STA 35C Statistical Data Science III

Midterm exam 2 solution

Instructor: Dogyoon Song

Problem 1: Solution (20 points)

(a) Using more folds in K-fold CV increases computational cost.

True. Each fold requires retraining the model, so more folds means more total fits.

(b) Every point must appear at least once in each bootstrap sample.

False. Sampling with replacement can skip some points entirely, while others are duplicated.

(c) Forward stepwise can remove a predictor added earlier.

False. Standard forward stepwise *only adds* predictors; it doesn't drop them once included.

(d) Lasso shrinks correlated predictors together, whereas Ridge might zero one out.

False. It's usually Ridge that "groups" correlated predictors in similar shrinkage, while Lasso may set one to zero and keep another.

(e) FDR at 0.05 means no false positives occur with 95% probability.

False. FDR ≤ 0.05 ensures the *expected fraction* of false positives is at most 5%, not that we have zero false positives 95% of the time.

Problem 2: Solution (20 points)

(a) (6 pts) 5-fold CV vs. single split

- **Advantage:** 5-fold CV averages multiple train/validation splits, typically reducing the variance of the estimated test error compared to a single split.
- **Disadvantage:** It is more computationally expensive (you must train a model 5 times instead of once).

(b) (14 pts total) Comparing two models via LOOCV

We have three data points ($n = 3$):

$$(-2, 1),\ (0, 3),\ (3, 9).$$

(i) **(10 pts)** Leave-One-Out CV means: each time leave out 1 point, train on the other 2, predict the omitted point, compute squared error, then average over all 3 folds.

- **LOOCV for *Linear* model $f(x) = \beta_0 + \beta_1 x$.**
  - **Fold 1**: Omit $(-2, 1)$. Train on $(0, 3)$ and $(3, 9)$.
    - $\beta_1 = \frac{9-3}{3-0} = 2,\ \beta_0 = 3 - (2 \times 0) = 3.$
    - Predict omitted point $(-2, 1)$: $\hat{f}(-2) = 3+2(-2) = 3-4 = -1,\ e = 1-(-1) = 2,\ e^2 = 4.$
  - **Fold 2**: Omit $(0, 3)$. Train on $(-2, 1), (3, 9)$.
    - $\beta_1 = \frac{9-1}{3-(-2)} = \frac{8}{5} = 1.6,\ \beta_0 = 1 - 1.6 \times (-2) = 1 + 3.2 = 4.2.$
    - Predict omitted point $(0, 3)$: $\hat{f}(0) = 4.2,\ e = 3 - 4.2 = -1.2,\ e^2 = 1.44.$
  - **Fold 3**: Omit $(3, 9)$. Train on $(-2, 1), (0, 3)$.
    - $\beta_1 = \frac{3-1}{0-(-2)} = \frac{2}{2} = 1,\ \beta_0 = 3 - (1 \times 0) = 3.$
    - Predict omitted point $(3, 9)$: $\hat{f}(3) = 3 + (1 \times 3) = 6,\ e = 9 - 6 = 3,\ e^2 = 9.$

  So LOOCV MSE for linear model:

$$\text{MSE}_{\text{linear}} = \frac{4 + 1.44 + 9}{3} = \frac{14.44}{3} = 4.8133 \approx 4.81.$$

- **LOOCV for *Quadratic* model $g(x) = \beta_0 + \beta_1 x^2$.**
  - **Fold 1**: Omit $(-2, 1)$. Train on $(0, 3), (3, 9)$.
    - $\beta_1 = \frac{9-3}{3^2-0} = \frac{2}{3},\ \beta_0 = 3 - (\frac{2}{3} \times 0) = 3.$
    - Predict omitted point $(-2, 1)$: $\hat{g}(-2) = 3 + 0.6667 \times 4 = 3 + 2.6667 = 5.6667,\ e = 1 - 5.6667 = -4.6667,\ e^2 \approx 21.78.$
  - **Fold 2**: Omit $(0, 3)$. Train on $(-2, 1), (3, 9)$.
    - $\beta_1 = \frac{9-1}{3^2-(-2)^2} = \frac{8}{5},\ \beta_0 = 1 - (\frac{8}{5} \times (-2)^2) = -\frac{27}{5}.$
    - Predict omitted point $(0, 3)$: $\hat{g}(0) = -5.4 + 1.6 \times 0 = -5.4,\ e = 3 - (-5.4) = 8.4,\ e^2 = 70.56.$
  - **Fold 3**: Omit $(3, 9)$. Train on $(-2, 1), (0, 3)$.
    - $\beta_1 = \frac{3-1}{0^2-(-2)^2} = -\frac{1}{2},\ \beta_0 = 3 - (-\frac{1}{2} \times (0)^2) = 3.$
    - Predict omitted point $(3, 9)$: $\hat{g}(3) = 3 + 0.5 \times 3^2 = -1.5,\ e = 9 - (-1.5) = 10.5,\ e^2 = 110.25.$

