STA 35C Statistical Data Science III

Midterm 2 practice B solutions

Instructor: Dogyoon Song

Problem 1. Basic concepts

(a) No. Since $b_1X_1 + b_2X_2 = (b_1 + 2b_2)X_1$, adding X_2 leaves the set of attainable fitted values unchanged; the minimum training RSS is unchanged.

(b) A and C. CV holds each evaluation observation out of its model fit, and refitting on all development data after tuning does not use the untouched test set. B is false because the validation responses influence the model choice; D is false because neither CV scheme must give the smaller estimate.

(c) 1/2. The mean is 3 for the ordered resamples $(1, 5)$ and $(5, 1)$, each with probability $1/4$. A bootstrap mean can differ from every original observation even though resampled observations cannot.

(d) Decrease. With positive RSS fixed, dividing it by fewer residual degrees of freedom increases the numerator in the adjusted-R^2 formula; $\mathrm{TSS}/(n-1)$ is unchanged.

(e) B and D. Centering alone does not remove scale differences, but dividing by each predictor's own standard deviation does. CV can select the penalty; a nonzero lasso coefficient is not a significance test.

(f) C and D. FDR is an expectation, not a per-dataset or per-rejection guarantee. Writing R for the number of rejections and V for the number of false rejections,

$$0 \le \frac{V}{\max\{R, 1\}} \le \mathbf{1}_{\{V \ge 1\}} \quad \Longrightarrow \quad \mathrm{FDR} \le \mathrm{FWER}\,.$$

Problem 2. Prediction assessment and uncertainty

(a) Expected squared prediction errors. For each method, add squared bias, prediction variance, and independent new-noise variance:

$$\mathrm{MSE}_A = (5-5)^2 + 4 + 2 = 6,$$
$$\mathrm{MSE}_B = (4-5)^2 + 1 + 2 = 4.$$

Choose **Method B**, since $4 < 6$: its variance advantage outweighs its nonzero squared bias.

(b) LOOCV error. The held-out residuals are $1, 0, -2, 1$, so

$$\mathrm{CV}_{\mathrm{LOO}} = \frac{(3-2)^2 + (5-5)^2 + (4-6)^2 + (8-7)^2}{4} = \frac{6}{4} = \mathbf{1.5}.$$

No: the estimate is deterministic conditional on the observed dataset and method, but it can vary when a new dataset is collected.

(c) Bootstrap confidence interval. Using the supplied normal-approximation rule,

$$12 \pm 1.96(1.5) = 12 \pm 2.94 = \mathbf{[9.06,\ 14.94]} \text{ minutes.}$$

Using this approximation, we are 95% confident that the **population mean delivery time** is between 9.06 and 14.94 minutes, not that 95% of individual deliveries fall in that range.

Problem 3. Model selection and shrinkage

(a) Training fit versus bias and prediction error. The claim is not justified: training MSE measures fit to the observed training responses, whereas squared bias compares the average prediction across repeated training samples with the true regression function. The degree-8 model can reproduce every linear fit by setting its higher-order coefficients to zero, so its minimum training MSE cannot be larger; however, its additional flexibility can increase prediction variance by fitting training noise. If that variance increase outweighs any reduction in squared bias, expected test MSE increases; the displayed results are consistent with this explanation but do not determine the separate bias and variance components.

(b) Backward selection versus the best pair. At the first step, compare the four three-predictor models:

Predictors	X_1, X_2, X_3	X_1, X_2, X_4	X_1, X_3, X_4	X_2, X_3, X_4
RSS	26	12	11	16

The smallest RSS is 11, so remove X_2 and retain $\{X_1, X_3, X_4\}$. The next step compares only its three two-predictor subsets:

$$\mathrm{RSS}(X_1, X_3) = 38, \qquad \mathrm{RSS}(X_1, X_4) = 20, \qquad \mathrm{RSS}(X_3, X_4) = 43.$$

Remove X_3, giving the path

$$\{X_1, X_2, X_3, X_4\} \longrightarrow \{X_1, X_3, X_4\} \longrightarrow \boldsymbol{\{X_1, X_4\}}.$$

No: the best two-predictor subset is $\{X_2, X_4\}$, with RSS $18 < 20$, but backward selection cannot reintroduce the previously removed predictor X_2.

(c) Lasso and the prediction–simplicity tradeoff. The pattern is characteristic of **lasso**, which can set fitted slopes exactly to zero as the penalty increases. If minimizing the observed CV MSE is the only criterion, $\lambda = 6$ **is preferred**, with CV MSE 17.0 rather than 17.4. However, $\lambda = 30$ **retains only** X_1 **and** X_3 **instead of all four predictors**, so it may be preferred when the simpler model or lower predictor-collection cost is worth the increase of 0.4 in observed CV MSE.

Problem 4. Reviewing a customer-renewal analysis

(a) Pooled classification errors. Every held-out customer receives equal weight:

$$\widehat{R}_A = \frac{1 + 3 + 15}{20 + 30 + 50} = \mathbf{0.19}, \qquad \widehat{R}_B = \frac{3 + 5 + 8}{20 + 30 + 50} = \mathbf{0.16}.$$

Choose **Method B**, whose observed CV misclassification rate is 16%, compared with 19% for A.

(b) Benjamini–Hochberg. With $m = 5$ and $q = 0.10$, compare $p_{(i)}$ with $iq/m = 0.02i$ for *every* rank:

Rank i	$p_{(i)}$	iq/m	$p_{(i)} \le iq/m$?
1	0.010	0.020	Yes
2	0.045	0.040	No
3	0.055	0.060	Yes
4	0.090	0.080	No
5	0.300	0.100	No

The largest passing rank is $k = 3$, so reject all hypotheses of ranks 1 through 3:

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

Rank 2 is included even though its own comparison fails; BH does not stop at the first failure.

(c) The claim is false. Increasing the number of bootstrap replicates reduces Monte Carlo variability in the approximation, but does not reduce the original error-rate estimator's sampling uncertainty across customer samples.

Problem 5. Optional bonus questions

(a) An alternative lasso minimizer is $(5, 1, 1)$**.** Because $x_{i1} = x_{i2}$, the fitted training values agree:

$$5 + 2x_{i1} + 0x_{i2} = 5 + x_{i1} + x_{i2}.$$

The penalties also agree:

$$\lambda(|2| + |0|) = 2\lambda = \lambda(|1| + |1|).$$

Thus $(5, 1, 1)$ has the same objective value as the given minimizer and is also a minimizer, despite retaining both predictors. **Lasso need not determine a unique selected predictor set when predictors are identical.**

(b) FDR equals FWER when every null is true. In that case $V = R$, and

$$\frac{V}{\max\{R, 1\}} = \begin{cases} 0, & R = 0, \\ 1, & R \geq 1. \end{cases}$$

Consequently,

$$\mathrm{FDR} = \mathbb{E}[\mathbf{1}_{\{R \geq 1\}}] = P(R \geq 1) = P(V \geq 1) = \mathrm{FWER}\,.$$

No independence assumption is needed for this identity.