

STA 35C Statistical Data Science III

Practice 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

Provide the requested answer for each part. When an explanation is requested, limit it to one sentence.

- (a) (4 points) The values $Y = 1, 2, 3$ label three product types. Is predicting the product type from its measured features a **regression** or **classification** problem? State your choice (2 points) and give a one-sentence reason (2 points).

- (b) (4 points) Suppose $P(A) = 1/5$, $P(B | A) = 1/2$, and $P(B | A^c) = 1/8$. Find $P(B)$.

- (c) (4 points) If $X \sim \text{Bernoulli}(0.3)$, find $\mathbb{E}[X^2]$.

- (d) (4 points) Suppose X and Y are independent, with $\text{Var}(X) = 4$ and $\text{Var}(Y) = 9$. Find $\text{Var}(2X - Y)$.

- (e) (4 points) A model predicts a household's monthly electricity bill by

$$\hat{y} = 40 + 20x,$$

where the bill is in dollars and x is monthly usage in units of 100 kWh. If one household uses 200 kWh more than another, how much higher is its predicted monthly bill? Include units.

(f) (4 points) A 95% confidence interval for the *mean* monthly bill of households using 200 kWh is [\$75, \$85]. A colleague interprets this as an interval containing 95% of *individual* household bills at that usage. Is this interpretation correct? State **Yes** or **No** and explain in one sentence.

(g) (4 points) An OLS regression with an intercept has training $R^2 = 0.64$. Which statement correctly interprets this value? Circle one.

- **A.** The model correctly predicts 64% of future observations.
- **B.** The fit accounts for 64% of training-response variation about the mean.
- **C.** Every predictor has correlation 0.64 with the response.
- **D.** The sample correlation between the training responses and fitted values is 0.64.

(h) (4 points) A fitted logistic regression uses a real-valued predictor x and models

$$p(x) = \frac{1}{1 + \exp[-(\beta_0 + \beta_1 x)]}, \quad \beta_1 > 0.$$

Does a one-unit increase in x produce the same increase in $p(x)$ at every starting value of x ? State **Yes** or **No** and explain in one sentence.

(i) (4 points) A regression predicts height measured in centimeters and has residual standard error 5. In one sentence, state the units of the residual standard error and what the value 5 summarizes.

(j) (4 points) A screening classifier labels all 100 observations negative; 10 observations are actually positive. Report both quantities:

- the overall classification error rate;
- the fraction of actually positive observations that it misses.

Problem 2 (20 points). Dependence and prediction

A machine produces a daily record (X, Y) , where X is the number of alerts and $Y = 1$ if a special inspection occurs ($Y = 0$ otherwise). Their joint probability distribution is:

	$Y = 0$	$Y = 1$
$X = 0$	$1/4$	0
$X = 1$	0	$1/2$
$X = 2$	$1/4$	0

(a) (10 points) Compute $\mathbb{E}[X]$ and $\text{Cov}(X, Y)$.

(b) (10 points) Can a classifier using X predict Y with zero error under this joint distribution, meaning $P\{\hat{Y}(X) \neq Y\} = 0$?

- State **Yes** or **No**. If yes, give a prediction rule; a piecewise rule is acceptable. If no, explain why not in one sentence.
- In one sentence, reconcile your answer with the covariance in part (a).

Problem 3 (30 points + 5 bonus). Electricity bills and pricing plans

Let Y be a household's monthly electricity bill in dollars and X its monthly usage in units of 100 kWh. Let $D = 0$ or 1 indicate its pricing plan.

- (a) **(10 points)** In a simple regression analysis of 20 households, assume independent normal errors with mean zero and a common positive variance. The fitted coefficients are:

Coefficient	Estimate	Standard error
Intercept	40	12
Usage X	24	4

Test $H_0 : \beta_1 = 30$ against $H_1 : \beta_1 \neq 30$ at the 5% level. Use $t = (\hat{\beta}_1 - 30) / \text{SE}(\hat{\beta}_1)$. The 97.5th percentile of the t -distribution with 18 degrees of freedom is 2.10, so reject H_0 if $|t| > 2.10$. Report:

- the test statistic and your test decision;
- a one-sentence conclusion about the population slope in context.

