STA 35C Statistical Data Science III

Final Exam

Instructor: Dogyoon Song

Name: ____________________ **Student ID:** ____________________

Instructions: This is a **closed-book final exam**. You may bring a pen or pencil, three letter-sized sheets of *hand-written* notes (both sides), and a *non-graphing* calculator. No other materials (e.g., textbooks) are allowed. **You have 120 minutes** to complete all problems, and **the total score is 200 points**, with *up to 6 bonus points available.* Once you receive this exam problem set, please **confirm you have all 14 pages**.

- Make sure to clearly write your name and ID above.
- Present answers succinctly, but include all relevant steps for full credit. Partial credit is possible only if your reasoning is clearly presented and can be easily followed 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, and you only earn the 2 points if you check *all* correct options (and no incorrect ones). **In each subproblem, there may be one or two correct answers.**

(a) Linear regression: R^2

□ R^2 measures the proportion of variability in Y explained by the model.

□ R^2 can take on any values between -1 and 1.

□ If $R^2 = 1$, then the model's predicted values match the actual response perfectly.

□ A large R^2 always guarantees excellent out-of-sample performance.

(b) Classification thresholds & errors

□ In a discriminative approach (e.g. logistic regression), we explicitly model $p(X \mid Y)$ and $p(Y)$.

□ A false negative occurs when the actual class is negative (Y=0) but we predict positive (Y=1).

□ A generative method (e.g. LDA) often models $p(X \mid Y)$ and $p(X)$.

□ Lowering p^* generally increases the number of predicted positives.

(c) 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 cannot choose a different best model size than backward stepwise selection.

(d) Cross-validation

□ k-fold cross-validation splits the data into k folds, using 1 fold for training and $k-1$ folds for testing.

□ Leave-one-out cross-validation is a special case of k-fold where each fold contains exactly 1 observation.

□ Cross-validation can be used to select tuning parameters or compare models' predictive performance.

□ k-fold cross-validation always guarantees the absolute minimum test error.

(e) Regularization

□ Ridge regression penalizes coefficients by their absolute values (ℓ_1-penalty).

□ Increasing the regularization parameter λ generally shrinks coefficients toward zero.

□ Lasso always outperforms ordinary least squares in terms of test error.

□ Ridge can set some coefficients exactly to zero, similar to Lasso.

(f) Regression splines: Degree-3 spline

- □ A degree-3 spline is a piecewise cubic polynomial 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 splines with additional degrees of freedom for a more natural fit.

(g) PCA (Principal Component Analysis)

- □ The first and the second principal components are orthogonal to each other.
- □ The first principal component always aligns with the direction of smallest variance in X.
- □ PCA can reduce the dimensionality of X by projecting onto a few components.
- □ PCA requires a labeled response Y to identify the principal axes.

(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 it; **if false**, correct or briefly explain the error. **Each question is worth 4 points**; no partial credit without a justification.

(a) "If the first principal component explains 80% of the variance in X, then it must be the single best predictor of the response Y among all principal components"

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) "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:

(d) "Adding more predictors to a linear model always reduces the test error."

True / **False**

Reason:

(e) "In k-means clustering, the final partition is guaranteed to be the global minimum of the objective function."

True / **False**

Reason:

(f) "When we want to keep the probability of making at least one false rejection below 5%, we typically control the False Discovery Rate (FDR) at 5%."

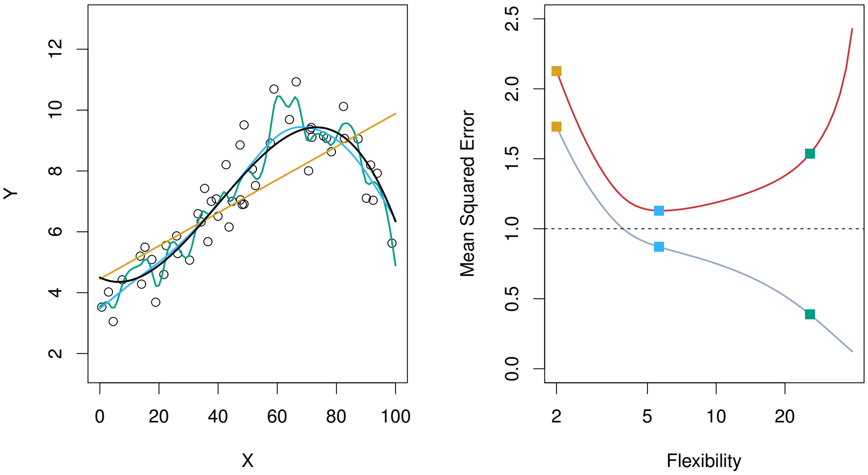
True / **False**

Reason:

Problem 3 (20 points total). Statistical Learning

(a) **(6 points)** Define *training error* and *test error*, and explain their roles in fitting and evaluating a model.

