STA 35C Statistical Data Science III

(Practice for the Final Exam)

Instructor: Dogyoon Song

Name: _____ **Student ID:** _____

Instructions: The problems here are designed to help students practice and prepare for the final exam; however, the actual exam may differ in content or format. You have 120 minutes to complete all problems, and the total score is 200 points.

- Make sure to clearly write your name and ID above.
- The actual final exam will be a **closed-book** exam. You may bring only a pen/pencil, a non-graphing calculator, and up to three letter-sized sheets of handwritten notes (both sides).
- 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 three decimal places.

Problem	**Score**
Problem 1	
Problem 2	
Problem 3	
Problem 4	
Problem 5	
Problem 6	
Problem 7	
Problem 8	
Total	

Problem 1 (16 points total). Multiple-choice Questions

In each subproblem, *check all* the boxes in front of the statements that you believe are **correct**. Each subproblem is worth 2 points total, and you only receive the 2 points if you check *all* correct options (and no incorrect options). At least one answer in each subproblem is correct, and there may be more.

(a) Probability, conditional probability, Bayes' rule

- ☐ If events A and B are independent, then $\Pr(A \mid B) = \Pr(A)$.
- ☐ Bayes' rule states $\Pr(A \mid B) = \dfrac{\Pr(B \mid A)\ \Pr(A)}{\Pr(B)}$.
- ☐ For disjoint events A and B, $\Pr(A \cap B) = \Pr(A)\ \Pr(B)$.
- ☐ $\Pr(A \cup B) = \Pr(A) + \Pr(B)$, regardless of whether A and B are disjoint.

(b) Linear regression: R^2

- ☐ R^2 measures the proportion of variability in Y explained by the model.
- ☐ R^2 can take on values between -1 and 1.
- ☐ If $R^2 = 1$, then the model's fitted values match the actual response perfectly.
- ☐ A large R^2 always guarantees excellent out-of-sample performance.

(c) Classification thresholds & errors

- ☐ Increasing the decision threshold p^* typically increases the false positive rate.
- ☐ A false negative occurs when the actual class is 0 but we predict 1.
- ☐ Setting $p^* = 0.5$ always optimizes both sensitivity and specificity.
- ☐ Lowering p^* generally increases the number of predicted positives.

(d) Best subset selection

- ☐ Best subset selection fits *all* possible subsets of predictors of each size to find the best subset.
- ☐ It always returns the model with the lowest test error among all subsets.
- ☐ It can suffer from high computational cost if the number of predictors is large.
- ☐ It may choose a different best model size than forward stepwise selection.

(e) Resampling methods: Bootstrap

- ☐ Typically the bootstrap procedure repeatedly samples *without replacement* from the original data.
- ☐ In each bootstrap sample of size n, all original observations must appear at least once.
- ☐ The bootstrap can be used to estimate the uncertainty (standard error) of an estimator.
- ☐ A single bootstrap sample is guaranteed to provide lower variance than the original sample.

(f) Regularization: Lasso

- ☐ Lasso uses an ℓ_1-penalty on the regression coefficients and can set some coefficients exactly to zero.
- ☐ Increasing the penalty parameter λ shrinks coefficients toward zero.
- ☐ Lasso always provides lower test error than ordinary least squares.
- ☐ If $\lambda = 0$, Lasso reduces to the usual least squares solution.

(g) Regression splines: Degree-3 spline

- ☐ A degree-3 spline is a piecewise cubic function.
- ☐ It is continuous up to its second derivative at each knot.
- ☐ It must have a continuous third derivative at each knot.
- ☐ "Natural splines" are degree-3 splines with additional boundary constraints.

(h) Clustering

- ☐ Clustering is a supervised learning task.
- ☐ In clustering, each observation has a numeric response Y used to form clusters.
- ☐ k-means and hierarchical clustering are two common approaches.
- ☐ Clustering can only be done with at most two features per observation.

Problem 2 (24 points total). True/False with Justification

For each statement below, circle **True** or **False**, and provide a brief justification in one sentence. **If true**, explain why, e.g., by stating a principle or example that supports the statement. **If false**, correct it or briefly explain why it is incorrect. **Each question is worth 3 points**; no partial credit without a justification.

(a) "If the first principal component accounts for most of the variation in the predictor variables X, then it must also be the single best predictor of the response Y."

True / False

Reason:

(b) "When using cross-validation, having more folds (e.g., $k = 10$ instead of $k = 5$) always guarantees a lower test error."

True / False

Reason:

(c) "In Lasso regression, all highly correlated predictors are shrunk equally toward zero."

True / False

Reason:

(d) "Lowering the decision threshold p^* in a logistic model will generally increase both the false positive rate and the true positive rate."

