STA 35C Statistical Data Science III

Midterm exam 2 solution

Instructor: Dogyoon Song

Problem 1

(a) **True.** In k-fold cross-validation, we fit k models. Thus, 10-fold CV requires fitting 10 models, while 5-fold CV requires fitting 5 models.

(b) **False.** A bootstrap sample is drawn with replacement, so some original observations may appear multiple times, while others may not appear at all.

(c) **False.** Forward stepwise selection starts from the null model and adds predictors one at a time. It does not remove predictors added in earlier steps.

(d) **False.** The roles are reversed: lasso can set some coefficients exactly equal to zero, while ridge regression generally shrinks coefficients toward zero without setting them exactly to zero.

(e) **False.** FDR control at $q = 0.05$ controls the expected fraction of false discoveries among rejected hypotheses. It does not guarantee, with probability 95%, that there are no false positives.

Problem 2

(a) One advantage of 5-fold cross-validation compared to a single train/validation split is its stability: it averages validation errors over several held-out folds rather than relying on one random split.

One disadvantage is that it is more computationally expensive, since it requires fitting 5 models instead of only one model from a single split.

(b) We compute the held-out prediction errors fold by fold.

1) Linear model.

When F_1 is the training fold, the fitted linear model is $\hat{y} = -\frac{1}{3} + 2x$. We evaluate this model on the held-out fold $F_2 = \{(0,0),(1,2),(2,6)\}$. The squared prediction residuals are

$$\left(0 - \left(-\frac{1}{3}\right)\right)^2 = \frac{1}{9}, \qquad \left(2 - \frac{5}{3}\right)^2 = \frac{1}{9}, \qquad \left(6 - \frac{11}{3}\right)^2 = \left(\frac{7}{3}\right)^2 = \frac{49}{9}.$$

Therefore, the validation MSE for this fold is

$$\mathrm{MSE}_{F_2} = \frac{1/9 + 1/9 + 49/9}{3} = \frac{17}{9}.$$

When F_2 is the training fold, the fitted linear model is $\hat{y} = -\frac{1}{3} + 3x$. We evaluate this model on the held-out fold $F_1 = \{(0,0),(1,1),(2,4)\}$. The squared prediction residuals are

$$\left(0 - \left(-\frac{1}{3}\right)\right)^2 = \frac{1}{9}, \qquad \left(1 - \frac{8}{3}\right)^2 = \left(-\frac{5}{3}\right)^2 = \frac{25}{9}, \qquad \left(4 - \frac{17}{3}\right)^2 = \left(-\frac{5}{3}\right)^2 = \frac{25}{9}.$$

Therefore, the validation MSE for this fold is

$$\mathrm{MSE}_{F_1} = \frac{1/9 + 25/9 + 25/9}{3} = \frac{17}{9}.$$

Hence the 2-fold CV estimate for the linear model is

$$\widehat{\mathrm{MSE}}_{\mathrm{CV,linear}} = \frac{1}{2}\left(\mathrm{MSE}_{F_1} + \mathrm{MSE}_{F_2}\right) = \frac{1}{2}\left(\frac{17}{9} + \frac{17}{9}\right) = \frac{17}{9} \approx 1.889.$$

2) Quadratic model.

When F_1 is the training fold, the fitted quadratic model is $\hat{y} = x^2$. We evaluate it on $F_2 = \{(0,0), (1,2), (2,6)\}$. The squared prediction residuals are

$$(0-0)^2 = 0, \qquad (2-1)^2 = 1, \qquad (6-4)^2 = 4.$$

Therefore,

$$\mathrm{MSE}_{F_2} = \frac{0+1+4}{3} = \frac{5}{3}.$$

When F_2 is the training fold, the fitted quadratic model is $\hat{y} = x^2 + x$. We evaluate it on $F_1 = \{(0,0), (1,1), (2,4)\}$. The squared errors are

$$(0-0)^2 = 0, \qquad (1-2)^2 = 1, \qquad (4-6)^2 = 4.$$

Therefore,

$$\mathrm{MSE}_{F_1} = \frac{0+1+4}{3} = \frac{5}{3}.$$

Hence the 2-fold CV estimate for the quadratic model is

$$\widehat{\mathrm{MSE}}_{\mathrm{CV,quad}} = \frac{1}{2}\left(\mathrm{MSE}_{F_1} + \mathrm{MSE}_{F_2}\right) = \frac{1}{2}\left(\frac{5}{3} + \frac{5}{3}\right) = \frac{5}{3} \approx 1.667.$$

(c) We choose the quadratic model for prediction because it has the smaller estimated test MSE:

$$\frac{5}{3} < \frac{17}{9}.$$

Cross-validation estimates out-of-sample prediction performance, so the model with smaller CV MSE is preferred for prediction.

