STA 35C Statistical Data Science III

Midterm exam 2

Instructor: Dogyoon Song

Name: _____ **Student ID:** _____

Instructions: This midterm exam is **closed-book**, except for the permitted note sheet described below. 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 10 bonus points available*. Once you receive this exam, please confirm that you have all 8 pages.

- Make sure to clearly write your name and student 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 by the grader.
- If necessary, round all numerical answers to three decimal places; you may leave fractions such as 8/55 or logarithms such as log(3/4) unevaluated.
- Bonus problems can be more challenging; consider attempting them after you finish the main problems.

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

Problem 1 (20 points). 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, such as 10-fold instead of 5-fold, 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) Ridge can set some coefficients exactly equal to zero, while lasso regression generally shrinks coefficients toward zero without setting them exactly to zero.

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 (24 points). Cross-validation

Suppose that we have a dataset of (x, y) pairs and we want to choose between two regression models:

$$\textbf{Linear model: } f(x) = \beta_0 + \beta_1 x \qquad \text{and} \qquad \textbf{Quadratic model: } g(x) = \beta_0 + \beta_1 x + \beta_2 x^2.$$

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

(b) (12 points) Suppose our dataset $\big\{(0,0),(0,0),(1,1),(1,2),(2,4),(2,6)\big\}$ is split into two folds:

$$F_1 = \{(x_1, y_1) = (0,0),\ (x_2, y_2) = (1,1),\ (x_3, y_3) = (2,4)\},$$
$$F_2 = \{(x_4, y_4) = (0,0),\ (x_5, y_5) = (1,2),\ (x_6, y_6) = (2,6)\}.$$

The two $(0,0)$ pairs are distinct observations with the same values. For the 2-fold cross-validation calculation, the fitted models from each training fold are given below:

Training fold	**Linear fit**	**Quadratic fit**
F_1	$\hat{y} = -\frac{1}{3} + 2x$	$\hat{y} = x^2$
F_2	$\hat{y} = -\frac{1}{3} + 3x$	$\hat{y} = x^2 + x$

Using these estimates, compute the 2-fold CV estimate of test MSE for the linear and quadratic models.

(c) (6 points) Based on the two cross-validation MSE values from part (b), which model would you choose for prediction? Briefly justify your choice.

Problem 3 (16 points). Bootstrap

Suppose that we roll a fair 6-sided die 5 times, and observe the following sample of five values:

$$x = \{1, 2, 3, 4, 6\}.$$

Let the statistic of interest be the sample median. For a sample of size 5, the median is the middle value, i.e., the 3rd value after sorting the sample in increasing order.

(a) (10 points) Five bootstrap samples are shown below:

	Bootstrap 1	Bootstrap 2	Bootstrap 3	Bootstrap 4	Bootstrap 5
Value 1	1	1	1	2	1
Value 2	1	2	2	3	4
Value 3	2	2	3	4	6
Value 4	2	3	4	4	6
Value 5	3	6	6	6	6

Construct a normal-approximation 95% confidence interval for the population median, centered at the original sample median, using $\widehat{\mathrm{SE}}_{\mathrm{boot}}$ computed from the five bootstrap medians.
(*Hint*: $z_{0.9} \approx 1.28$, $z_{0.95} \approx 1.64$, $z_{0.975} \approx 1.96$, $z_{0.99} = 2.33$.)

(b) (6 points) In this context, how should we interpret "95%" in a 95% confidence interval? State what is random and which probability is intended to be approximately 95% clearly and succinctly.

Problem 4 (40 points + 5 bonus). Model selection and regularization

You have 4 candidate predictors X_1, X_2, X_3, X_4 and a response Y. The following table gives the *training Residual Sum of Squares (RSS)* for all $2^4 = 16$ possible subsets, computed from a sample of size $n = 10$.

