

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.
- For Problem 1, answer as directed and explain only when requested. For all other problems, present answers succinctly and include the relevant steps. Partial credit is possible only if your reasoning is clearly shown and traceable by the grader.
- If necessary, round numerical answers to three decimal places; fractions, logarithms, and exponentials may be left 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 (40 points). Basic concepts

Answer each part as directed. Give an explanation only when requested, within the stated sentence limit.

(a) (4 points) For each question, write **Inference** or **Prediction**. (2 points each)

- How is study time associated with mean exam score?
- What exam score should we predict for a new student?

(b) (4 points) Suppose $\mathbb{E}[X] = 2$ and $\mathbb{E}[Y] = -1$. Find $\mathbb{E}[3X - 2Y]$.

(c) (4 points) Assume X and Y have finite variances. Is the following statement **true** or **false** (2 points)?
Give a reason in **one sentence** (2 points).

$$X \text{ and } Y \text{ are independent} \iff \text{Cov}(X, Y) = 0.$$

(d) (4 points) Which objective does ordinary least squares minimize? Circle one.

$$\text{A. } \sum_i (y_i - \hat{y}_i) \qquad \text{B. } \sum_i (y_i - \hat{y}_i)^2 \qquad \text{C. } \sum_i |y_i - \hat{y}_i| \qquad \text{D. } \max_i |y_i - \hat{y}_i|$$

(e) (4 points) A regression predicts exam score from weekly study time. At 5 hours per week, it reports a 95% confidence interval $[75, 83]$ and a 95% prediction interval $[60, 98]$. Interpret each interval in context in **one sentence**, identifying the quantity it concerns. (2 points per interval)

- Confidence interval:
- Prediction interval:

- (f) (4 points) Suppose that transportation mode takes one of four values: car, bus, bicycle, or walking. For a regression of commute time on transportation mode with an intercept and one baseline mode, give the **number of indicator variables needed** and a **one-sentence justification**.
- (g) (4 points) A logistic regression assigns an email an estimated spam probability of 0.80. In **one sentence**, **interpret this estimate** and state **whether it guarantees that the email is spam**.
- (h) (4 points) A classifier estimates the feature distribution within each class and the class proportions, then applies Bayes' rule. Is this a **generative** or **discriminative** approach?
- (i) (4 points) A fraud detector labels a transaction as fraudulent (class 1) or legitimate (class 0). For each scenario below, select and write **one correct label** from **True positive**, **True negative**, **False positive**, or **False negative**. (2 points each)
- A legitimate transaction is labeled fraudulent.
 - A fraudulent transaction is labeled legitimate.
- (j) (4 points) In the logistic model below, the *odds* are the ratio $p(x)/(1 - p(x))$:

$$\log\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + (\log 2)x.$$

For a one-unit increase in x , state whether the odds change by a **fixed additive amount** or a **fixed multiplicative factor**, and give that amount or factor.

Problem 2 (20 points). Learning a coin's type

A box contains equal numbers of ***H*-type** and ***L*-type** coins. An *H*-type coin has probability $3/4$ of heads; an *L*-type coin has probability $1/4$ of heads. Choose one coin uniformly at random and toss **that same coin twice**. The tosses are independent **conditional on the coin's type**.

Let $T \in \{H, L\}$ denote the chosen coin's unobserved type, and define

$$A = \{\text{the first toss is heads}\}, \quad B = \{\text{the second toss is heads}\}.$$

(a) (10 points) Given A , **compute** $P(T = H \mid A)$ and **identify** the more likely coin type.

(b) (10 points) State whether A and B are independent **without conditioning on T** in **one sentence**, supported by a calculation comparing $P(B \mid A)$ and $P(B)$.

Problem 3 (30 points + 5 bonus). Study time and exam scores

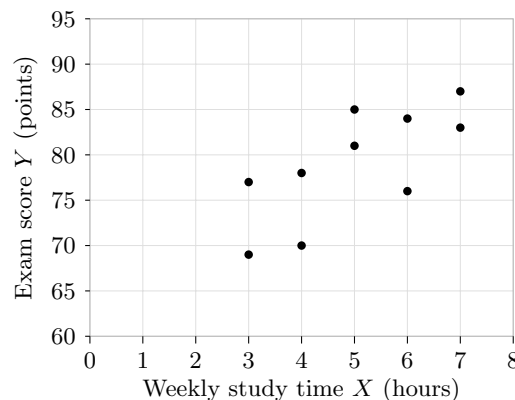
For part (a), use these hypothetical data for ten students. X is weekly study time in hours, and Y is exam score in points.

Student	1	2	3	4	5	6	7	8	9	10
X	3	3	4	5	5	4	6	6	7	7
Y	77	69	78	85	81	70	84	76	87	83

The following summaries are provided; you do not need to recompute them:

$$n = 10, \quad \bar{x} = 5, \quad \bar{y} = 79, \quad S_{xx} = \sum_i (x_i - \bar{x})^2 = 20, \quad S_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y}) = 60.$$

- (a) (10 points) Find and report the least-squares line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ by regressing exam score on study time, and predict $\hat{y}_{\text{test}}$ for a student with $x_{\text{test}} = 4$. Then draw the fitted line and label $(x_{\text{test}}, \hat{y}_{\text{test}})$ on the axes below.



