STA 35C Statistical Data Science III

Midterm exam 2

Instructor: Dogyoon Song

Name: ____________________ **Student ID:** ____________________

Instructions: This midterm exam is a **closed-book** exam. 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 5 bonus points available***. Once you receive this exam problem set, please **confirm you have all 7 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 shown and traceable.
- If necessary, round all numerical answers to three decimal places.

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

Problem 1 (20 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 4 points**; no partial credit without a justification.

(a) Using more folds in k-fold cross-validation (e.g., 10 vs. 5) generally increases the computational cost.

True / **False**

Reason:

(b) In each bootstrap sample drawn *with replacement*, every original data point must appear at least once.

True / **False**

Reason:

(c) Forward stepwise selection can remove a predictor added in an earlier step if it later becomes non-significant.

True / **False**

Reason:

(d) As λ increases in Lasso regression, correlated predictors are often shrunk *together*, whereas in Ridge one might be set to zero while another is kept.

True / **False**

Reason:

(e) When controlling the False Discovery Rate (FDR) at $q = 0.05$, we guarantee that with probability 95%, there are no false positives among the rejected null hypotheses.

True / **False**

Reason:

Problem 2 (20 points total): Cross-Validation

(a) **(6 points)** Briefly explain one advantage and one disadvantage of *5-fold cross-validation* compared to using a single train/validation split.

(b) **(14 points total)** Suppose that we have a dataset

$$(x_1, y_1) = (-2, 1), \qquad (x_2, y_2) = (0, 3), \qquad (x_3, y_3) = (3, 9).$$

We want to compare two regression models:

Linear model: $f(x) = \beta_0 + \beta_1 x + \epsilon$ or **Quadratic model:** $g(x) = \beta_0 + \beta_1 x^2 + \epsilon$.

(i) **(10 points)** Use *leave-one-out cross-validation (LOOCV)* to estimate the test MSE for each model.

(ii) **(4 points)** Decide which model (f or g) you would select, and briefly justify your choice.

Problem 3 (20 points total): Bootstrap

You have a coin with unknown head probability $p \in [0,1]$. After 5 flips, you observed the sequence:

$$H, \quad T, \quad T, \quad H, \quad T \qquad\qquad \text{(i.e., 2 heads out of 5).}$$

(a) **(6 points)** If you resample from these 5 flips *with replacement* to generate a new bootstrap sample of size 5, what is the probability of drawing the *exact same* sequence (H, T, T, H, T) in that sample?

(b) **(8 points)** Suppose we generated 4 bootstrap samples (each of size 5) as shown below:

	Bootstrap 1	Bootstrap 2	Bootstrap 3	Bootstrap 4
Flip 1	T	H	T	H
Flip 2	H	H	H	T
Flip 3	T	T	T	H
Flip 4	H	T	H	T
Flip 5	H	H	T	T

Construct a 95% confidence interval for the Head probability p, using:

- $\hat{p}$ estimated from the original sample (H, T, T, H, T), and
- the normal approximation, with the standard deviation estimated from the four bootstrapped samples. (*Hint*: $z_{0.9} \approx 1.28$, $z_{0.95} \approx 1.64$, $z_{0.975} \approx 1.96$, $z_{0.99} = 2.33$.)

(c) **(6 points)** In this context, how do we interpret "95%" in the 95% confidence interval for p? Specifically, describe which probability is intended to be approximately 95% succinctly.