(b) **(7 points)** In the right figure below, state which curve (upper or lower) represents *test error*. Then, explain the bias-variance tradeoff in one or two sentences, describing how (squared) bias, variance, and irreducible error might change as the model flexibility increases. Lastly, choose the model you would use (among three options) and briefly justify.



(c) **(7 points)** State the purpose of cross-validation in one sentence, and describe how *k-fold cross-validation* computes validation error using a training dataset of size n.

Problem 4 (40 points total + 3 bonus points). Regression

(a) (10 points total). Suppose you have a dataset $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, and want to regress Y on X.

(i) **(5 points)** Describe how to obtain least squares estimates for the regression coefficients in the model:

$$\text{(Model 1)} \qquad Y \approx \beta_0 + \beta_1 X.$$

Either write down the objective function to be minimized, or explain visually (e.g. with a scatter plot) how the best-fit line and parameters are determined.

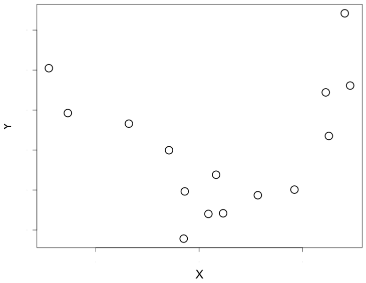
(ii) **(5 points)** Suppose the fitted model yields $\hat{\beta}_0 = -3$, $\hat{\beta}_1 = 5$. Predict Y at $x_{\text{new}} = 2$, and explain why the actual y-value (y_{new}) may differ from the predicted value.

(b) (15 points total + 3 bonus points). Now consider two more flexible models:

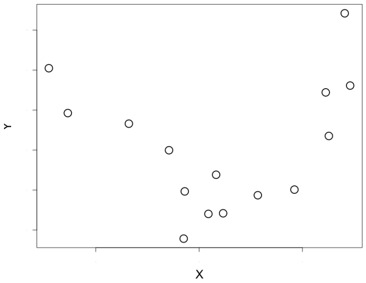
$$\begin{aligned} &\text{(Model 2)} \qquad Y \approx \beta_0 + \beta_1 X + \beta_2 X^2, \\ &\text{(Model 3)} \qquad Y \approx \begin{cases} \beta_{1,0} + \beta_{1,1} X, & X \le c, \\ \beta_{2,0} + \beta_{2,1} X, & X > c. \end{cases} \qquad (c : \text{hyperparameter}) \end{aligned}$$

(i) **(5 points)** Explain under what circumstances these might be preferred over a simple linear model (Model 1). Discuss their relative flexibility and possible impacts on bias, variance, or irreducible error.

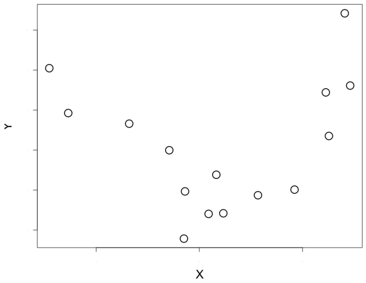
(ii) **(5 points)** Suppose you fit Models 1, 2, and 3 by least squares to the same dataset depicted below. Sketch each regression function below. For Model 3, choose a breakpoint c appropriately.



Model 1



Model 2



Model 3

(iii) **(5 points)** Define a *cubic spline* and a *natural cubic spline*, specifying the constraints they must satisfy and how they differ.

(iv*) **(3 bonus points*)** What is a *smoothing spline*? Briefly explain what the fitted curve look like in the two limits: (a) as $\lambda \to 0$ and (b) as $\lambda \to \infty$.