  So LOOCV MSE for quadratic:

$$\text{MSE}_{\text{quad}} = \frac{21.78 + 70.56 + 110.25}{3} = \frac{202.59}{3} = 67.53 \approx 67.53.$$

(ii) **(4 pts)**

$$\text{MSE}_{\text{linear}} \approx 4.81, \quad \text{MSE}_{\text{quad}} \approx 67.53.$$

The linear model has a much smaller LOOCV MSE, so we select $f(x)$ (the linear one).

Problem 3: Solution (20 points)

We have 5 flips of a coin, observed H, T, T, H, T (2 heads out of 5).

(a) (6 points) The original dataset has 2 heads, 3 tails. Thus, $\Pr(H) = 2/5$ and $\Pr(T) = 3/5$. Thus, the probability of drawing the exact same sequence (H, T, T, H, T) is

$$\Pr(H) \times \Pr(T) \times \Pr(T) \times \Pr(H) \times \Pr(T) = \Pr(H)^2 \times \Pr(T)^3 = \left(\frac{2}{5}\right)^2 \left(\frac{3}{5}\right)^3 = \frac{108}{3125}.$$

(b) (8 points) First of all, we have $\hat{p} = 0.4$ from the original sample.

We have 4 bootstrap samples of 5 flips. Observe that

$$\hat{p}_1^* = 0.6, \quad \hat{p}_2^* = 0.6, \quad \hat{p}_3^* = 0.4, \quad \hat{p}_4^* = 0.4.$$

Thus, the sample standard deviation of $\hat{p}_1^*, \hat{p}_2^*, \hat{p}_3^*, \hat{p}_4^*$ is

$$\hat{\sigma} = \sqrt{\frac{1}{4-1}\sum_{i=1}^{4}(\hat{p}_i^* - \bar{p}^*)} = \sqrt{\frac{4}{300}} \approx 0.115.$$

Ultimately, we can construct a 95% confidence interval via a normal approximation:

$$\hat{p} \pm z_{0.975} \times \hat{\sigma},$$

where $z_{0.975} \approx 1.96$. Therefore, we obtain a 95% CI, approximately $[0.174, 0.626]$.

(c) (6 points) A 95% bootstrap confidence interval means that if we repeated the entire experiment + bootstrap procedure many times, about 95% of such intervals would contain the true p. It's about *long-run* coverage of p under repeated sampling.

Problem 4: Solution (20 points)

We have $n = 11$ data points, TSS $= 100$, and 4 predictors. The RSS table is given.

(a) (7 points) Best Subset

- For each model size $k = 0, 1, 2, 3, 4$, pick the subset with the smallest RSS:

$$k = 0: \quad \emptyset \quad (\text{RSS} = 100).$$
$$k = 1: \quad X_2 \quad (\text{RSS} = 40).$$
$$k = 2: \quad X_1, X_4 \quad (\text{RSS} = 25).$$
$$k = 3: \quad X_1, X_3, X_4 \quad (\text{RSS} = 20).$$
$$k = 4: \quad (X_1, X_2, X_3, X_4) \ (\text{RSS} = 19).$$

- Here $n = 11$, so $(n-1) = 10$. For each k, we plug in:

$$\text{adj}\, R_k^2 = 1 - \frac{\text{RSS}_k}{100} \times \frac{10}{11-k-1}.$$

Hence:

$$k = 0: \ \text{RSS} = 100 \quad \Rightarrow R^2_{\text{adj},0} = 1 - \frac{100}{100} \times \frac{10}{10} = 1 - 1 = 0.$$
$$k = 1: \ \text{RSS} = 40 \quad \Rightarrow R^2_{\text{adj},1} = 1 - \frac{40}{100} \times \frac{10}{9} = 1 - 0.4 \times 1.1111 = 1 - 0.4444 = 0.5556.$$
$$k = 2: \ \text{RSS} = 25 \quad \Rightarrow R^2_{\text{adj},2} = 1 - 0.25 \times 1.25 = 1 - 0.3125 = 0.6875.$$
$$k = 3: \ \text{RSS} = 20 \quad \Rightarrow R^2_{\text{adj},3} = 1 - \frac{20}{100} \times \frac{10}{7} = 1 - 0.2 \times 1.4286 = 1 - 0.2857 = 0.7143.$$
$$k = 4: \ \text{RSS} = 19 \quad \Rightarrow R^2_{\text{adj},4} = 1 - \frac{19}{100} \times \frac{10}{6} = 1 - 0.19 \times 1.6667 = 1 - 0.3167 = 06833.$$