- (b) **(10 points)** In a separate analysis, three models are fitted to the same training data and evaluated on the same separate test data:

Model	Predictors	Training R^2	Test MSE
A	X	0.55	36
B	X, D	0.61	31
C	X, D, XD	0.65	33

Which model would you favor for predicting bills of new households from the same population? Justify your recommendation in at most two sentences, using the table and distinguishing training fit from prediction on new observations.

(c) (10 points) Model C in part (b) has fitted equation

$$\hat{y} = 50 + 30x + 60D - 15xD.$$

An adviser recommends Plan 0 to every household because the coefficient of D is 60. Address both of the following, using at most two sentences in addition to your algebra:

- Is the adviser's recommendation supported by the fitted model?
- Find the usage at which the plans have equal predicted bills, and state which plan has the lower predicted bill on either side. Express usage in kWh.

(d*) (5 bonus points) *This bonus is independent of the required questions and the other bonus.*

The fitted function $f_1(x) = x$ has zero training RSS on the two observations $(0, 0)$ and $(1, 1)$. Construct a polynomial f_2 of degree at most two that also has zero RSS on these observations but satisfies $f_2(2) = 10$.

Problem 4 (30 points + 5 bonus). Quality control

A quality-control team uses a measurement score Z to classify a unit as defective ($Y = 1$) or good ($Y = 0$). The team models the score within each class as Gaussian, with different means and a common variance. It wants to combine these class-conditional models with the class proportions to classify units.

Each part supplies the numerical information it needs.

- (a) (10 points) Among **ordinary linear regression, logistic regression, and LDA**, which method most directly implements the team's stated modeling approach? In one sentence, name the method and give one reason linking its assumptions to this approach. No equations are needed.

- (b) (10 points) A fitted classifier uses the estimates

$$\hat{\mu}_0 = 0, \quad \hat{\mu}_1 = 2, \quad \hat{\sigma}^2 = 2, \quad \hat{\pi}_0 = \frac{3}{4}, \quad \hat{\pi}_1 = \frac{1}{4}$$

and the class-conditional Gaussian densities

$$f_k(z) = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\left[-\frac{(z - \hat{\mu}_k)^2}{2\hat{\sigma}^2}\right], \quad k = 0, 1.$$

Under this fitted model, predict class 1 when $P(Y = 1 \mid Z = z) \geq P(Y = 0 \mid Z = z)$. Find the cutoff and give the inequality in z that defines the region classified as defective, including ties. You may leave logarithms unevaluated, or use $\log 3 \approx 1.099$.

- (c) **(10 points)** A separate fitted classifier predicts class 1 when its estimated probability of a defect is at least the chosen threshold. Two thresholds give these results on the same test set of 1,000 units, keeping the fitted probabilities fixed. A false positive diverts a good unit for further inspection; a false negative lets a defective unit pass.

Threshold	False positives	False negatives	Overall error rate
0.5	30	10	4.0%
0.8	10	35	4.5%

The team favors threshold 0.5 because it has the smaller error rate. Let each false positive cost $r > 0$ units and each false negative cost 1 unit; correct classifications cost nothing.

On this test set, for which values of r should the team instead favor threshold 0.8 on the basis of total cost? Include the equal-cost case. Use at most two sentences in addition to your algebra.

- (d*) **(5 bonus points)** *This bonus is independent of the required questions and the other bonus.*

Can two one-dimensional classifiers based on Gaussian class distributions, each with equal class priors and common variance 1, give the same class label for every real x but different posterior probabilities $P(Y = 1 | X = x)$ at $x = 1$? Each predicts the class with the larger posterior probability, choosing class 1 at ties. If yes, give two choices of class means (μ_0, μ_1) and justify both requirements. If no, explain why. Use equations and at most two sentences.