True / **False**

Reason:

(e) "A high training R^2 guarantees that the model will also generalize well to new data."

True / **False**

Reason:

(f) "The main purpose of the bootstrap is to reduce the bias of an estimator, rather than to assess variability or build confidence intervals."

True / **False**

Reason:

(g) "When performing clustering, labeled data are critical for computing within-cluster variance."

True / **False**

Reason:

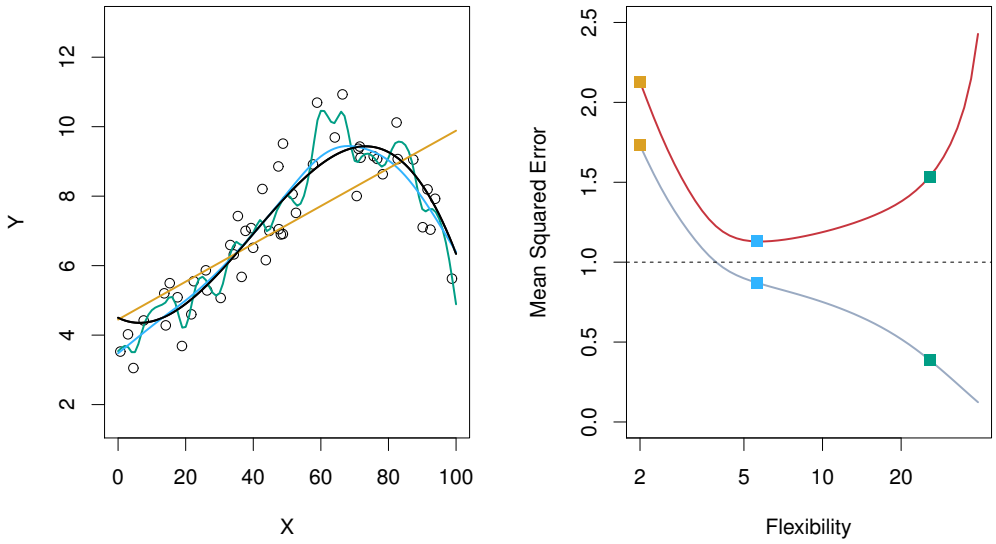
(h) "Benjamini-Hochberg procedure can control the family-wise error rate (FWER) when testing multiple hypotheses." **True** / **False**

Reason:

Problem 3 (20 points total). Statistical Learning

(a) (8 points) Describe briefly the objectives of *supervised learning* and *unsupervised learning*, emphasizing their main difference. Provide *one* example for each type.

(b) (6 points) Describe *training error* and *test error* and explain what they are used for. In the figure below, indicate which curve (red or black) shows test error. Then explain the bias-variance tradeoff in one or two sentences.



(c) (6 points) Explain how cross-validation uses the training dataset for validation, how it estimates performance on unseen test data, and which assumption is necessary for this estimation to be valid.

Problem 4 (40 points total). Regression

(a) (10 points total).

(i) **(5 points)** Given a dataset of (X_1, X_2, Y), describe how to obtain the least squares estimates for the linear model

$$Y \;=\; \beta_0 \;+\; \beta_1 X_1 \;+\; \beta_2 X_2.$$

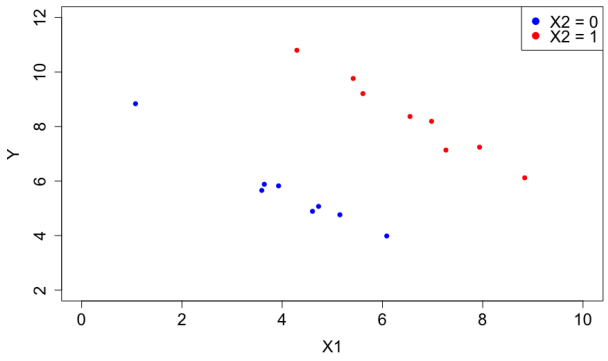
(*Hint:* You may either show the objective function to minimize or describe visually with a scatter plot how best-fit lines/planes are determined.)

(ii) **(5 points)** Suppose the fitted model yields $\hat{\beta}_0 = -3$, $\hat{\beta}_1 = 5$, and $\hat{\beta}_2 = 2$. What value would you predict for Y at a new point $x_{\text{new}} = (x_{\text{new},1}, x_{\text{new},2}) = (1, 2)$?

(b) (10 points total).

(i) **(5 points)** In the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$, interpret β_1. How does this interpretation differ from its counterpart in the simpler model $Y = \beta_0 + \beta_1 X_1$ (if at all)?

(ii) **(5 points)** Discuss how interpretations in multiple vs. simple regression might differ when explaining X_1's effect on Y, referring to the figure below. Explain how such confusion (confounding) can arise if X_2 is also related to Y. Suppose X_2 is a binary indicator ($X_2 = 1$ for treatment and $X_2 = 0$ for control), and interpret the observed trends in the figure.

