STA 35C Statistical Data Science III

(Mock exam for midterm 1)

Instructor: Dogyoon Song

Name: ____________________ **Student ID:** ____________________

Instructions: This mock exam is designed to illustrate the approximate structure, length, and style of Midterm 1. However, the actual Midterm 1 may differ in content or format from this practice exam.

- Make sure to clearly write your name and ID above.
- The actual Midterm 1 will be a **closed-book** exam. You may bring only a pen/pencil, one letter-sized sheet of handwritten notes (both sides), and a non-graphing calculator.
- You have 50 minutes to complete all problems. The total score is 100 points.
- Show all relevant steps in your solutions for full credit. Partial credit is possible only if your reasoning is clearly presented, and can be easily traced by the grader.
- If necessary, please round all numerical answers to two decimal places.

Problem	**Score**
Problem 1	
Problem 2	
Problem 3	
Problem 4	
Problem 5	
Total	

Problem 1 (20 points in total).

(a) (4 points) Let X be a random variable with pdf f_X defined by

$$f_X(x) = \begin{cases} 3x^2, & \text{if } 0 \leq x \leq 1, \\ 0, & \text{otherwise.} \end{cases}$$

(i) Compute $\mathbb{E}[X]$.

(ii) Compute $\mathrm{Var}(X)$.

(b) (8 points) Suppose X and Y are two random variables with

$$\mathbb{E}[X] = 2, \quad \mathrm{Var}(X) = 4, \quad \mathbb{E}[Y] = 5, \quad \mathrm{Var}(Y) = 1.$$

(i) If X and Y are *independent*, compute $\mathbb{E}[X + 2Y]$ and $\mathrm{Var}(X + 2Y)$.

(ii) Now assume $\mathrm{corr}(X, Y) = 0.5$. Recompute $\mathbb{E}[X + 2Y]$ and $\mathrm{Var}(X + 2Y)$.

(iii) Compare these two results, and briefly comment on why knowledge of correlation matters.

(c) **(8 points)** Let S denote the event "email is spam." A filter flags an email as spam (F) if it detects suspicious terms.

(i) Suppose $\Pr(S) = 0.05$, $\Pr(F \mid S) = 0.90$, and $\Pr(\text{not } F \mid \text{not } S) = 0.95$. If an email is flagged, what is $\Pr(S \mid F)$, the conditional probability of the flagged email being a spam?

(ii) Why might this probability be lower than one would intuitively expect (say, at a similar level to $\Pr(F \mid S) = 0.90$), despite the filter's seemingly good performance?

Problem 2 (15 points in total).

(a) (6 points) Prediction vs. Inference:

(i) In your own words, succinctly differentiate "prediction" and "inference" in building a statistical model.

(ii) Give an example where *prediction* accuracy is crucial, and another where *inference* is more important.

(b) (6 points) Regression vs. Classification: For each scenario, decide if it is a regression problem or a classification problem. Briefly justify your decision.

(i) Predicting a patient's blood pressure.

(ii) Predicting whether a customer will default on a loan.

(iii) Forecasting the number of phone calls to a hotline.

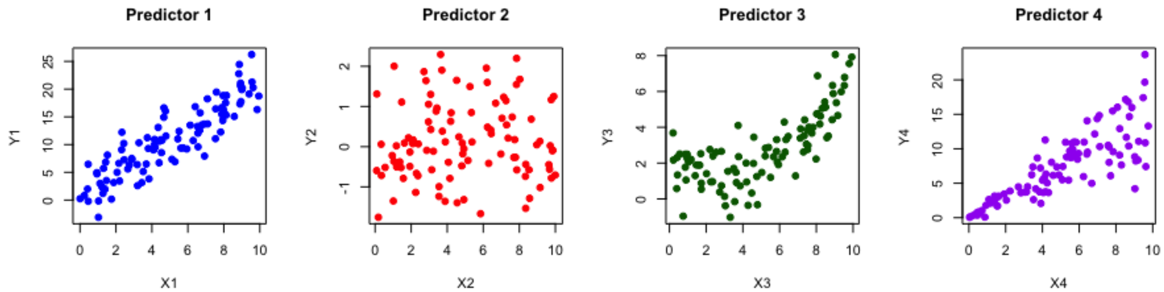
(iv) Predicting which of three cell phone plans (basic, advanced, or unlimited) a new user will choose.