(c) (15 points total). Finally, you want to predict Y using three predictors: D (binary, e.g. treatment indicator), X_1, and X_2, all numeric except D being categorical (0/1). You fit three models:

$$\begin{aligned}&\text{Model A: } Y \approx \alpha_0 + \alpha_1\, D,\\ &\text{Model B: } Y \approx \beta_0 + \beta_1\, X_1,\\ &\text{Model C: } Y \approx \gamma_0 + \gamma_1\, D + \gamma_2\, X_1 + \gamma_3\, X_2,\end{aligned}$$

and obtain the following R^2 values and coefficient estimates:

Model A: $R^2 = 0.4$ $\quad \alpha_1 = 2,$ with SE $= 0.75;$

Model B: $R^2 = 0.6$ $\quad \beta_1 = 4,$ with SE $= 1.25;$

Model C: $R^2 = 0.9$ $\quad \gamma_1 = -4,\ \gamma_2 = 1,\ \gamma_3 = 3,$ with SEs $(1.33,\ 2.5,\ 0.5)$, $\quad \mathrm{corr}(X_1, X_2) = 0.8.$

(i) **(5 points)** Which model likely has the best *predictive* fit? Define R^2 and explain its meaning briefly.

(ii) **(5 points)** How do we interpret β_1 in Model B? How does this differ from interpreting γ_2 in Model C?

(iii) **(5 points)** Based on the information above, discuss what these data suggest about Y's relationship to D, X_1, X_2. (*Hint:* Discuss the significance of correlation, sign of the effect, etc.)

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

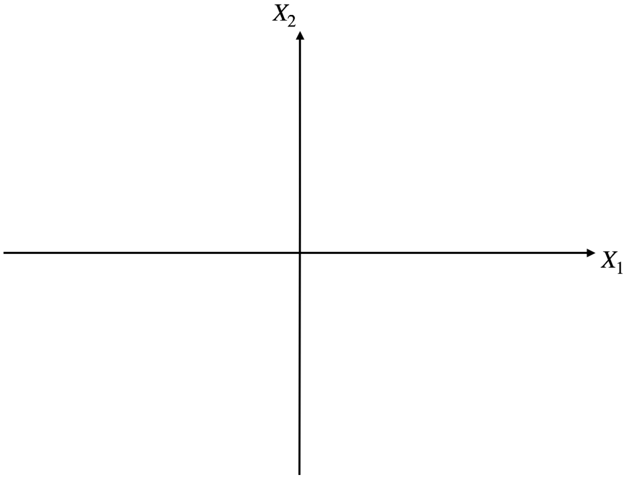
(a) (10 points total). Consider a two-dimensional logistic regression model:

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

Suppose the estimated coefficients are $\hat\beta_0 = -4,\ \hat\beta_1 = 3,\ \hat\beta_2 = 2.$

(i) **(5 points)** For $x_{\text{test}} = (1,1)$, compute $\hat p(x_{\text{test}})$ and decide whether $\hat y_{\text{test}} = 1$ or 0 using the decision rule "$\hat y = 1$ if and only if $\hat p(x) \geq p^* = 0.5$".

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



(b) (15 points total). The PDF of a Gaussian random variable with mean μ and variance σ^2 is

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

(i) **(5 points)** Explain briefly how *generative modeling* approaches for classification use Bayes' rule to formulate the conditional probability $\Pr[Y|X]$, and discuss how it differ from *discriminative* approaches.

(ii) **(5 points)** Suppose that we have a dataset

$$\text{Class 0: } x = \{-3,\ -2,\ 0,\ 1\}, \qquad \text{Class 1: } x = \{4,\ 5,\ 6\}.$$

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

(iii) **(5 points)** If you additionally collect two points, $x = \{3.5, 6.5\}$ from Class 1, does that change your class prediction with LDA at $x = 2$? Justify briefly.

(c) (15 points total + 3 bonus points).

