STA 35C Statistical Data Science III

Practice Midterm 2 Solution

Instructor: Dogyoon Song

Problem 1: Solution (24 points)

(1) False. A single train/validation split (one-shot approach) typically yields a *higher*-variance error estimate because it relies on just one particular split of the data. In contrast, 5-fold CV averages multiple splits, usually leading to a more stable (lower-variance) test-error estimate.

(2) True. In a bootstrap sample (size n, drawn with replacement), some original points appear multiple times, while others are omitted; e.g., $\{x_1, x_1, x_3, x_5, ...\}$.

(3) False. Forward stepwise *starts with no predictors* and adds them one by one. (Starting with all predictors and removing them is *backward* stepwise.)

(4) True. Because of the L1 penalty geometry, a large λ can drive some coefficients exactly to zero, effectively performing variable selection.

(5) False. Performing many tests at $\alpha = 0.05$ inflates the chance of a false positive (Type I error), not the power.

(6) True. Overfitted models often show very low training error but degrade significantly on test or cross-validation data, indicating poor generalization.

Problem 2: Solution (18 points)

(a) (12 points) Two-Fold CV with Four Data Points We have four points:

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

split into two folds:

$$\text{Fold 1: } \{(2, 3), (7, 10)\}, \quad \text{Fold 2: } \{(4, 5), (9, 14)\}.$$

We compare two models: - **Linear**: $f(x) = \beta_0 + \beta_1 x$, - **Quadratic**: $g(x) = \beta_0 + \beta_1 x^2$.

Linear Model

- **Train on Fold 1**, test on Fold 2:

$$\beta_1 = \frac{10 - 3}{7 - 2} = 1.4, \ \beta_0 = 3 - 1.4 \times 2 = 0.2.$$

Predict on $(4,5)$ and $(9,14)$:

$$\hat{f}(4) = 5.8 \text{ (error} = 5 - 5.8 = -0.8,\ e^2 = 0.64), \quad \hat{f}(9) = 12.8 \text{ (error} = 14 - 12.8 = 1.2,\ e^2 = 1.44).$$

$\text{MSE}_1 = \frac{0.64+1.44}{2} = 1.04.$

- **Train on Fold 2**, test on Fold 1:

$$\beta_1 = \frac{14-5}{9-4} = 1.8,\ \beta_0 = 5 - 1.8 \times 4 = -2.2.$$

Predict on $(2,3)$ and $(7,10)$:

$$\hat{f}(2) = 1.4\ (e = 1.6,\ e^2 = 2.56), \quad \hat{f}(7) = 10.4\ (e = -0.4,\ e^2 = 0.16).$$

$\text{MSE}_2 = \frac{2.56+0.16}{2} = 1.36.$

Hence, 2-fold CV MSE for the linear model is

$$\tfrac{1.04+1.36}{2} = 1.20.$$

Quadratic Model

- **Train on Fold 1**, test on Fold 2:

$$3 = \beta_0 + 4\,\beta_1, \quad 10 = \beta_0 + 49\,\beta_1 \implies \beta_1 = \tfrac{7}{45},\ \beta_0 \approx 2.376.$$

Predict $(4,5)$, $(9,14)$:

$$\hat{g}(4) = 4.872,\ e^2 = (5 - 4.872)^2 = 0.01638, \quad \hat{g}(9) = 15.012,\ e^2 = (14 - 15.012)^2 = 1.02414.$$

$\text{MSE}_1 \approx 0.52026.$

- **Train on Fold 2**, test on Fold 1:

$$5 = \beta_0 + 16\,\beta_1, \quad 14 = \beta_0 + 81\,\beta_1 \implies \beta_1 = \tfrac{9}{65},\ \beta_0 \approx 2.78464.$$

Predict $(2,3)$, $(7,10)$:

$$\hat{g}(2) = 3.33848,\ e^2 = 0.11457, \quad \hat{g}(7) = 9.56918,\ e^2 = 0.18560.$$

$\text{MSE}_2 = \frac{0.11457+0.18560}{2} = 0.150085.$

Hence, 2-fold CV MSE for the quadratic model is

$$\tfrac{0.52026+0.150085}{2} \approx 0.335.$$

Conclusion Since $1.20 > 0.335$, the **quadratic model** is preferred based on 2-fold CV.