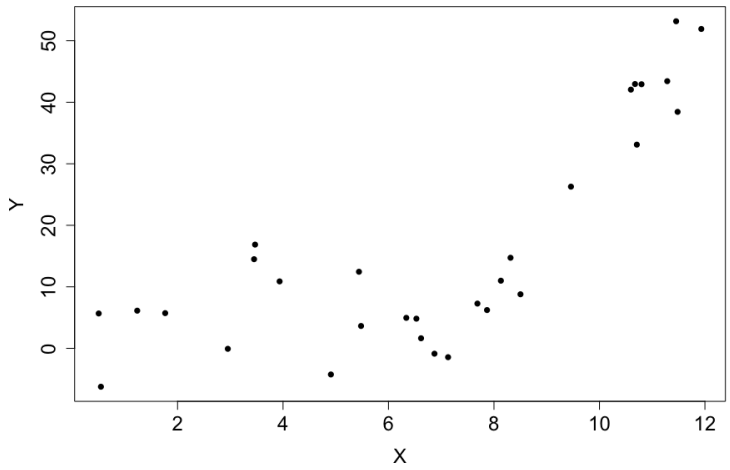


(c) (10 points total). For additional flexibility, suppose we model Y as a degree-3 polynomial in X_1, while still including X_2 linearly:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_1^3 + \beta_4 X_2.$$

(i) **(5 points)** Explain why including polynomial terms in X_1 might be beneficial compared to a purely linear form.

(ii) **(5 points)** Using the figure below, sketch and describe how linear regression vs. a degree-3 polynomial fit might differ for the same dataset. Also, discuss how these may vary in bias, variance, and flexibility.



(d) (10 points total). Finally, consider a piecewise (spline) approach with cubic segments in X_1, still including X_2 linearly:

$$Y = \begin{cases} \alpha_0 + \alpha_1 X_1 + \alpha_2 X_1^2 + \alpha_3 X_1^3 + \alpha_4 X_2, & X_1 \le c_1, \\ \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_1^3 + \beta_4 X_2, & c_1 < X_1 \le c_2, \\ \gamma_0 + \gamma_1 X_1 + \gamma_2 X_1^2 + \gamma_3 X_1^3 + \gamma_4 X_2, & X_1 > c_2. \end{cases}$$

(i) **(5 points)** Write down the constraints among α, β, and γ that ensure the model is a degree-3 spline.

(ii) **(5 points)** Explain why a *natural cubic spline* can be preferable to a single degree-3 polynomial, focusing on flexibility (degrees of freedom), and behavior at X_1 values near the boundaries.

Problem 5 (40 points total + 3 bonus points). Classification

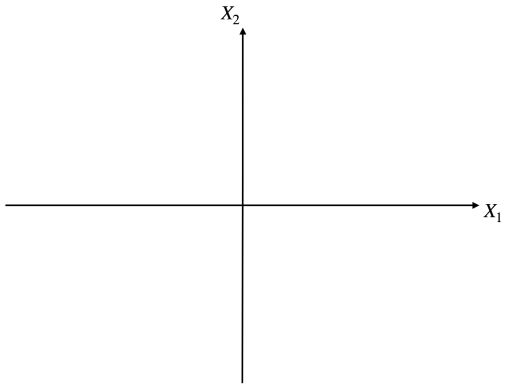
(a) (10 points total). Consider a two-dimensional logistic regression model:

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

Suppose the estimated coefficients are $\hat{\beta}_0 = -2,\ \hat{\beta}_1 = -1,\ \hat{\beta}_2 = 2$.

(i) **(5 points)** For $x_{\text{test}} = (1, 1)$, compute $\hat{p}(x_{\text{test}})$ and decide $\hat{y}_{\text{test}} = 1$ or 0 if $p^* = 0.5$.

(ii) **(5 points)** On the figure below, draw the decision boundary ($p^* = 0.5$), mark intercepts, and indicate the region where $\hat{Y} = 1$.



(b) (10 points total). The PDF of a normal with mean μ and variance σ^2 is

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

(i) **(5 points)** Suppose that we are given a dataset

$$\text{Class 0: } x = \{-2,\ 0,\ 1,\ 2\}, \quad \text{Class 1: } x = \{3,\ 4,\ 6\}.$$

What class would LDA predict for $X = 2$?

(ii) **(5 points)** Explain why LDA is *generative*: how does it assign probability (probability density function) to a new sample $(x_{\text{new}}, y_{\text{new}}) = (2, 1)$ given this dataset?

(c) (10 points total).