(c) (3 points) Parametric vs. Nonparametric: Suppose you have a small dataset but a strong theoretical reason to expect a linear relationship. Would you prefer a parametric linear model or a more flexible nonparametric method (e.g., kNN)? Name one advantage and one drawback of your choice.

Problem 3 (20 points in total).

Consider a response variable Y and four possible predictors X_1, X_2, X_3, X_4. You are exploring these relationships with plots and basic statistical measures.

(a) (9 points) You have four scatterplots:



(i) For what variables does the linear assumption look reasonable?

(ii) In each plot, do the errors (not prediction residuals) appear to be independent of the predictors?

(iii) What can you comment on the variances of the errors by comparing the plots? (e.g., one looks larger than another, seems to depend on X_i, etc.)

(b) (5 points) Suppose that the correlation matrix among X_1, X_2, X_3, X_4 in part **(a)** is given as follows. Choose two best predictor variables to include in a linear regression model for Y. Explain your choice.

	X_1	X_2	X_3	X_4
X_1	1.00	0.08	-0.95	-0.06
X_2		1.00	0.11	0.27
X_3			1.00	0.18
X_4				1.00

(c) (6 points) Three candidate models (f_1, f_2, f_3) yield:

Model	f_1	f_2	f_3
Training MSE	4.0	3.2	2.1
Test MSE	3.2	2.8	5.4

(i) If only training data were available, which model would you choose?

(ii) Does that choice remain optimal once you see the test MSE? Why or why not?

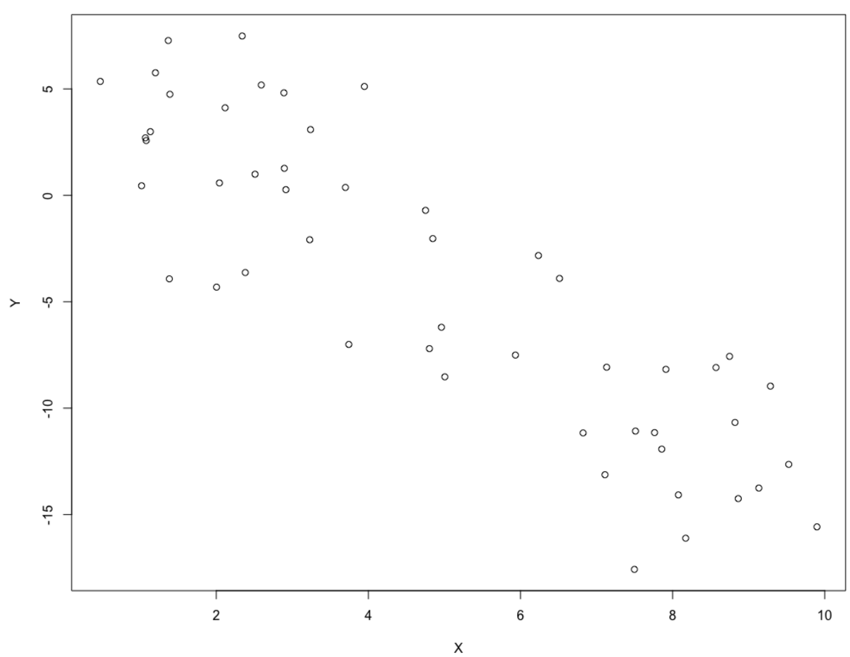
Problem 4 (20 points in total).

(a) (6 points) You fit a simple linear regression of Y on X:

$$\hat{Y} = -2.0 + 1.5\,X.$$

(i) Suppose $X = 6$. What is the predicted value of Y?

(ii) Now suppose that you have a fresh training dataset as below. Sketch the regression line you would obtain by least squares on this scatter plot of (X, Y) data. Also, visualize how you predict value of Y at $X = 6$ using the regression line.



(b) (6 points) A partial regression output is given:

Coefficient	**Estimate**	**Std. Error**	*t*-statistic	*p*-value
Intercept	-2.0	0.5	??	??
X	$+1.5$	0.4	??	??

