STA 35C Statistical Data Science III

Midterm exam 1

Instructor: Dogyoon Song

Name: ____________________ **Student ID:** ____________________

Instructions: This midterm exam is a **closed-book** exam. 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 8 bonus points available*. Once you receive this exam problem set, please confirm you have all 12 pages.

- Make sure to clearly write your name and 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.
- If necessary, round all numerical answers to three decimal places.
- 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 (20 points in total + 2 bonus points).

A manufacturing company ships products in boxes of **two** items each. We define two random variables:

$$X = \text{number of defective items in a box of size 2},$$
$$Y = \text{time (hours) for the final inspection of the box.}$$

(a) (5 points) Suppose each of the 2 items has a $\frac{1}{3}$ chance of being defective, independently. Then

$$X \sim \text{Binomial}(2, \tfrac{1}{3}).$$

Compute $\mathbb{E}[X]$ and $\text{Var}(X)$.
(*Hint:* use the PMF $p_X(x) = \binom{2}{x}\left(\frac{1}{3}\right)^x\left(\frac{2}{3}\right)^{2-x}$, or let $X = X_1 + X_2$ where X_1, X_2 are i.i.d. Bernoulli($\frac{1}{3}$).)

(b) (5 points) Consider a cost variable

$$W \;=\; X \;+\; 2Y \;+\; 2,$$

where X is a per-item defect penalty, Y is an hourly inspection cost, and 2 is a fixed operating cost. In reality, more defects might delay inspection. Suppose $\mathbb{E}[Y] = \text{Var}(Y) = 9$, and the correlation coefficient $\rho \coloneqq \text{corr}(X, Y) = 0.3$.

Compute $\mathbb{E}[W]$ and $\text{Var}(W)$. (*Hint:* $\text{Cov}(X,Y) = \rho\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}$.)

(c) (10 points total) Each box (with 2 items) is produced by **Factory A** or **Factory B**, each with *equal probability*.

- If it is from **Factory A**, each item is defective with probability $\frac{1}{3}$.
- If it is from **Factory B**, each item is defective with probability $\frac{1}{10}$.

(i) (5 points) What is the probability that a randomly chosen box is from Factory A *and* has exactly one defective item ($X = 1$)?

(ii) (5 points) You observe $X = 1$ defective item in a newly received box. What is the posterior probability that this box came from **Factory A**?

(iii*) (*2 bonus points) Now assume there are *four* factories: A, B, C, and D, with unknown probabilities.

- Factory C makes all items defective (probability of defect = 1).
- Factory D makes all items non-defective (probability of defect = 0).

You observe $X = 1$ defective item in a new box. What is the posterior probability that the box is from **Factory D**?

Problem 2 (25 points in total).

(a) **(12 points total)** For each of the following four scenarios:

- Identify the predictor(s) X and the response Y.
- State whether it is a *regression* or a *classification* problem.
- Briefly discuss whether the primary goal is *prediction* or *inference* (and why).

(i) **(3 points)** A nutritionist wants to forecast a patient's daily protein intake (grams) from age, weight, and exercise routine.

(ii) **(3 points)** A market analyst wants to predict which of three smartphone plans (A, B, or C) a new customer will choose, based on browsing habits.

(iii) **(3 points)** An admissions officer wants to estimate a student's final exam score from prior homework grades, aiming to see which assignments are most influential.

(iv) **(3 points)** A real estate agent wants to assess how each factor (location, bedrooms, floor area, building age) influences monthly rent, in order to identify the most significant effect.

(b) (13 points total) You have two models for predicting a numeric outcome:

- Model (1): $f_1 : x \to y$, a simple, interpretable linear model.
- Model (2): $f_2 : x \to y$, a more complex "black-box" model (e.g., a deep learning method).

Assume you initially have only *training data* $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$.

(i) (5 points) How would you compare their predictive performance? Specify what metric(s) you would use, how you would compute it using f_1, f_2, and $\mathcal{D}$, and how you would decide which model is better (e.g., lower value is better or worse).

(ii) (4 points) If Model (2) outperforms Model (1) *on training data*, should you always choose Model (2)? If yes, justify. If no, give one reason you might still prefer Model (1).

(iii) (4 points) Now suppose you test each model on a separate test dataset. If Model (2) consistently outperforms Model (1) *on both training and test data*, should you *always* choose Model (2)? If yes, justify. If no, provide one reason you might still prefer Model (1).

Problem 3 (40 points in total + 2 bonus points).

You are studying how long it takes participants to solve a particular logic puzzle. Let:

Y : time (in minutes) to finish the puzzle,

X_1 : indicator (0 or 1) if the participant has at least 2 years of prior puzzle-solving experience,

X_2 : short-term memory test score.