(i) **(5 points)** From 100 data points, you obtained the following confusion matrix (at $p^* = 0.5$):

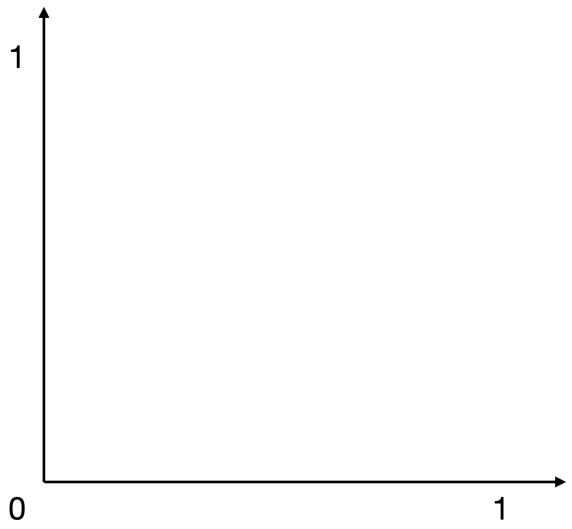
	Pred = 1	Pred = 0
$Y = 1$	52	8
$Y = 0$	6	34

Compute the true positive rate (TPR) and false positive rate (FPR).

(ii) **(5 points)** If p^* increases from 0.5 to 0.9, do you expect more, fewer, or the same number of false positives/negatives? Provide a reason briefly.

(iii) **(5 points)** In some scenarios (e.g. medical tests, fraud detection), p^* might be set to a value other than 0.5. Give an example scenario in one sentence where $p^* > 0.5$ might be preferable, then explain how that choice affects decisions and why it could lead to a more desirable outcome.

(iv*) **(3 bonus points*)** How do we pick p^* to maximize TPR while keeping FPR $\leq 10\%$? Draw an example ROC curve below, and illustrate your decision on it, marking your chosen operating point with each axis and coordinate clearly labeled.



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

(a) **(10 points)** You have an estimate $\hat{\theta} = 10$ of a parameter θ, which lacks a standard error formula available. You bootstrap and acquire 5 estimates:

$$6,\ 8,\ 8,\ 11,\ 12.$$

Using a *normal approximation*, construct a 95% confidence interval for θ, with $\hat{\theta}$ and the standard deviation estimated from the 5 bootstrap estimates. (*Hint*: $z_{0.975} \approx 1.96$.) Then state how to interpret the "95%" in this confidence interval—i.e., which probability is intended to be about 95%.

(b) **(10 points total).** For a single null hypothesis H_0, the table on the left 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) **(5 points)** If we reject H_0 at level α (e.g. 0.05), express this requirement as an inequality involving p_1, p_2, p_3, p_4, and α.

(ii) **(5 points)** Define the *family-wise error rate (FWER)* and the *false discovery rate (FDR)* using these probabilities/counts. Explain their meanings in one sentence each.

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

(a) (7 points) Compare and contrast *principal component analysis (PCA)* against *best subset selection.* In your answer, (i) state the main goal of PCA vs. best subset selection, and (ii) outline how each procedure operates.

(b) (6 points) Suppose you have a two-dimensional dataset of five points:

$$\mathcal{X} = \big\{(-4,\, 2),\ (-1,\, 3),\ (0,\, 4),\ (2,\, 6),\ (3,\, 5)\big\}.$$

Compute the *directional variance* of this dataset along the vector $\mathbf{e}_1 = (1,\, 0)$ and $\mathbf{e}_2 = (0,\, 1)$. Then, decide whether $\mathbf{u}_1$ (the first principal component) is "closer" to $\mathbf{e}_1$ or $\mathbf{e}_2$, and briefly justify.

(c) (7 points). Performing PCA on a dataset yields 7 principal components with variances:

$$28,\ 16,\ 9,\ 3,\ 2,\ 1,\ 1.$$

Sketch a scree plot and compute the cumulative proportion of variance explained (PVE) by the top 3 components. Then briefly discuss the trade-offs of keeping too few or too many components.

Problem 8 (20 points total). Clustering

(a) (14 points total). You have a two-dimensional dataset:

$$z_1 = (0,\ 2), \quad z_2 = (2,\ 1), \quad z_3 = (3,\ -2), \quad z_4 = (6,\ 0).$$

(i) **(7 points)** Briefly explain the k-means algorithm and illustrate *two iterations* of k-means with $k = 2$ on $\{z_1, z_2, z_3, z_4\}$, assuming the initial cluster assignment:

$$C_1 = \{z_1,\ z_3\}, \quad C_2 = \{z_2,\ z_4\}.$$

(ii) **(7 points)** Using *complete linkage*, construct a dendrogram for $\{z_1, z_2, z_3, z_4\}$. Then explain how you would form two clusters from this dendrogram and identify those clusters.

(b) (6 points). State one advantage and one drawback of k-means compared to hierarchical clustering.