STA 35C – Homework 4

Submission due: Tue, May 5 at 11:59 PM PT

Instructor: Dogyoon Song

Instructions: Upload a **single PDF file** to Gradescope via Canvas ("Homework 4" under "Assignments"). Name the file using the prefix of your UC Davis email ID and the homework number (e.g., `dgsong_hw4.pdf`). Include "STA 35C," your name, and the last four digits of your student ID on the front page. For coding problems, include your R code, relevant output, and any requested plots/tables in the PDF. No late submissions will be accepted; any submission received after the deadline will receive 0 points. For full information about submission requirements and the late submission policy, see the syllabus.

When preparing your submission:

- Complete any **two** of Problems 1–4 that you want to review the most (**50 points total**). Clearly indicate which two you want to get graded.
- Then complete **Problem 5** and **Problem 6** (**50 points total**).

The regular total score is **100 points**, with **up to 10 bonus points** available.

Problem 1: Probability (25 points in total).

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 4 randomly selected predictions and records

$$X = \text{number of incorrect predictions among the 4 audited predictions.}$$

Let A denote the event that the audited batch is from Model A.

(a) (8 points) First suppose the audited batch was produced by Model A. Each prediction is incorrect independently with probability 1/4, so

$$X \mid A \sim \text{Binomial}(4, \tfrac{1}{4}).$$

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$(4, \frac{1}{4})$ random variable is the sum of four independent Bernoulli$(\frac{1}{4})$ random variables.)

(b) (7 points) Continue under the Model A setting from part **(a)**, so that X has the same mean and variance as in part **(a)**; for notational brevity, we omit the $\mid A$ that indicates the conditioning on A here. Suppose the processing time Y (in minutes) for each batch is a continuous random variable with PDF

$$f_Y(y) = \begin{cases} \frac{1}{4}e^{-\frac{1}{4}y}, & y \geq 0, \\ 0, & y < 0. \end{cases}$$

We define the total cost

$$W = 10 + 3\,X + 2\,Y.$$

Compute $\mathbb{E}[W]$ and $\text{Var}(W)$, assuming $\text{corr}(X, Y) = 0.2$.
(*Hint:* You may use that Y has mean 4 and variance 16, and $\text{Cov}(X, Y) = \rho\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}$.)

(c) (10 points total) 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 1/8, so

$$X \mid B \sim \text{Binomial}(4, \tfrac{1}{8}).$$

(i) (5 points) What is the probability that a randomly chosen batch is from Model A *and* has all predictions correct?

(ii) (5 points) Given that you observe $X = 0$ incorrect predictions, find the posterior probability that the batch is from Model A. Briefly explain why this posterior probability is smaller than $P(A) = 1/2$. (*Hint:* Use Bayes' rule.)

Problem 2: Regression (25 points in total).

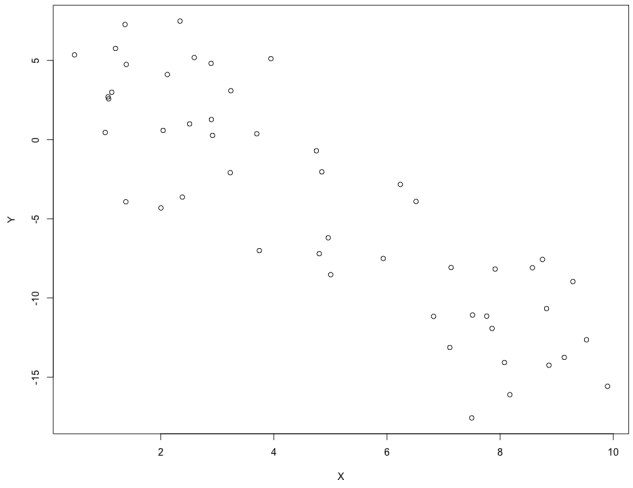
(a) (8 points) Answer the following two short items independently.

