STA 35C Statistical Data Science III

Midterm exam 2 solutions

Instructor: Dogyoon Song

Problem 1. Basic concepts

(a) A and B. New-response noise gives the irreducible component, and prediction variance measures sensitivity to the training sample.

(b) A and C. Resample rows independently with replacement while keeping each predictor–response pair intact.

(c) A and D. With V false rejections and R total rejections,

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

FDR is an expectation; controlling it does not guarantee the same bound for each dataset's false-discovery proportion or for FWER.

(d) The prediction at $x = 2$ approaches **18**: the slope approaches zero, while the unpenalized intercept approaches the training-response mean $\overline{y} = 18$.

(e) No. A nonzero lasso coefficient indicates selection by the penalized fitting procedure, not rejection of a null hypothesis at the 5% significance level.

(f) False. LOOCV is deterministic for fixed data and a deterministic fitting method, but its error estimate can vary across independently collected datasets.

Problem 2. Regularization and cross-validation

(a) Qualitative ridge trends: $(2, 4, 2, 3, 1)$.

(i) **(2)** Training MSE generally increases.
(ii) **(4)** Expected test MSE decreases, then increases.
(iii) **(2)** Squared bias generally increases.
(iv) **(3)** Prediction variance generally decreases.
(v) **(1)** Irreducible error remains constant.

These are the requested qualitative lecture patterns; in particular, a U-shaped test-error curve is not a universal guarantee.

(b) Two-fold CV errors. Evaluate fits trained on F_1 using F_2, and vice versa. The held-out residuals $y - \widehat{y}$ are

Training → validation	Linear residuals	Quadratic residuals
$F_1 \to F_2$	$(1, 1, 7)$	$(0, 3, 6)$
$F_2 \to F_1$	$(1, -5, -5)$	$(0, -3, -6)$

Pooling all six held-out observations gives

$$\mathrm{CV}_{\text{linear}} = \frac{1+1+49+1+25+25}{6} = \mathbf{17},$$
$$\mathrm{CV}_{\text{quadratic}} = \frac{0+9+36+0+9+36}{6} = \mathbf{15}.$$

Problem 3. Bootstrap uncertainty

(a) Bootstrap standard error. The five bootstrap medians are $6, 6, 8, 10, 10$, with mean $\overline{m}^* = 8$. Their sample standard deviation is

$$\widehat{\mathrm{SE}}_{\text{boot}} = \sqrt{\frac{(6-8)^2+(6-8)^2+(8-8)^2+(10-8)^2+(10-8)^2}{5-1}} = \sqrt{\frac{16}{4}} = \mathbf{2} \text{ minutes.}$$

(b) Confidence interval and confidence level. The original sample median is $\widehat{m} = 9$, so the supplied rule gives

$$9 \pm 1.96(2) = 9 \pm 3.92 = \mathbf{[5.08,\ 12.92]} \text{ minutes.}$$

With a valid normal approximation, about 95% of intervals constructed from repeated independent samples would contain the fixed population median delivery time.

Problem 4. Subset selection

(a) Best subset and forward stepwise selection. Here TSS = 100 and $n = 10$, so maximize adjusted R^2 by minimizing $C_k = \mathrm{RSS}/(9-k)$. Forward selection first adds X_1 because $40 < 46 < 76$, then X_2 because $\mathrm{RSS}(X_1, X_2) = 30 < 40 = \mathrm{RSS}(X_1, X_3)$. The candidates and criterion values are:

k	Best subset	C_k	Forward subset	C_k
0	$\varnothing$	100/9	$\varnothing$	100/9
1	$\{X_1\}$	5	$\{X_1\}$	5
2	$\{X_2, X_3\}$	22/7	$\{X_1, X_2\}$	30/7
3	$\{X_1, X_2, X_3\}$	22/6	$\{X_1, X_2, X_3\}$	22/6

Best subset selection chooses $\{X_2, X_3\}$, with $C_2 = 22/7$ and adjusted $R^2 = 251/350 \approx 0.717$. **Forward selection chooses** $\{X_1, X_2, X_3\}$, with $C_3 = 22/6$ and adjusted $R^2 = 0.670$.

(b) Evaluate the entire selection-and-fitting procedure. For each CV split, use best subset selection to choose the two-predictor model with the smallest RSS on the training observations, and fit OLS on those same observations. Evaluate predictions on the held-out fold, repeating selection and fitting for every split.

Problem 5. From predictive models to conclusions

(a) Lasso; choose $\lambda = 15$. Exact zeros in the fitted slopes as the penalty increases are characteristic of lasso's variable selection. The setting $\lambda = 3$ retains five predictors and is ineligible; among the eligible settings, $\lambda = 15$ retains X_2, X_4 and has lower observed CV MSE than $\lambda = 75$ ($2.41 < 10.57$).

(b) The claim is false. The original regularized coefficients are feasible in the OLS refit, so OLS cannot increase training RSS. Removing shrinkage can increase sensitivity to the training sample, so improved training fit does not guarantee lower expected error on new observations.

(c) Bonferroni and Holm. Bonferroni: the cutoff is $0.05/8 = 0.00625$, so reject $\boldsymbol{\{H_{0,1}, H_{0,2}\}}$.

Holm: compare ordered p-values with $0.05/(8 - i + 1)$, stopping at the first failure:

Rank i	$p_{(i)}$	Holm cutoff	Decision
1	0.003	$0.05/8 = 0.00625$	Pass
2	0.006	$0.05/7 \approx 0.007143$	Pass
3	0.008	$0.05/6 \approx 0.008333$	Pass
4	0.012	$0.05/5 = 0.010$	Fail; stop
5–8			Not reached

Thus Holm rejects $\boldsymbol{\{H_{0,1}, H_{0,2}, H_{0,3}\}}$. Passing comparisons at later ranks do not restart Holm after a failure.

Problem 6. Optional bonus questions

(a) The actual coverage probability is zero. Every bootstrap observation is at most the original sample maximum M, so every bootstrap maximum, and hence the interval's upper endpoint U^*, is at most M. For a finite sample from the continuous uniform distribution on $(0, \theta)$,

$$U^* \leq M < \theta \quad \text{with probability 1.}$$

The proposed interval therefore cannot contain θ, so $\boldsymbol{P\{\theta \text{ is in the interval}\} = 0}$.

(b) No; the two minimum objective values are equal. For any duplicated-predictor fit, set $b_0 = a_0$ and $b = a_1 + a_2$; this preserves fitted values, while the triangle inequality gives

$$Q(a_0, a_1 + a_2) \leq \widetilde{Q}(a_0, a_1, a_2), \qquad |a_1 + a_2| \leq |a_1| + |a_2|.$$

Thus the original minimum is no larger than the duplicated-model minimum. Conversely, every original fit is reproduced with exactly the same objective by $(a_0, a_1, a_2) = (b_0, b, 0)$, proving equality of the minima.