(a) (15 points total) Suppose you have two separate *simple* linear regression models:

$$\text{Model A:} \quad Y = 11.2 - 5\,X_1,$$
$$\text{Model B:} \quad Y = 10 - 0.6\,X_2.$$

(i) (5 points) For a new participant with $(X_1, X_2) = (1, 7)$, compute $\hat{Y}_A$ (Model A) and $\hat{Y}_B$ (Model B).

(ii) (5 points) Model A has $R^2 = 0.50$, Model B has $R^2 = 0.64$. Which model gives a better *predictive* fit? Also, explain what R^2 represents about Y's variation or residuals in one or two sentences.

(iii) (5 points) The multiple regression model including both X_1 and X_2 yields $R^2 = 0.70$. Does that mean the combined model is always "better"? Give one justification why it *may* be better and one reason it might *not* be strictly better.

(b) (15 points total) Suppose you fit two models: (1) a simple model $Y \sim X_1$, (2) a multiple model $Y \sim X_1 + X_2$. You have the following partial outputs, where t-statistics and p-values are not included:

Model (simple, $Y \sim X_1$):

Coefficient	Estimate	Std. Error	t-statistic
X_1	-5	1.67	?

Model (multiple, $Y \sim X_1 + X_2$):

Coefficient	Estimate	Std. Error	t-statistic
X_1	-1.6	1.6	?

Here are approximate two-sided p-values for standard normal z (or large-sample t) at selected points:

z	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

(i) (5 points) Compute the t-statistics for X_1 in the simple model and also in the multiple model, then decide if each is statistically significant at the 5% level.

(ii) (5 points) Interpret the coefficient for X_1 in the *simple* vs. the *multiple* model, in the context of puzzle-solving time. Provide your explanation in *one or two sentences* each.

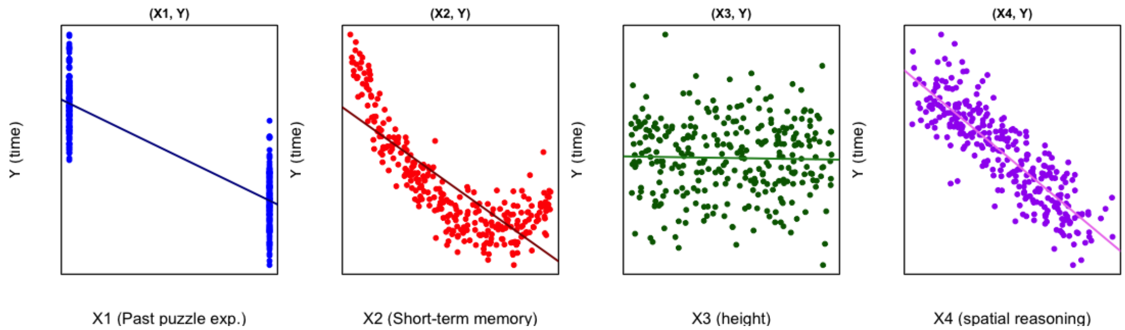
(iii) (5 points) Explain how the infuence of X_1 (puzzle-solving experience) on Y (puzzle-solving time) might be *confounded* by X_2 (short-term memory score). State why controlling for X_2 could change X_1's estimated effect in *one or two sentences.*

(iv*) (*2 bonus points) Suppose you suspect different slopes of Y against X_2 for $X_1 = 0$ vs. $X_1 = 1$, and include an interaction term $X_1 \times X_2$ in the regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \times X_2).$$

If $\hat{\beta}_1 = -7$, interpret this value in *one or two sentences.*

(c) (10 points total) You now consider X_3 (height) and X_4 (spatial reasoning test score) as additional predictors. Below is a figure of four scatterplots (Y vs. X_i) and the correlation matrix for (X_1, X_2, X_3, X_4)



	X_1	X_2	X_3	X_4
X_1	1.00	-0.06	-0.04	-0.03
X_2		1.00	0.07	0.94
X_3			1.00	0.01
X_4				1.00

(i) (4 points) If you only have X_1 and X_2, would you modify or add anything in your multiple linear regression model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ to improve the predictive power for Y? Explain in *one or two sentences.*

(ii) (3 points) If you have X_3 in addition to X_1, X_2, would you add X_3 to the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$? Provide your reasoning in *one or two sentences.*

(iii) (3 points) If you have X_4 in addition to X_1, X_2, would you add X_4 to the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$? Provide your reasoning in *one or two sentences.*

Problem 4 (35 points in total + 4 bonus points).

