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**. Problems 1–5 total **120 points**; Problem 6 offers **up to 10 bonus points**. Please confirm that you have all 8 pages.

- Make sure to clearly write your name and student ID above.
- Present answers succinctly. Show relevant steps unless a question requests answers only.
- If necessary, round only final 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	
Problem 6 (bonus)	
Total	

Problem 1 (30 points). Basic concepts

For parts (a)–(c), **circle all correct options**. Each part has **two** correct options. Correct selections earn points, and incorrect selections lose points. For parts (d)–(f), explain only where requested.

(a) (5 points) At a fixed input x_0, let $Y_0 = f(x_0) + \varepsilon_0$, where ε_0 is independent of the training data, with mean zero and variance σ^2. The expected squared prediction error decomposes as

$$\mathbb{E}[(Y_0 - \widehat{f}(x_0))^2] = \big(\mathbb{E}[\widehat{f}(x_0)] - f(x_0)\big)^2 + \mathrm{Var}(\widehat{f}(x_0)) + \text{irreducible error}.$$

Which statements are correct?

A. Irreducible error is the variance of the new-response noise.
B. Prediction variance measures sensitivity to the training sample.
C. Observed training MSE equals squared bias.
D. Greater model flexibility always reduces expected test MSE.

(b) (5 points) An i.i.d. dataset contains n predictor–response pairs (x_i, y_i). Which statements about ordinary **(paired) bootstrap resampling** of size n are correct?

A. The n rows are drawn independently with replacement.
B. Every original row appears exactly once.
C. Each selected predictor–response pair is kept together.
D. The resample consists of n new observations from the population.

(c) (5 points) Suppose that we have 100 hypotheses, $H_1, \ldots, H_{100}$ to test. Define the false-discovery proportion as zero when there are no rejections. Which statements are correct?

A. FDR is the expected false-discovery proportion.
B. FDR control at 0.10 guarantees that at most 10% of all rejections are false in every dataset.
C. FDR control at 0.10 guarantees FWER at most 0.10.
D. FWER control at 0.10 guarantees FDR at most 0.10.

(d) (5 points) Ridge regression of Y on X is fitted to training data with $\overline{y} = 18$, with the intercept unpenalized. **What value does the prediction at $x = 2$ approach as the slope penalty $\lambda \to \infty$?**

(e) (5 points) A lasso fit assigns a nonzero coefficient to X_1. Does this alone establish that X_1 is statistically significant at the 5% level? Answer **Yes** or **No** (3 points) and give a **one-sentence reason** (2 points).

(f) (5 points) True or False: State your answer (3 points) and give a **one-sentence reason** (2 points).

> "Because LOOCV has no random split, its error estimate cannot vary across independently collected datasets."

Problem 2 (20 points). Regularization and cross-validation

Parts (a) and (b) concern separate analyses.

(a) (10 points) Ridge regression is fitted to the **same training observations and predictors**, with an unpenalized intercept. As λ increases from 0 toward ∞, how do you expect each quantity below to behave? Use the **typical qualitative bias–variance patterns discussed in lecture**.

For each quantity, write the number of the most appropriate option: (1) Remain constant; (2) Generally increase; (3) Generally decrease; (4) Decrease, then increase (U-shaped); or (5) Increase, then decrease (inverted U-shaped). Options may be used more than once.

Each answer is worth 2 points; no justification is required.

(i) Training MSE.

(ii) Expected test MSE.

(iii) Squared bias.

(iv) Variance of the fitted prediction.

(v) Irreducible error.

(b) (10 points) Six observations are divided into two folds:

$$F_1 = \{(0,1),(1,4),(2,13)\}, \qquad F_2 = \{(0,1),(1,7),(2,19)\}.$$

The two $(0,1)$ pairs are distinct observations. The linear and quadratic OLS fits, each with an intercept, trained on each fold are:

Training fold	Linear fit	Quadratic fit
F_1	$\widehat{y} = 6x$	$\widehat{y} = 1 + 3x^2$
F_2	$\widehat{y} = 9x$	$\widehat{y} = 1 + 3x + 3x^2$

Compute the two-fold CV MSEs for the linear and quadratic models.

Problem 3 (20 points). Bootstrap uncertainty

An i.i.d. sample of delivery times, in minutes, is $6, 8, 9, 10, 12$. The table shows five bootstrap resamples, each formed by five independent draws with replacement from the original sample and then sorted.

Resample	Sorted observations
1	$6, 6, 6, 8, 12$
2	$6, 6, 6, 9, 10$
3	$6, 8, 8, 10, 12$
4	$8, 9, 10, 10, 12$
5	$6, 10, 10, 12, 12$

(a) (10 points) Estimate the bootstrap standard error of the sample median.

Hint: For five observations, the median is the third value after sorting.

