64 | 3.2 |
| $C$ | $X, D, X^2$ | 0.66 | 4.5 |

Which model would you choose for prediction? Briefly justify. Also explain why Model C can have the largest training $R^2$ and the largest test MSE at the same time.

**(d*) (5 bonus points)** Suppose we fit

$$Y = \beta_0 + \beta_1 D + \beta_2 X + \beta_3 (D \times X) + \varepsilon,$$

and obtain $\hat{\beta}_0 = 35$, $\hat{\beta}_1 = 5$, $\hat{\beta}_2 = 1$ and $\hat{\beta}_3 = -0.5$. Find the value of $x$ at which the estimated attendance effect is zero (if it exists), and briefly interpret what $\hat{\beta}_3 = -0.5$ means in context.

## Problem 4 (30 points + 5 bonus). Classification

A bank wants to predict whether a loan applicant will default. Let

$$Y = \begin{cases} 1, & \text{default,} \\ 0, & \text{no default.} \end{cases}$$

Let $X_1$ and $X_2$ be two standardized financial predictors.

**(a) (10 points)** A fitted logistic regression model is

$$\log\left(\frac{\hat{p}(x)}{1-\hat{p}(x)}\right) = -2 + x_1 + x_2, \qquad \text{where} \qquad \hat{p}(x) = \widehat{P}(Y = 1 \mid X = x).$$

For $x_{\text{new}} = (1, 2)$, compute $\hat{p}(x_{\text{new}})$ and predict $\hat{y}_{\text{new}}$ using $p^* = 0.5$. Then write the decision boundary for $p^* = 0.5$, and state which side of the boundary is classified as default.

*Hint:* $e \approx 2.718$.

**(b) (10 points)** On a test set of 100 applicants, the classifier gives the following confusion matrix:

| | Predicted 1 | Predicted 0 |
|---|---|---|
| $Y = 1$ | 18 | 7 |
| $Y = 0$ | 10 | 65 |

Compute the error rate, true positive rate (TPR), false positive rate (FPR), and false negative rate (FNR). If the bank lowers the threshold from $p^* = 0.5$ to $p^* = 0.2$, what would you generally expect to happen to FPR and FNR? Can we also determine whether the overall error rate increases or decreases? Briefly explain.

**(c) (10 points)** The bank also considers a one-dimensional LDA classifier using a proprietary risk score $Z$. Suppose the bank collected the following data, written as $(z, y)$:

$$(0,0), \quad (1,0), \quad (2,0), \quad (2,0), \quad (3,0), \quad (4,0), \quad (4,1), \quad (6,1).$$

Estimate the class priors and class means using the data; suppose the pooled variance estimate is already given as $\hat{\sigma}^2 = 2$. Using these estimates, compute the linear discriminant functions for Class 0 and Class 1. Find the LDA cutoff $z^*$, and classify a new applicant with $z = 4.0$ using the rule you derive.

*Hint:* $\log(3) \approx 1.099$ and the PDF of Gaussian distribution with mean $\mu$ and variance $\sigma^2$ is $f(z) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(z-\mu)^2}{2\sigma^2}\right\}$.

**(d*) (5 bonus points)** In the same dataset as in part **(b)**, there are 25 actual defaults and 75 actual non-defaults. For several classification thresholds $p^*$, suppose the classifier produces the following numbers of true positives and false positives:

| $p^*$ | 1.0 | 0.8 | 0.6 | 0.4 | 0.2 | 0 |
|---|---|---|---|---|---|---|
| $TP$ | 0 | 10 | 15 | 20 | 22 | 25 |
| $FP$ | 0 | 2 | 6 | 12 | 35 | 75 |

Compute and sketch the ROC points corresponding to these $p^*$ values in the (FPR, TPR)-plane, with the axes clearly labeled. If the bank requires FPR $\leq 0.20$, which threshold among these six values should it choose to maximize TPR?