Problem 3

(a) The original sample is $\{1, 2, 3, 4, 6\}$, so the original sample median is $\hat{\theta} = 3$.

The bootstrap samples are already sorted in the table. Their medians are:

$$2, \quad 2, \quad 3, \quad 4, \quad 6.$$

The mean of the bootstrap medians is

$$\bar{\theta}^* = \frac{2+2+3+4+6}{5} = \frac{17}{5} = 3.4.$$

The bootstrap standard error is the sample standard deviation of these five medians:

$$\begin{aligned}\widehat{\mathrm{SE}}_{\mathrm{boot}} &= \sqrt{\frac{(2-3.4)^2+(2-3.4)^2+(3-3.4)^2+(4-3.4)^2+(6-3.4)^2}{5-1}} \\ &= \sqrt{\frac{1.96+1.96+0.16+0.36+6.76}{4}} \\ &= \sqrt{2.8} \approx 1.673.\end{aligned}$$

A normal-approximation 95% confidence interval is

$$\hat{\theta} \pm 1.96\,\widehat{\mathrm{SE}}_{\mathrm{boot}} = 3 \pm 1.96 \times 1.673 \approx 3 \pm 3.279.$$

Thus, the resulting confidence interval is approximately

$$[-0.279,\ 6.279].$$

(b) The population median is fixed but unknown. The confidence interval is random because it depends on the random sample.

The statement "95% confidence" means that if we repeatedly collected new samples and constructed intervals using the same procedure, then approximately 95% of those intervals would contain the true population median.

It does not mean that, after this particular interval has been computed, the probability that the fixed population median lies in this particular interval is 95%.

Problem 4

(a) For best subset selection, we choose the lowest RSS among all subsets of each size.

The best subsets by model size are:

p	**Best subset**	**RSS**	R^2_{adj}
0	$\emptyset$	100.0	0
1	$\{X_1\}$	50.0	$1 - \frac{50/(10-1-1)}{100/(10-1)} \approx 0.438$
2	$\{X_2, X_3\}$	38.0	$1 - \frac{38/(10-2-1)}{100/(10-1)} \approx 0.511$
3	$\{X_1, X_3, X_4\}$	36.0	$1 - \frac{36/(10-3-1)}{100/(10-1)} \approx 0.460$
4	$\{X_1, X_2, X_3, X_4\}$	35.5	$1 - \frac{35.5/(10-4-1)}{100/(10-1)} \approx 0.361$

The largest adjusted R^2 among these is approximately 0.511, achieved by

$$\{X_2, X_3\}.$$

Thus, best subset selection with adjusted R^2 chooses the model with predictors X_2 and X_3.

(b) Forward stepwise selection starts with the null model:

$$\mathcal{M}_0 = \emptyset.$$

Step 1. Among one-predictor models, the smallest RSS is for X_1, so

$$\mathcal{M}_1 = \{X_1\}.$$

Step 2. Add one predictor to $\{X_1\}$. The candidates are:

$$\{X_1, X_2\} : 48.0, \qquad \{X_1, X_3\} : 46.0, \qquad \{X_1, X_4\} : 49.0.$$

The best is

$$\mathcal{M}_2 = \{X_1, X_3\}.$$

Step 3. Add one predictor to $\{X_1, X_3\}$. The candidates are:

$$\{X_1, X_2, X_3\} : 36.5, \qquad \{X_1, X_3, X_4\} : 36.0.$$

The best is

$$\mathcal{M}_3 = \{X_1, X_3, X_4\}.$$

Step 4. Add the remaining predictor:

$$\mathcal{M}_4 = \{X_1, X_2, X_3, X_4\}.$$

Now compare adjusted R^2 along the forward stepwise path:

p	**Forward stepwise model**	**RSS**	R^2_{adj}
0	$\emptyset$	100.0	0
1	$\{X_1\}$	50.0	0.438
2	$\{X_1, X_3\}$	46.0	$1 - \frac{46/(10-2-1)}{100/(10-1)} \approx 0.409$
3	$\{X_1, X_3, X_4\}$	36.0	$1 - \frac{36/(10-3-1)}{100/(10-1)} \approx 0.460$
4	$\{X_1, X_2, X_3, X_4\}$	35.5	$1 - \frac{35.5/(10-4-1)}{100/(10-1)} \approx 0.361$

The largest adjusted R^2 along this path is approximately 0.460, achieved by

$$\{X_1, X_3, X_4\}.$$

This differs from best subset selection, which selected $\{X_2, X_3\}$.