(i) Suppose a fitted simple linear regression model is

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

What is the predicted value of y at $x = 6$?

(ii) Now consider the fresh training dataset shown below. Roughly sketch a least-squares regression line on the scatterplot, and indicate how you would use that line to predict Y at $X = 6$.



(b) (8 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. If your test statistic falls between two listed values, an approximate p-value between the corresponding entries is acceptable:

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

(i) Compute the t-statistics and approximate two-sided p-values for the two regression coefficients above.

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

(c) (9 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 the original predictor X changes sign from $+1.5$ to -0.2.

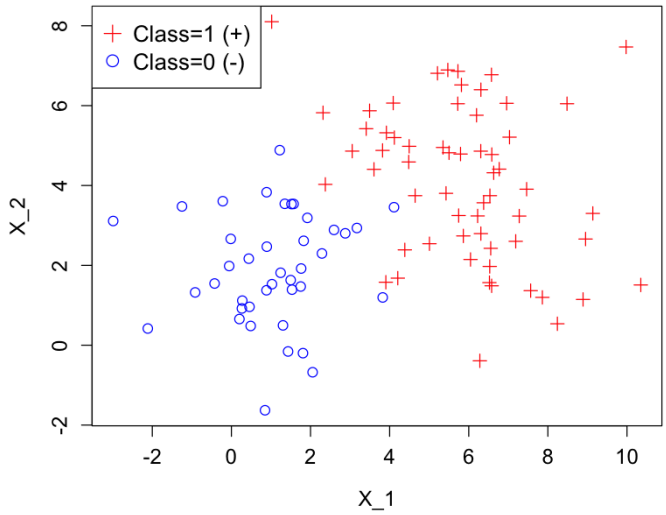
(i) How can adding X_2 cause the direction of X's effect to reverse?

(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 3: Logistic regression and classification errors (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 a reasonable decision boundary (for example, corresponding to $p^* = 0.5$).

(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) With respect to the boundary you sketched in part (a), define a *false positive* and a *false negative* in this context. Identify one example point of each from the scatterplot if possible; 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 the error rate, TPR and FPR from this confusion matrix.

(ii) If you increase the decision threshold from 0.5 to 0.9, 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 (equivalently, TPR ≥ 0.9), how would you use the ROC curve to choose p^*? Describe your choice verbally, or draw a hypothetical ROC curve and mark an operating point that satisfies this requirement with a brief justification.

Problem 4: Linear discriminant analysis (25 points in total).

Suppose that a bank wants to classify a potential customer using a one-dimensional LDA classifier based on a proprietary risk score Z. Suppose the bank collected the following data, written as (z, y):

$$(0,0), \quad (1,0), \quad (2,0), \quad (3,0), \quad (4,0), \quad (4,1), \quad (5,1), \quad (6,1).$$

Use these data throughout this problem.

(a) (5 points) Estimate the class priors, class means, and the pooled sample variance s^2.

(b) (5 points) Write the linear discriminant functions $\delta_0(z)$ and $\delta_1(z)$, using the estimates obtained above.

(c) (5 points) If $z_{\text{new}} = 3.5$, which class would you predict? Show your reasoning.

(d) (5 points) Suppose the bank wants to be more conservative and predict "no default" (class 0) only if

$$\Pr(Y = 0 \mid Z = z) \geq 0.8,$$

and otherwise predicts class 1. How does this modify the LDA decision rule in terms of $\delta_0(z)$ and $\delta_1(z)$? State the new decision boundary and apply it to $z_{\text{new}} = 3.5$.

(e) (5 points) If you additionally collect three more data points for default customers ($y = 1$) with roughly the same mean and variance as the current class-1 scores, so that the main change is in the estimated class prior, would your prediction at $z_{\text{new}} = 3.5$ change? Explain briefly.

Problem 5: Cross-validation (25 points in total + 5 bonus points).