(b) (10 points) Use the normal-approximation rule

$$\widehat{m} \ \pm \ 1.96\,\widehat{\mathrm{SE}}_{\mathrm{boot}},$$

where $\widehat{m}$ is the **original sample median**. Report the **approximate 95% confidence interval for the population median**. In **one sentence**, explain what the intended "95% confidence level" means.

Problem 4 (20 points). Subset selection

OLS models with an intercept are fitted to the same $n = 10$ observations. The training RSS values for all subsets of three candidate predictors are below; $\varnothing$ denotes the intercept-only model.

Predictors	RSS	Predictors	RSS
$\varnothing$	100	X_1, X_2	30
X_1	40	X_1, X_3	40
X_2	46	X_2, X_3	22
X_3	76	X_1, X_2, X_3	22

(a) (12 points) Apply **best subset selection (6 points)** and **forward stepwise selection (6 points)**, using training RSS to compare eligible subsets at each size. For each method, **list the selected subset for $k = 0, 1, 2, 3$** and **choose one final model from its four candidates using the largest adjusted R^2**. Show the criterion values used to choose each final model; shared calculations need not be repeated.

Hint: For k predictors, $R^2_{\text{adj}} = 1 - \frac{n-1}{\text{TSS}} \frac{\text{RSS}}{n-k-1}$. So comparing adjusted R^2 is equivalent to comparing $\text{RSS}/(n-k-1)$.

(b) (8 points) An analyst selects two predictors by best subset selection using all observations in the dataset, then performs five-fold CV, fitting OLS with that same pair of predictors in every data fold. This does **not** correctly assess the complete selection-and-fitting procedure.

Describe the corrected CV procedure in **at most two sentences**, specifying which observations are used for predictor selection, fitting, and evaluation within each fold.

Problem 5 (30 points). From predictive models to conclusions

A team predicts energy demand using five standardized predictors $X_1, \ldots, X_5$. It fits either ridge or lasso at three penalty values, always leaving the intercept unpenalized. The table gives full-development-data slopes and CV MSEs obtained using the **same four folds**; the displayed zeros are exact.

λ	$\widehat{b}_1$	$\widehat{b}_2$	$\widehat{b}_3$	$\widehat{b}_4$	$\widehat{b}_5$	CV MSE (rounded)
3	0.3	2.4	0.2	-1.4	0.1	1.17
15	0	2.0	0	-1.0	0	2.41
75	0	0	0	0	0	10.57

(a) (10 points) Is this more likely **ridge regression** or **lasso**? If the team wants the final model to use **at most two predictors**, **which of the listed λ values should the team use?** Justify both answers in **at most two sentences**.

(b) (8 points) The team drops predictors with zero slopes in the selected fit, then refits OLS on the remaining predictors using the **same development observations** and an intercept. An analyst claims,

> "Because OLS minimizes training RSS, this refit must have lower expected prediction error on new observations."

State whether this claim is **True** or **False** and justify your answer in **at most two sentences**, distinguishing training fit from prediction on new observations from the same population.

(c) (12 points) A **separate confirmatory study** tests eight prespecified associations with energy demand. These eight tests form the entire testing family and were not selected using the fitted coefficients above. The valid p-values are already in increasing order:

Null hypothesis	$H_{0,1}$	$H_{0,2}$	$H_{0,3}$	$H_{0,4}$	$H_{0,5}$	$H_{0,6}$	$H_{0,7}$	$H_{0,8}$
p-value	0.003	0.006	0.008	0.012	0.014	0.016	0.022	0.040

At family-wise level 0.05, **report the rejection sets for Bonferroni (6 points) and Holm (6 points)**, showing the relevant cutoff comparisons.

Problem 6 (10 bonus points). Optional bonus questions

These questions are independent of the required questions and of each other.

(a) (5 points) Let $X_1, \ldots, X_n$ be i.i.d. uniform on $(0, \theta)$, where $\theta > 0$ is unknown. An analyst estimates θ by the sample maximum $M = \max_i X_i$. Each ordinary bootstrap resample draws n observations with replacement from the original sample, and the analyst records its maximum.

A proposed 95% interval has endpoints equal to the 2.5th and 97.5th percentiles of these bootstrap maxima. **What is this interval's actual coverage probability for θ?** Justify by comparing its upper endpoint with M and θ.

(b) (5 points) On fixed training data, a one-predictor lasso fit minimizes

$$Q(b_0, b) = \sum_{i=1}^{n} (y_i - b_0 - b x_i)^2 + \lambda |b|, \qquad \lambda > 0.$$

An analyst adds an identical copy of the predictor and minimizes

$$\widetilde{Q}(a_0, a_1, a_2) = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i - a_2 x_i)^2 + \lambda (|a_1| + |a_2|).$$

Can adding this duplicate strictly reduce the minimum penalized objective value? Justify by comparing fitted values and penalties.