(b) (6 points) k-Fold CV vs. LOOCV

- **Advantages of k-fold:**
  - Less computation than LOOCV (fewer total fits).
  - Typically lower variance in the estimated error than a single train/test split.
- **Disadvantages:**
  - Slightly more bias than LOOCV, since each training set is smaller than $n-1$.
  - Must decide on the hyperparameter k; results can vary if k is too small or large.

Problem 3: Solution (20 points)

We have 5 data points (not explicitly shown), plus 3 bootstrap samples. Our tasks involve computing *sample means* and using them to form a confidence interval.

(a) (8 points) Sample Means - Let the **original sample** be $\{2, 3, 5, 7, 8\}$. Then

$$\hat{\mu}_{\text{orig}} = \frac{2+3+5+7+8}{5} = 5.0.$$

- **Bootstrap 1**: The table shows $\{2, 2, 5, 7, 8\}$ (top to bottom in column 2).

$$\hat{\mu}_{B_1} = \frac{2+2+5+7+8}{5} = 4.8.$$

- **Bootstrap 2**: $\{3, 5, 7, 8, 8\}$ etc. Suppose that column 3 reads $\{3, 5, 7, 8, 8\}$ (the middle row is 5, etc.). Then

$$\hat{\mu}_{B_2} = \frac{3+5+7+8+8}{5} = 6.2.$$

- **Bootstrap 3**: $\{2, 3, 5, 5, 8\}$ yields

$$\hat{\mu}_{B_3} = \frac{2+3+5+5+8}{5} = 4.6.$$

(b) (6 points) Std. Dev. of the Four Means We have four mean values:

$$\hat{\mu}_{\text{orig}} = 5.0, \quad \hat{\mu}_{B_1} = 4.8, \quad \hat{\mu}_{B_2} = 6.2, \quad \hat{\mu}_{B_3} = 4.6.$$

Compute their standard deviation:

$$\bar{m} = \frac{5.0+4.8+6.2+4.6}{4} = 5.15,$$

$$s_{\hat{\mu}} = \sqrt{\frac{(5.0-5.15)^2 + (4.8-5.15)^2 + (6.2-5.15)^2 + (4.6-5.15)^2}{4-1}} \approx 0.719.$$

(c) (6 points) 95% CI for μ

- **Percentile approach**: If you had many bootstraps, you'd sort their means and pick the 2.5% and 97.5% quantiles as the confidence bounds. With only 3 bootstraps, we can't truly do percentile method reliably.
- **Normal approximation approach**:

  $$\hat{\mu}_{\text{orig}} \pm z_{0.975} \times s_{\hat{\mu}} \approx 5.0 \pm 1.96 \times 0.71.$$

  That might give an interval roughly $(3.59,\ 6.41)$.

Problem 4: Solution (20 points)

(a) (8 points) Best Subset.

- $k = 0$: Choose $\emptyset$ (RSS=40.0).
- $k = 1$: Minimizes RSS at X_1 (RSS=10.0).
- $k = 2$: Minimizes RSS at X_1, X_2 (RSS=8.0).
- $k = 3$: Full model X_1, X_2, X_3 (RSS=7.5).

(b) (6 points) Forward & Backward Stepwise.

(i) Forward:

- Start with $\emptyset$. Among $\{X_1\}, \{X_2\}, \{X_3\}$, best is X_1 (RSS=10.0).
- Then among $\{X_1, X_2\}, \{X_1, X_3\}$, best is (X_1, X_2) (RSS=8.0).
- Checking (X_1, X_2, X_3) is next: RSS=7.5, so final includes all three if we keep going until no improvement is meaningful.