- Then we might pick the $k = 3$ model (X_1, X_3, X_4) because it has the largest R^2_{adj} among the k-wise best models.

(b) (7 points) Forward Stepwise Here we start from the null model ($\emptyset$, RSS=100) and at each size k, we pick only from those subsets containing the previously chosen subset. Then we compute R^2_{adj} for the model we get at each step, and finally pick the best among them.

- Path of forward selection:
  - $k = 0$: Subset is $\emptyset$ (RSS=100).
  - $k = 1$: Among singletons, X_2 yields RSS=40 (best). So we choose (X_2).
  - $k = 2$: From (X_2), test adding $X_1 \to 30, X_3 \to 35, X_4 \to 32$.
    The best is (X_1, X_2) with RSS=30.
  - $k = 3$: From (X_1, X_2), test adding $X_3 \to 28$ or $X_4 \to 21$.
    The best is (X_1, X_2, X_4) with RSS=21.
  - $k = 4$: From (X_1, X_2, X_4), adding X_3 yields RSS=19, so final is (X_1, X_2, X_3, X_4).
- Here $n = 11$, so $(n - 1) = 10$. For each k, we plug in:

$$\text{adj}\, R^2_k = 1 - \frac{\text{RSS}_k}{100} \times \frac{10}{11 - k - 1}.$$

Hence:

$$\begin{aligned}
k = 0: &\quad \emptyset,\ \text{RSS} = 100 \\
&\Rightarrow R^2_{\text{adj},0} = 0. \\
k = 1: &\quad (X_2),\ \text{RSS} = 40 \\
&\Rightarrow R^2_{\text{adj},1} = 1 - \frac{40}{100} \times \frac{10}{9} = 1 - 0.4 \times 1.1111 = 1 - 0.4444 = 0.5556. \\
k = 2: &\quad (X_1, X_2),\ \text{RSS} = 30 \\
&\Rightarrow R^2_{\text{adj},2} = 1 - \frac{30}{100} \times \frac{10}{8} = 1 - 0.30 \times 1.25 = 1 - 0.375 = 0.625. \\
k = 3: &\quad (X_1, X_2, X_4),\ \text{RSS} = 22 \\
&\Rightarrow R^2_{\text{adj},3} = 1 - \frac{21}{100} \times \frac{10}{7} = 1 - 0.21 \times 1.4286 = 1 - 0.3000 = 0.7000. \\
k = 4: &\quad (X_1, X_2, X_3, X_4),\ \text{RSS} = 19 \\
&\Rightarrow R^2_{\text{adj},4} = 1 - \frac{19}{100} \times \frac{10}{6} = 1 - 0.19 \times 1.6667 = 1 - 0.3167 = 06833.
\end{aligned}$$

- Then we might pick the $k = 3$ model (X_1, X_2, X_4) because it has the largest R^2_{adj} among the k-wise best models. Note that this is different from the subset chosen via the Best Subset Selection due to the greedy nature of Forward Stepwise Selection procedure

(c) (6 points) Advantage/Disadvantage of Backward Stepwise

- **Advantage**: Less expensive than enumerating all 2^p subsets, provided $n > p$. You systematically remove unneeded predictors.
- **Disadvantage**: May fail to find the global best subset if a crucial predictor was dropped early in the sequence. Also requires $n > p$ so you can start with the full model.

Problem 5: Solution (20 points)

(a) (10 points) Behavior as λ increases (Ridge)

(i) **Training MSE**: *Steadily increases* (larger λ imposes more shrinkage, so the fit on training data worsens due to underfitting).

(ii) **Test MSE**: often *U-shaped*; at $\lambda = 0$ we might overfit, leading to high test error; at some middle λ it's minimized, then at very large λ it underfits.