Problem 4 (20 points total): Subset Selection

You have 4 predictors (X_1, X_2, X_3, X_4) and a response Y. Below is a table of the *Residual Sum of Squares* (RSS) for **all 16 possible subsets** (including the null model), computed from a sample of size $n = 11$:

Predictors	RSS	Predictors	RSS	Predictors	RSS	Predictors	RSS	Predictors	RSS
$\emptyset$	100.0	X_1	50.0	X_1, X_2	30.0	X_1, X_2, X_3	28.0	X_1, X_2, X_3, X_4	19.0
		X_2	40.0	X_1, X_3	45.0	X_1, X_2, X_4	21.0		
		X_3	60.0	X_1, X_4	25.0	X_1, X_3, X_4	20.0		
		X_4	45.0	X_2, X_3	35.0	X_2, X_3, X_4	25.0		
				X_2, X_4	32.0				
				X_3, X_4	40.0				

Hint: Recall $R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$ and $R^2_{\text{adj}} = 1 - \frac{\text{RSS}}{\text{TSS}} \cdot \frac{n-1}{n-p-1}$ for a model with p predictors. Note TSS = 100 here.

(a) **(7 points)** Using **Best Subset Selection**, which subset is chosen for each $k = 0, 1, 2, 3, 4$? Ultimately, which model might you pick to use and why?

(b) **(7 points)** Using **Forward Stepwise**, list which subset is chosen at each size $k = 0, 1, 2, 3, 4$. Finally, which model might you select to use and why?

(c) **(6 points)** Briefly state *one advantage* and *one disadvantage* of using **Backward Stepwise** instead of Best subset selection.

Problem 5 (20 points total): Regularization

(a) (10 points) Recall the ridge regression estimates for a linear model is obtained by minimizing

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2.$$

As we increase λ from 0 to ∞, how do you expect each of the following to behave?

Pick among the five options: (1) "Remain constant," (2) "Steadily increase," (3) "Steadily decrease," (4) "Decrease initially, and then eventually start increasing in a U shape," or (5) "Increase initially, and then eventually start decreasing in an inverted U shape."

Each question is worth 2 points; you don't need to justify your choice.

(i) Training RSS (=training MSE)

(ii) Test RSS (=test MSE)

(iii) (Squared) bias

(iv) Variance

(v) Irreducible error

(b) (10 points) Suppose you fit two methods (Method A and Method B) — one is **Ridge**, the other is **Lasso** — at three values of λ each, obtaining the following 5-fold CV errors and coefficient estimates $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ for a 2-predictor model:

	Method A				**Method B**			
λ	CV error	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	CV error	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
0.1	1.10	0.2	0.80	0.40	1.10	0.3	0.75	0.10
1.0	1.05	0.3	0.65	0.25	1.15	0.5	0.70	0.00
5.0	1.30	0.6	0.40	0.10	1.35	0.8	0.40	0.00

(i) Which method (A or B) is likely Lasso, and which is likely Ridge? Briefly justify your choice.

(ii) Based on the above table, which λ among $\{0.1, 1.0, 5.0\}$ might you pick for each method? If you care about achieving simpler models, how could that possibly change your decision?

Problem 6 (20 points total + 5 bonus points): Multiple Testing

(a) **(10 points total)** Consider 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)** Often we reject H_0 at significance level α (e.g. 0.05). Write this requirement as an inequality involving p_1, p_2, p_3, p_4 and α.

(ii) **(5 points)** Suppose instead we aim to control the *false discovery rate (FDR)* at level q (e.g. 0.10). Express this goal as an inequality involving N_1, N_2, N_3, N_4 and q.

(b) **(10 points total + 5 bonus points)** You have 8 hypotheses to test (each at $\alpha = 0.05$) with p-values:

$$H_{0,1} : 0.001, \quad H_{0,2} : 0.01, \quad H_{0,3} : 0.02, \quad H_{0,4} : 0.04,$$
$$H_{0,5} : 0.06, \quad H_{0,6} : 0.10, \quad H_{0,7} : 0.15, \quad H_{0,8} : 0.25.$$

(i) **(5 points)** With *no correction*, which hypotheses are rejected at $\alpha = 0.05$?

(ii) **(5 points)** With the *Bonferroni correction* to achieve FWER $\leq \alpha$, which hypotheses are rejected?

(iii*) **(5 bonus points*)** Apply the Benjamini-Hochberg procedure to control FDR at 10%. Which hypotheses are rejected?