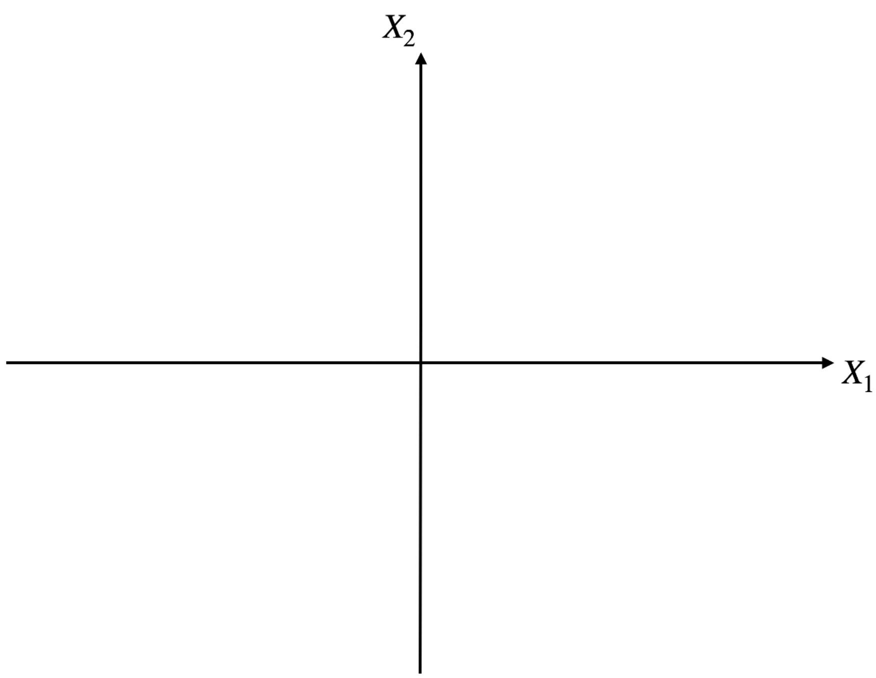
(a) **(20 points total)** Consider a two-dimensional logistic model for binary classification:

$$\log \frac{p(X)}{1-p(X)} = \beta_0 + \beta_1 X_1 + \beta_2 X_2, \qquad \text{where} \quad p(X) = \Pr[Y = 1 \mid X].$$

Given $X = x$, you predict $Y = 1$ if and only if $p(x) \geq p^*$. Suppose the estimated coefficients are $\hat{\beta}_0 = -2,\ \hat{\beta}_1 = -1,\ \hat{\beta}_2 = 2$.

(i) **(5 points)** You have a new test point $x_{\text{test}} = (x_1, x_2) = (1, 1)$. Compute $\hat{p}(x_{\text{test}})$ and decide $\hat{y}_{\text{test}} = 1$ or 0 for $p^* = 0.5$.

(ii) **(5 points)** Draw the decision boundary (for $p^* = 0.5$) in the figure below, clearly marking intercepts with numbers specified. Indicate where your logistic model predicts $\hat{Y} = 1$ (e.g., shade or label "+").



(iii*) **(*2 bonus points)** Describe how the decision boundary changes if $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) = (-1, -1, 1)$ instead of $(-2, -1, 2)$. If possible, draw this on the figure above with a different style (e.g., dashed line).

(iv) (5 points) Suppose you obtained the following confusion matrix from 100 data points:

	Pred = 1	Pred = 0
$Y = 1$	35	5
$Y = 0$	15	45

Compute the true positive rate (TPR/sensitivity) and false positive rate (FPR or $1 -$ specificity).

(v) (5 points) Suppose you lower p^* from 0.5 to 0.1. Would you expect more, fewer, or the same number of false positives and false negatives? Briefly explain your answers in *one or two sentences*.

(b) (15 points total) You have caught 5 crabs of two different species and recorded their weights (pounds):

$$\text{Species A: } x = \{1.5,\ 2.5\}, \quad \text{Species B: } x = \{2.0,\ 3.0,\ 4.0\}.$$

You want to classify a new crab by weight, assuming each species' weight follows a normal distribution with potentially different means but the same variance. The PDF of a normal with mean μ and variance σ^2 is

$$f(x) = \frac{1}{\sqrt{2\pi\,\sigma^2}} \exp\Big(-\frac{(x-\mu)^2}{2\,\sigma^2}\Big).$$

(i) (4 points) Compute the sample means $\bar{x}_A, \bar{x}_B$ and the pooled standard deviation s.

(ii) (4 points) Write down the linear discriminant functions $\delta_A(x)$ and $\delta_B(x)$.
(*Hint:* $\log(\frac{2}{5}) \approx -0.916,\ \log(\frac{3}{5}) \approx -0.511$.)

(iii) (4 points) Suppose you observe $x_{\text{new}} = 2.5$. Which species would you predict based on your linear discriminant? Show your reasoning briefly.

(iv) (3 points) Would this prediction change if you gathered 4 more data points for Species A (with the same mean and variance)? Explain why or why not.

(v*) (*2 bonus points) Suppose you do *not* want to miss any crabs of Species B (e.g., for taste or invasive reasons). You decide to predict A only if $\Pr(Y = A \mid X) \geq p^*$ with $p^* > 0.5$ (e.g. $p^* = 0.9$). How does this change the LDA decision rule in terms of $\delta_A(x)$ and $\delta_B(x)$? Specifically, state your new decision rule and boundary, and apply it to $x'_{\text{new}} = 2.0$.