(c) This is more likely **lasso**, because some coefficients are exactly zero for $\lambda = 1.0$ and $\lambda = 10.0$. Ridge generally shrinks coefficients toward zero but does not typically set coefficients exactly equal to zero.

If prediction accuracy is the only goal, choose the λ with the smallest CV MSE:

$$\lambda = 0.1, \qquad \text{CV MSE} = 20.$$

If interpretability is more important, then $\lambda = 1.0$ may be preferred despite its larger CV MSE of 23, because it sets $\hat{\beta}_2 = 0$ and $\hat{\beta}_4 = 0$, producing a simpler sparse model.

Thus, the choice between $\lambda = 0.1$ and $\lambda = 1.0$ depends on whether prediction accuracy or sparsity/interpretability is prioritized.

(d) As λ increases from 0 to ∞, the expected qualitative behavior is:

Quantity	**Behavior**	**Option**
Training error	Generally increases	(2)
Expected test error	Decreases initially, then eventually increases	(4)
Squared bias	Generally increases	(2)
Variance	Generally decreases	(3)
Irreducible error	Remains constant	(1)

(e*) Ridge and lasso can behave differently when predictors are highly correlated.

Ridge tends to keep both X_1 and X_2 in the model and shrink their coefficients together. This can be preferable for prediction because it can give more stable predictions when correlated predictors carry similar information.

Lasso may select one of X_1 or X_2 and set the other coefficient to zero. This can be preferable for interpretation because it produces a simpler model, but the particular selected predictor may be unstable when the predictors are nearly identical.

Problem 5

(a)(i) Controlling the familywise error rate (FWER) at level α means

$$\Pr(N_1 \geq 1) \leq \alpha.$$

(ii) The false discovery proportion is

$$\text{FDP} = \frac{N_1}{N_1 + N_2}.$$

Controlling the false discovery rate (FDR) at level q means

$$\text{FDR} = \mathbb{E}[\text{FDP}] = \mathbb{E}\left[\frac{N_1}{N_1 + N_2}\right] \leq q.$$

(b) We test $m = 8$ hypotheses with p-values:

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

(i) With no correction at significance level $\alpha = 0.05$, reject all hypotheses with $p_j < 0.05$. Thus, reject

$$H_{0,1},\ H_{0,2},\ H_{0,3},\ H_{0,4},\ H_{0,5}.$$

(ii) With the Bonferroni correction, the threshold is

$$\frac{\alpha}{m} = \frac{0.05}{8} = 0.00625.$$

Reject hypotheses with $p_j < 0.00625$. Thus, reject

$$H_{0,1},\ H_{0,2}.$$

(iii) For the Benjamini–Hochberg procedure at $q = 0.05$, compute thresholds

$$\frac{qj}{m} = \frac{0.05j}{8}, \qquad j = 1, \ldots, 8.$$

These thresholds are

$$0.00625,\ 0.0125,\ 0.01875,\ 0.025,\ 0.03125,\ 0.0375,\ 0.04375,\ 0.050.$$

Compare:

j	1	2	3	4	5	6	7	8
$p_{(j)}$	0.001	0.004	0.020	0.024	0.045	0.080	0.125	0.200
qj/m	0.00625	0.0125	0.01875	0.025	0.03125	0.0375	0.04375	0.050

Although $p_{(3)} = 0.020$ is larger than its own threshold 0.01875, the largest j satisfying

$$p_{(j)} \leq \frac{qj}{m}$$

is $j = 4$, since $p_{(4)} = 0.024 \leq 0.025$. Therefore, the BH procedure rejects

$$H_{0,1},\ H_{0,2},\ H_{0,3},\ H_{0,4}.$$

(c*) No. If Benjamini–Hochberg at $q = 0.10$ rejects 40 hypotheses, it does not guarantee that at most 4 of the 40 rejected hypotheses are false discoveries.

FDR control is an average guarantee over repeated applications of the procedure:

$$\mathbb{E}[\text{FDP}] \leq 0.10.$$

The realized false discovery proportion in one particular dataset may be larger than 0.10. Thus, more than 4 of the 40 rejected hypotheses could be false discoveries in a particular study.