(iii) **(Squared) Bias**: *Steadily increases* with λ (more shrinkage means systematically underestimating true effects).

(iv) **Variance**: *Steadily decreases* as λ grows (heavy shrinkage lowers the model's sensitivity to training noise, thereby lowering variance).

(v) **Irreducible Error**: *Remains constant* (it doesn't depend on the model or λ).

(b) (10 points) Table with two methods (Ridge vs. Lasso)

	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 is Lasso, which is Ridge?**

- *Method A* never zeros out β_2.
- *Method B* sets $\hat{\beta}_2 = 0$ for $\lambda \geq 1.0$.

Hence, **Method A** is *Ridge*, **Method B** is *Lasso*.

(ii) **Which λ to pick?**

- For Ridge (A), $\lambda = 1.0$ yields the lowest CV error (1.05 vs. 1.10 or 1.30).
- For Lasso (B), $\lambda = 0.1$ is yields minimal CV error (1.10 vs. 1.15, 1.35).
- If we want a simpler model (fewer nonzero coefficients), $\lambda \geq 1.0$ for Method B zeroes out β_2, though that raises CV error slightly (from 1.10 to 1.15 or 1.35). We might accept that if interpretability/simplicity of the model is important.

Problem 6: Solution (20 points + 5 bonus points)

(a) (10 points)

We have a *single* hypothesis H_0, with probabilities $\{p_1, p_2, p_3, p_4\}$ in the table (left). For *multiple* tests (e.g. m of them), let N_1, N_2, N_3, N_4 be the respective counts of outcomes in the right table.

(i) Often we set $\Pr(\text{reject } H_0 \mid H_0 \text{ true}) \leq \alpha$. In the single-test table, that means

$$\frac{p_1}{p_1 + p_3} \;\leq\; \alpha \quad\Longleftrightarrow\quad p_1 \;\leq\; \alpha\,(p_1 + p_3).$$

(ii) To control the false discovery rate (FDR) at level q, we want the expected fraction of false positives among all rejections $\leq q$. In the multiple-test table, if N_1 is the total FP among the "true" nulls and $N_1 + N_2$ is the number of rejections, then

$$\mathbb{E}\Big[\tfrac{N_1}{N_1+N_2}\Big] \;\leq\; q.$$

(b) (10 points + 5 bonus points)

You have *8* hypotheses (each at $\alpha = 0.05$) with p-values

$$\{\,0.001,\ 0.01,\ 0.02,\ 0.04,\ 0.06,\ 0.10,\ 0.15,\ 0.25\}.$$

(i) **No correction**: We reject all hypotheses whose p-values < 0.05. Hence 0.001, 0.01, 0.02, 0.04 are each below 0.05. So we reject 4 hypotheses: $H_{0,1}, H_{0,2}, H_{0,3}, H_{0,4}$.

(ii) **Bonferroni correction** ($m = 8$): The new threshold is $\alpha^* = \frac{0.05}{8} = 0.00625$. Only $0.001 < 0.00625$. So 1 rejection: $H_{0,1}$.

(iii) **Benjamini–Hochberg (BH) at FDR=10%**: Sort the p-values:

$$0.001,\ 0.01,\ 0.02,\ 0.04,\ 0.06,\ 0.10,\ 0.15,\ 0.25.$$

We look for the largest j such that

$$p_{(j)} \;\leq\; \frac{q\,j}{m} = \frac{0.10 \times j}{8}.$$

- $j = 1$: check $0.001 \leq 0.10 \times \frac{1}{8} = 0.0125$? yes.
- $j = 2$: check $0.01 \leq 0.10 \times \frac{2}{8} = 0.025$? yes.
- $j = 3$: check $0.02 \leq 0.10 \times \frac{3}{8} = 0.0375$? yes.
- $j = 4$: check $0.04 \leq 0.10 \times \frac{4}{8} = 0.05$? yes.
- $j = 5$: check $0.06 \leq 0.10 \times \frac{5}{8} = 0.0625$? yes.
- $j = 6$: check $0.10 \leq 0.10 \times \frac{6}{8} = 0.075$? no ($0.10 > 0.075$). stop.

Hence $j = 5$. Thus we reject $p_{(1)}, \ldots, p_{(5)}$:

$$\{\,0.001,\ 0.01,\ 0.02,\ 0.04,\ 0.06\}.$$

So 5 rejections: $H_{0,1}, H_{0,2}, H_{0,3}, H_{0,4}, H_{0,5}$.