(i) **(5 points)** In some scenarios (e.g. medical tests, fraud detection), p^* might be set to a value other than 0.5. Give a brief example illustrating why you might prefer $p^* < 0.5$ over $p^* > 0.5$, or vice versa.

(ii) **(5 points)** Suppose you add three times more data points, *all from one class*. Would this change the logistic regression decision boundary or the LDA boundary, or both? Explain briefly.

(d) (10 points total + 3 bonus points).

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

	Pred = 1	Pred = 0
$Y = 1$	45	5
$Y = 0$	10	40

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

(ii) **(5 points)** Decreasing p^* from 0.5 to 0.1: do you expect more, fewer, or unchanged false positives and false negatives? Explain briefly.

(iii*) (3 bonus points*) Describe how to choose p^* to minimize FPR while keeping TPR at least 90%, referencing the ROC curve.

Problem 6 (20 points total). Inference & Hypothesis Testing

(a) (7 points) You fit a simple linear regression $Y \sim X$ and obtain:

Coefficient	**Estimate**	**Std. Error**	**t-statistic**
X_1	4	1.5	?

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) **(4 points)** Compute the t-statistics for the coefficient of X_1 and determine if it is statistically significant at the 5% level.

(ii) **(3 points)** What does this imply about the relationship between Y and X_1?

(b) (7 points) You want to test a complex, nonlinear function of parameters (T) and obtain $\hat{T} = 10$. Without a standard error formula, you bootstrap to get 6 estimates:

$$7, 10, 8, 11, 15, 12.$$

(i) **(4 points)** Construct a 95% confidence interval for T.

(ii) **(3 points)** Clearly state how to interpret the "95%" in this confidence interval, explaining which probability is intended to be about 95%.

(c) (6 points). For a single null hypothesis H_0, the left table (below) shows the probabilities of each outcome ($p_1 + p_2 + p_3 + p_4 = 1$). Now suppose we have m (e.g. 100) hypotheses tested simultaneously; let N_1, N_2, N_3, N_4 count each outcome, so $N_1 + N_2 + N_3 + N_4 = m$.

Single	H_0 is true	H_0 is not true
Reject H_0	p_1	p_2
Not reject H_0	p_3	p_4

Multiple	H_0 is true	H_0 is not true
Reject H_0	N_1	N_2
Not reject H_0	N_3	N_4

(i) **(3 points)** Define the *family-wise error rate (FWER)*, using these probabilities/counts.

(ii) **(3 points)** Briefly explain how the *Bonferroni correction* controls FWER in multiple tests.

Problem 7 (20 points total). Model Selection & PCA

(a) (8 points) Compare and contrast the goals and procedures of **best subset selection** vs. **principal component analysis (PCA)**. In your response:

- Briefly describe what each method aims to achieve.
- Outline the basic steps involved in each procedure.

(b) (6 points) Suppose you have a 2D dataset of five points:

$$\mathcal{X} = \big\{(1,\,2),\ (2,\,3),\ (3,\,5),\ (4,\,6),\ (5,\,5)\big\}.$$

(i) **(3 points)** Compute the *directional variance* of this dataset along the vector $\mathbf{v} = (1,\,1)$. (*Hint*: Project each point onto $\mathbf{v}$ and compute the variance of these projections.)

(ii) **(3 points)** Is this variance necessarily less than, equal to, or greater than the variance along the *first principal component*? Briefly justify your answer.

(c) (6 points). You perform PCA on a dataset with 7 variables (thus 7 principal components), whose variances are

$$24,\ 12,\ 7,\ 3,\ 2,\ 1,\ 1.$$

(i) **(3 points)** Compute the cumulative proportion of variance explained when retaining the top 3 principal components.

(ii) **(3 points)** Discuss the potential benefits and drawbacks of retaining too few or too many principal components. Explain how you might decide on the number of components to keep.

Problem 8 (20 points total). Clustering

(a) (7 points).

(i) **(4 points)** Explain the main goal of clustering. In other words, what are we trying to accomplish when we cluster a set of points?

(ii) **(3 points)** Briefly compare clustering with classification. In what ways are they similar, and how do they differ?

(b) (7 points). Suppose you apply 2-means clustering to the points $(1, 1)$, $(2, 0)$, $(8, 6)$, and $(9, 9)$.

(i) **(4 points)** Identify the two clusters you would ideally obtain and compute their centroids. Show how each point is assigned.

(ii) **(3 points)** Would the k-means algorithm always produce this exact result? Briefly explain why or why not.

(c) (6 points).

(i) **(3 points)** Using complete linkage, draw a dendrogram for these four points. Then explain how hierarchical clustering would form clusters from the dendrogram.

(ii) **(3 points)** State one advantage and one drawback of k-means compared to hierarchical clustering.