- (b) (10 points) A separate sample of 60 students gives the simple linear regression results below. Assume independent normal errors with mean zero and a common positive variance.

Term	Estimate	Standard error
Intercept	60.0	3.0
Weekly study time	2.5	0.5

Report the **95% confidence interval** for the population slope β_1 and an **interpretation in context** in at most two sentences. (*Hint: Use 2.00 as the approximate 97.5th percentile of the Student t distribution with $60 - 2 = 58$ degrees of freedom.*)

- (c) (10 points) A **separate cohort** of students is used to study an optional study program. Participation was not randomly assigned. Let $D = 1$ for participants and $D = 0$ otherwise. The group averages and a fitted multiple regression model are

Group	Mean study time (hours)	Mean exam score
Participants ($D = 1$)	3	74
Nonparticipants ($D = 0$)	8	82

$$\hat{y} = 50 + 4x + 12D.$$

In **at most two sentences**, state **whether these results establish that the program lowers scores**, justifying your conclusion using the group study times, fitted model, and nonrandom assignment.

- (d*) (5 bonus points) *This bonus is independent of the numerical example above.*

An ordinary least-squares fit with an intercept gives $g(x) = b_0 + b_1x$. The training data contain at least two distinct values of x . For **every training or new input** x , define $z(x) = g(x)$, and refit the same training responses by ordinary least squares using

$$a_0 + a_1x + a_2z(x).$$

- Can adding this predictor change the fitted prediction at any x ?
- What does this imply about the training residual sum of squares and any additional overfitting caused by the new predictor?

Justify with a brief algebraic argument and **at most two sentences** for each.

Problem 4 (30 points + 5 bonus). Reading a spam-filter study

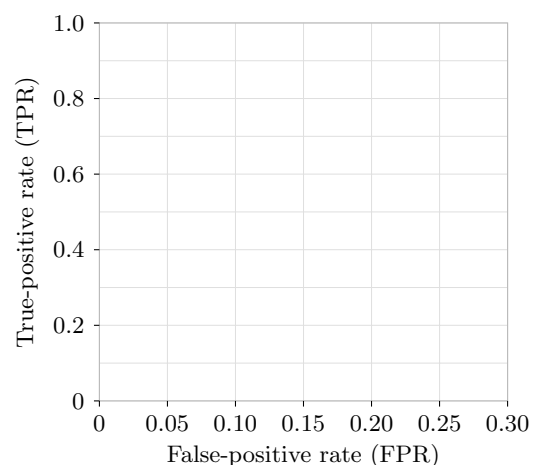
A team predicts whether an email is spam ($Y = 1$) or not spam ($Y = 0$) from features X . It wants probabilities $P(Y = 1 \mid X)$ between 0 and 1, modeled **directly** without specifying feature distributions within classes. An email is flagged when its fitted spam probability is at least p^* . The table compares three thresholds on the **same 1,000-email evaluation set**, with fitted probabilities held fixed.

Threshold p^*	Prediction	Actual spam ($Y = 1$)	Actual not spam ($Y = 0$)
0.2	Flagged as spam	90	180
	Not flagged	10	720
0.5	Flagged as spam	80	90
	Not flagged	20	810
0.8	Flagged as spam	60	18
	Not flagged	40	882

- (a) (10 points) Choose **one method** from ordinary linear regression, logistic regression, and linear discriminant analysis, and **justify your choice** using the team's modeling goals in **one sentence**.

- (b) (10 points) Complete both tasks below.

- Plot:** Mark the three ROC points (FPR, TPR) and label each with its p^* value.
- Choose:** Report the candidate threshold p^* with the highest TPR subject to $\text{FPR} \leq 0.10$, and justify your choice in **one sentence**.



- (c) (10 points) Keep $p^* = 0.5$. In a new email stream, only 1% of emails are spam. For this calculation, treat the TPR and FPR computed from the evaluation table as exact rates that remain unchanged.

A team member claims,

“A flagged email is just as likely to be spam as in the original evaluation set.”

Assess the claim as **True** or **False**, supporting your answer with the **probability that a flagged email in the new stream is spam** and a **one-sentence explanation**. (*Hint: Apply Bayes’ theorem.*)

- (d*) (5 bonus points) *This bonus is independent of the numerical example above.*

At a fixed feature value x , suppose $P(Y = 1 \mid X = x) = p$, where $1/2 < p < 1$. Consider two rules:

$$\text{Rule A: } \hat{Y}_A = 1,$$

$$\text{Rule B: } \hat{Y}_B = \begin{cases} 1, & \text{with probability } p, \\ 0, & \text{with probability } 1 - p. \end{cases}$$

At this fixed x , Rule B’s random draw is **independent of the true label**. Compute **both conditional error probabilities** given $X = x$ and state **which rule has lower error**, using at most two sentences in addition to your calculations.