(a) (25 points) Consider the `Auto` dataset from `ISLR2`. Consider polynomial regressions of `mpg` on `horsepower` of degrees $1, 2, 3, 4, 5$. You may want to use `R`.

(*Note:* This problem is adapted from a textbook problem [JWHT21, Chapter 5, Exercise 8], by replacing a synthetic dataset with a real dataset.)

(i) **(6 points)** Using the validation set approach with a random 50/50 split (`set.seed(2026)`), compute the validation MSE for each degree and state which degree you would choose. Use the same split for all five degrees.

(ii) **(6 points)** Repeat part (i) with a different random split (`set.seed(35)`). Again use the same split for all five degrees. Compare the selected degree and briefly comment on what changed.

(iii) **(6 points)** Use LOOCV to estimate the test MSE for each degree. Which degree would you choose based on the LOOCV estimates?

(iv) **(7 points)** Briefly compare the validation set approach and LOOCV in terms of stability and computational cost.

(b*) (*5 bonus points) In your own words:

(i) Explain how LOOCV is implemented.

(ii) Briefly discuss advantages/disadvantages of LOOCV vs. the single validation set approach and k-fold CV.

Problem 6: The Bootstrap (25 points in total + 5 bonus points).

(a) (6 points) Given a fixed sample $\{x_1, x_2, x_3, x_4, x_5\}$ with all values distinct, consider a bootstrap sample of size 5 drawn *with replacement* from these five observations. What is the probability of reproducing the original sample *exactly in the same order*? What is the probability of reproducing the original sample *ignoring the order*?

(b) (6 points) You flip a fair coin 10 times, observing 6 heads, 4 tails in a specific sequence:

$$\{H, T, H, H, T, H, T, T, H, H\}$$

For a bootstrap sample of size 10 drawn with replacement from these 10 observed tosses, find the probability of getting k heads. Compare this to flipping a fair coin 10 times independently. Evaluate these probabilities for $k \in \{4, 5, 6, 7\}$, and state which value of k is most likely under each scenario.

For the remaining parts, use R to create a dataset with:

```
set.seed(2026)
x <- rnorm(100)
y <- -1 + 2*x + rnorm(100)
```

(c) (6 points) For each $k = 1, \ldots, 200$, generate a fresh dataset using

```
set.seed(k)
x <- rnorm(100)
y <- -1 + 2*x + rnorm(100)
```

Then fit the simple linear regression model $Y \sim \beta_0 + \beta_1 X$ to each dataset and form the usual model-based 95% confidence interval for the slope. What fraction of these intervals contain the true value $\beta_1 = 2$? Is it close to 0.95?

(d) (7 points) Using the same 200 freshly generated datasets from part **(c)**, for each $k = 1, \ldots, 200$, fit the simple linear regression model and let $\hat{\beta}_1^{(k)}$ be the fitted slope. Then, using that same dataset, draw 500 bootstrap samples of size 100 with replacement, re-fit the same model on each bootstrap sample, and compute the bootstrap standard error $\widehat{\mathrm{SE}}_{\mathrm{boot}}^{(k)}$ of the slope. Form the normal-approximation bootstrap interval

$$\hat{\beta}_1^{(k)} \pm 1.96\, \widehat{\mathrm{SE}}_{\mathrm{boot}}^{(k)}.$$

What fraction of these 200 bootstrap intervals contain the true slope 2? Is it close to 0.95? Briefly compare your answer with part **(c)**.

(e*) (*5 bonus points) Repeat **(c)**–**(d)**, but now generate the data from the misspecified model below

```
set.seed(k)
x <- rnorm(100)
y <- -1 + 2*x - x^2 + rnorm(100)
```

for $k = 1, \ldots, 200$, and still fit the same simple linear regression model. Briefly discuss any difference in coverage you observe and why it may occur.

References

[JWHT21] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R*, volume 112 of *Springer Texts in Statistics*. Springer, New York, NY, 2nd edition, 2021.