0 predictor		**1 predictor**		**2 predictors**		**3 predictors**		**4 predictors**	
Predictors	**RSS**	**Predictors**	**RSS**	**Predictors**	**RSS**	**Predictors**	**RSS**	**Predictors**	**RSS**
$\emptyset$	100.0	X_1	50.0	X_1, X_2	48.0	X_1, X_2, X_3	36.5	X_1, X_2, X_3, X_4	35.5
		X_2	55.0	X_1, X_3	46.0	X_1, X_2, X_4	47.5		
		X_3	60.0	X_1, X_4	49.0	X_1, X_3, X_4	36.0		
		X_4	70.0	X_2, X_3	38.0	X_2, X_3, X_4	37.5		
				X_2, X_4	54.0				
				X_3, X_4	56.0				

Hint: Recall $R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$ and $R^2_{\text{adj}} = 1 - \frac{n-1}{\text{TSS}} \cdot \frac{\text{RSS}}{n-p-1}$ for a model with p predictors. Here, $n = 10$ and $\text{TSS} = 100$, so comparing R^2_{adj} across models is equivalent to comparing $\text{RSS}/(n-p-1)$.

(a) (10 points) Using **best subset selection**, identify the best subset for each model size $p = 0, 1, 2, 3, 4$, using RSS among subsets of that size. Then use R^2_{adj} to choose one final model among these five subsets.

(b) (10 points) Using **forward stepwise selection**, list the selected subset at each size $p = 0, 1, 2, 3, 4$. Then use R^2_{adj} along this forward stepwise path to choose one final model.

(c) (10 points) A regularized regression method produces the following coefficient estimates and cross-validation errors:

λ	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	CV MSE
0.1	2.8	1.2	0.4	0.3	20
1.0	2.0	0	0.8	0	23
10.0	0	0	0	0	55

Is this more likely ridge regression or lasso? Briefly justify. If prediction accuracy is the only goal, which λ would you choose? If interpretability is more important, how might that affect the choice of λ?

(d) (10 points) As λ increases from 0 to ∞, how do you expect each of the following to behave?

Pick among the five options: (1) "Remain constant," (2) "Generally increase," (3) "Generally 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 error (training MSE)

(ii) Expected test error (test MSE)

(iii) (Squared) bias

(iv) Variance

(v) Irreducible error

(e*) (5 bonus points) Suppose X_1 and X_2 are nearly identical and both are strongly associated with Y. Without doing calculations, explain how ridge and lasso might treat these two predictors differently. Why might one behavior be preferable for prediction, and the other for interpretation?

Problem 5 (20 points + 5 bonus). Multiple testing

(a) (8 points) 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) **(4 points)** Suppose we aim to control the *familywise error rate (FWER)* at level α (e.g. 0.05). Express this goal as an inequality, referring to the tables above.

(ii) **(4 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, referring to the tables above.

(b) (12 points) Suppose we test $m = 8$ null hypotheses with the following p-values, already listed in increasing order:

$$H_{0,1} : 0.001, \quad H_{0,2} : 0.004, \quad H_{0,3} : 0.020, \quad H_{0,4} : 0.024,$$
$$H_{0,5} : 0.045, \quad H_{0,6} : 0.080, \quad H_{0,7} : 0.125, \quad H_{0,8} : 0.200.$$

(i) **(4 points)** With *no correction*, which hypotheses are rejected at significance level $\alpha = 0.05$?

(ii) **(4 points)** With the *Bonferroni correction* to achieve FWER $\leq \alpha$, which hypotheses are rejected?

(iii) **(4 points)** At FDR level $q = 0.05$, apply the Benjamini–Hochberg procedure to the same 8 p-values. Which hypotheses are rejected? (*Hint:* compare $p_{(j)}$ to qj/m.)

(c*) (5 bonus points) Suppose Benjamini–Hochberg at $q = 0.10$ rejects 40 hypotheses. Does this guarantee that at most 4 of the 40 rejected hypotheses are false discoveries? Explain briefly.