For reference, here are approximate two-sided p-values for standard normal z (or large-sample t) at several points:

z	Approx. p-value	z	Approx. p-value
0.5	0.6171	3.0	0.0027
1.0	0.3173	3.5	0.000465
1.5	0.1336	4.0	6.3×10^{-5}
2.0	0.0455	4.5	6.8×10^{-6}
2.5	0.0124	5.0	5.7×10^{-7}

(i) Compute the t-statistic and the p-value for the two regression coefficient above.

(ii) Interpret the slope 1.5. Is X significantly associated with Y at the 5% level? Briefly explain.

(c) (8 points) You then add a second predictor, X_2 (e.g., competitor's marketing spend). In this new two-predictor model, the estimated slope for X_1 changes sign from $+1.5$ to -0.2.

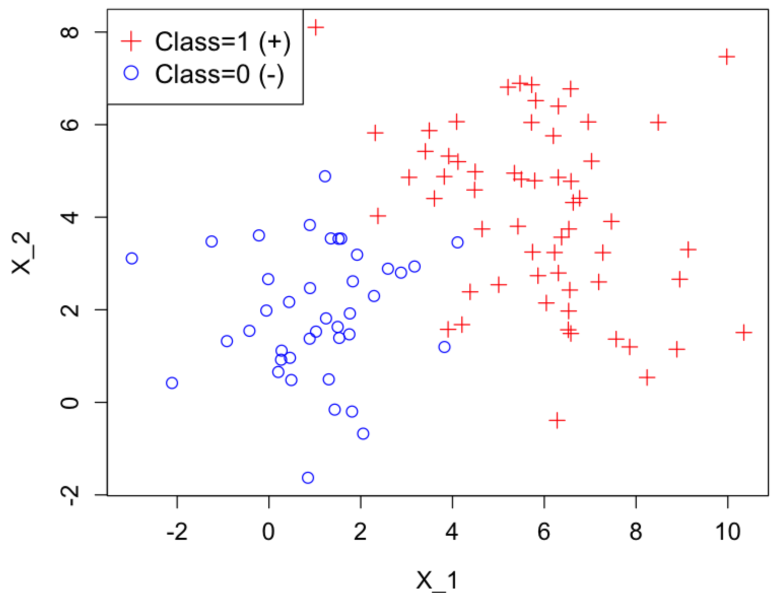
(i) How can adding X_2 cause the direction of X's effect to reverse?

(ii) Explain how you would interpret the new slope -0.2 in a two-predictor model.

(iii) What does this reveal about relationships among X, X_2, and Y?

Problem 5 (25 points in total).

A dataset of website visitors is labeled $Y = 1$ (subscriber) or $Y = 0$ (non-subscriber). Two predictors, X_1 and X_2, measure user behavior (e.g., time on page, pages viewed).



(a) **(6 points)** Suppose you fit a logistic model and decide $Y = 1$ if $\hat{p}(X_1, X_2) \geq p^*$. The figure above shows the dataset in the (X_1, X_2) plane.

(i) On the scatterplot, sketch the decision boundary assuming p^* is appropriately chosen (e.g., 0.5).

(ii) Mark the points $A = (6, 2)$ and $B = (1, 3)$ on the plot, and predict whether $Y = 1$ or $Y = 0$ for each.

(b) **(6 points)** Define a *false positive* and a *false negative* in this context. Identify one example datapoint of each in the scatterplot above. (Otherwise, construct a hypothetical point that would illustrate each.)

(c) (6 points) Let TPR = True Positive Rate, FPR = False Positive Rate. Suppose your current model yields the confusion matrix:

	Pred = 1	Pred = 0
$Y = 1$	54	6
$Y = 0$	5	35

(i) Compute TPR and FPR from this confusion matrix.

(ii) If you lower the decision threshold from 0.5 to 0.1, do you expect TPR to increase or decrease? What about FPR?

(d) (7 points) Suppose a false negative (missing a potential subscriber) is more costly than a false positive.

(i) Would you keep the cutoff at $p^* = 0.5$ or choose another value? Explain briefly.

(ii) If you want the false negative rate to stay below 0.1, how would you pick p^* using the ROC curve? Describe it verbally, or draw a hypothetical ROC curve and mark the operating point you would choose along with a brief justification.