STA 35C Statistical Data Science III

(Mock Exam for Midterm 2)

Instructor: Dogyoon Song

Name: ____________________ **Student ID:** ____________________

Instructions: This mock exam is designed to illustrate the approximate structure, length, and style of Midterm 2. However, the actual Midterm 2 may differ in content or format from this practice exam. You have 50 minutes to complete all problems. The total score is 120 points.

- Make sure to clearly write your name and ID above.
- The actual Midterm 2 will be a **closed-book** exam. You may bring only a pen/pencil, one letter-sized sheet of handwritten notes (both sides), and a non-graphing calculator.
- 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 two decimal places.

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

Problem 1 (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**, include one example or principle that supports the statement. **If false**, briefly correct it or give the main reason it is incorrect. **Each question is worth 4 points**; there is no partial credit without a justification.

1. Single-validation set approach typically exhibits a lower-variance estimate of test error compared to 5-fold CV.

   True / **False**

   Reason:

2. In each bootstrap sample drawn *with replacement* from the original dataset, some observations may appear multiple times while others do not appear at all.

   True / **False**

   Reason:

3. Forward stepwise selection starts with all predictors and removes them one by one based on criteria such as RSS or adjusted R^2.

   True / **False**

   Reason:

4. In Lasso (L1-penalized) regression, sufficiently large λ can force some coefficients to be exactly zero.

   True / **False**

   Reason:

5. Performing many hypothesis tests at a fixed $\alpha = 0.05$ inflates the *power* rather than the chance of any false positives.

   True / **False**

   Reason:

6. One symptom of overfitting is a much lower training error than test (or cross-validation) error.

   True / **False**

   Reason:

Problem 2 (18 points total): Cross-Validation

Suppose that we have a dataset

$$(x_1, y_1) = (2, 3), \qquad (x_2, y_2) = (4, 5), \qquad (x_3, y_3) = (7, 10), \qquad (x_4, y_4) = (9, 14).$$

(a) (12 points) We want to choose between:

a linear model: $f(x) = \beta_0 + \beta_1 x + \epsilon$ or a quadratic model: $g(x) = \beta_0 + \beta_1 x^2 + \epsilon$.

Suppose we fit these models using 2-fold cross-validation (CV), splitting the data into two folds: $\{1, 3\}$ and $\{2, 4\}$. Calculate the 2-fold CV estimates for test MSE for both models. Then state which model ($f(x)$ or $g(x)$) you would pick and why.

(b) (6 points) Briefly explain the advantage(s) and disadvantage(s) of k-fold CV over the LOOCV.

Problem 3 (20 points total): Bootstrap

Suppose that you have a dataset consisting of 5 numbers, and generated 3 bootstrap samples as follows.

Sample	Bootstrap 1	Bootstrap 2	Bootstrap 3
2	2	3	2
3	2	5	3
5	5	7	5
7	7	8	5
8	8	8	8

(a) **(8 points)** Compute the sample mean $\hat{\mu}$ for the *original sample* and for each of the three bootstrap samples.

(b) **(6 points)** Compute the sample standard deviation of these four $\hat{\mu}$ values.

(c) **(6 points)** Construct a 95% confidence interval for the population mean μ based on your bootstrap estimates. (You may use a percentile-based or normal-approximation approach, but state your method clearly.)

Problem 4 (20 points total): Subset Selection

You have 3 predictors (X_1, X_2, X_3) and a response Y. Below is a table of the *Residual Sum of Squares* (RSS) for **all 8 possible subsets** (including the null model):

Predictors in Model	RSS	Predictors in Model	RSS
$\emptyset$	40.0	X_1, X_2	8.0
X_1	10.0	X_1, X_3	12.0
X_2	15.0	X_2, X_3	14.5
X_3	20.0	X_1, X_2, X_3	7.5

(a) (8 points) For *Best Subset Selection*, which model is chosen for each of $k = 0, 1, 2, 3$ predictors?

(b) (6 points) For *Forward Stepwise*, show the path of how you add predictors starting from $\varnothing$. For *Backward Stepwise*, show how you remove predictors starting from (X_1, X_2, X_3).

(c) (6 points) Discuss one main advantage and one main drawback of using *Forward Stepwise* instead of *Best Subset* selection.

Problem 5 (20 points total): Regularization

(a) **(10 points)** Suppose You have 4 observations, 2 predictors (X_1, X_2), plus response Y. Below are the fitted coefficients at $\lambda = 1$:

Method	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
A	2.2	1.8	0.0
B	2.0	1.5	1.2

(i) Which method (A or B) is likely Lasso, and which is likely Ridge? Give a short reason.

(ii) Interpret why method A sets $\hat{\beta}_2 = 0$ while method B shrinks it to 1.2. What might this imply about X_2's correlation with X_1 or its importance?

(b) **(10 points)** Suppose you fit Ridge and Lasso on a bigger dataset (10 predictors) at three λ values each. The 5-fold CV errors are:

	Ridge			**Lasso**		
Regularization	$\lambda = 0.1$	$\lambda = 1.0$	$\lambda = 5.0$	$\lambda = 0.1$	$\lambda = 1.0$	$\lambda = 5.0$
CV Error	0.90	0.88	0.93	0.85	0.86	0.95

(i) Which λ would you pick for Ridge? For Lasso? Give the corresponding CV error for each choice.

(ii) Suppose Lasso at $\lambda = 1.0$ zeroes out 2 of the 10 predictors, but at $\lambda = 0.1$ it keeps all. If the CV errors are 0.85 vs. 0.86, how might you weigh a simpler model vs. a tiny difference in test error?

Problem 6 (18 points total + 2 bonus points): Multiple Testing

(a) (10 points) You have 10 hypotheses to test (each at $\alpha = 0.05$) with p-values

$$H_{0,1} : 0.001, \quad H_{0,2} : 0.01, \quad H_{0,3} : 0.02, \quad H_{0,4} : 0.03, \quad H_{0,5} : 0.04,$$
$$H_{0,6} : 0.10, \quad H_{0,7} : 0.15, \quad H_{0,8} : 0.20, \quad H_{0,9} : 0.25, \quad H_{0,10} : 0.50.$$

How many would look significant if you test these at $\alpha = 0.05$ with *no correction*? How many remain significant under Bonferroni correction? Comment in one sentence on any difference in the number of "discoveries."

(b) (8 points) Now you have 5 hypotheses to test, and obtain the following p-values:

$$H_{0,1} : 0.002, \quad H_{0,2} : 0.01, \quad H_{0,3} : 0.04, \quad H_{0,4} : 0.09, \quad H_{0,5} : 0.20$$

AApply the Benjamini-Hochberg procedure to control the *False Discovery Rate* at 5%. List which null hypotheses you reject (i.e., declare significant).

(c*) (*2 bonus points) Describe how controlling **FDR** differs from controlling the Familywise Error Rate (FWER). In which scenario might FDR be preferable, and why is it typically more powerful (i.e. yields more rejections) than Bonferroni?