(ii) Backward:

- Start with (X_1, X_2, X_3) (RSS=7.5).
- Removing $X_3 \Rightarrow (X_1, X_2)$ RSS=8.0, removing $X_2 \Rightarrow (X_1, X_3)$ RSS=12.0, removing $X_1 \Rightarrow (X_2, X_3)$ RSS=14.5. The best removal is X_3.
- Now we have (X_1, X_2). Could remove $X_1 \Rightarrow RSS = 15$, or $X_2 \Rightarrow RSS = 10$; best removal is X_2, and we move to (X_1).

(c) (6 points) Forward Stepwise vs. Best Subset.

- **Advantage (Forward):** Much faster in high p settings, not enumerating all 2^p subsets.
- **Drawback:** It can miss the overall best subset since it never revisits earlier decisions once it adds predictors.

Problem 5: Solution (20 points)

(a) (10 points) Ridge vs. Lasso Coefficients

(i) **Method A** is Lasso, because it sets $\hat{\beta}_2 = 0$. **Method B** is Ridge, which shrinks β_2 to 1.2 rather than zero.

(ii) Lasso can drive some coefficients exactly to zero, indicating X_2 is either less important or strongly correlated with X_1. Ridge merely reduces β_2 to 1.2, implying X_2 still has some effect but is penalized away from its OLS value.

(b) (10 points) CV for Different λ Values

	Ridge			**Lasso**		
λ	0.1	1.0	5.0	0.1	1.0	5.0
CV Error	0.90	**0.88**	0.93	**0.85**	0.86	0.95

(i) **Ridge**: The best λ is 1.0 (CV error 0.88). **Lasso**: The best λ is 0.1 (CV error 0.85).

(ii) At Lasso $\lambda = 1.0$, 2 of 10 predictors are set to zero (CV error 0.86). At $\lambda = 0.1$, none are zero (CV error 0.85). A difference of 0.01 in error may be negligible, so the simpler model (fewer predictors) might be preferable unless the absolute lowest test error is critical.

Problem 6: Solution (18 points + 2 bonus)

(a) (10 points) 10 p-values, no correction vs. Bonferroni.

$$\{0.001,\ 0.01,\ 0.02,\ 0.03,\ 0.04,\ 0.10,\ 0.15,\ 0.20,\ 0.25,\ 0.50\}$$

- **No Correction**: All p-values below 0.05 are declared significant, so we reject $H_{0,1}$ through $H_{0,5}$ (5 rejections).
- **Bonferroni**: Adjusted $\alpha^* = \frac{0.05}{10} = 0.005$. Then only $p = 0.001 < 0.005$ is significant, so 1 rejection.
- **Comment**: Bonferroni is more conservative, drastically reducing the number of discoveries.

(b) (8 points) 5 p-values, BH at FDR=5%.

$$\{0.002,\ 0.01,\ 0.04,\ 0.09,\ 0.20\}.$$

(i) Sort them: $0.002, 0.01, 0.04, 0.09, 0.20$.

(ii) BH critical values for each $p_{(i)}$ are $\alpha \frac{i}{m} = 0.05 \times \frac{i}{5} = 0.01i$.

$$i = 1 : 0.01; \quad i = 2 : 0.02; \quad i = 3 : 0.03; \quad i = 4 : 0.04; \quad i = 5 : 0.05.$$

(iii) Compare in ascending order:

$$\begin{aligned} p_{(1)} &= 0.002 < 0.01 \quad \text{(reject)}, \\ p_{(2)} &= 0.01 < 0.02 \quad \text{(reject)}, \\ p_{(3)} &= 0.04 > 0.03 \quad \text{(stop)}. \end{aligned}$$

Hence we reject $H_{0,1}$ and $H_{0,2}$ but not the rest.

(c*) (2 bonus points) FDR vs. FWER. FDR controls the fraction of false positives among the rejected hypotheses, typically more powerful when testing many hypotheses. FWER (Bonferroni/Holm) aims to keep the probability of *any* false positive near zero, so it may be too conservative in large-scale testing. FDR is generally preferred in scenarios like genomics with thousands of tests, where some false positives are tolerable, but we want to control their *proportion*.