STA 35C: Statistical Data Science III

Lecture 19: Review for Midterm 2

Dogyoon Song

Fall 2026, UC Davis

Announcements

Midterm 2: Mon, Nov 9, 9:00–9:50 AM (in class)

- Arrive early: The exam ends at 9:50 AM sharp.
- One hand-written, letter-sized (8.5×11 inches), double-sided sheet of formulas/notes.
- A simple non-graphing scientific calculator is allowed; no other reference materials or electronic devices.
- **SDC accommodations:** Confirm scheduling with AES online.

Preparation and support

- Primary coverage: Lectures 12–19.
- Core concepts from earlier may be assumed, especially regression and model assessment.
- Check Canvas and the course webpage for practice materials.

After midterm 2

- No class on Wed, Nov 11 (Veterans Day)
- The lecture on Fri, Nov 13 will also be held on Zoom; see Canvas for the link

Agenda: Review for Midterm 2

Problem map and study priorities:

- **Problem 1: Concept checks** Review core definitions, assumptions, and interpretations across the covered topics.
- **Problem 2: Bias–variance and cross-validation** Relate regularization to bias, variance, and prediction error; evaluate models using CV.
- **Problem 3: Bootstrap and uncertainty** Understand bootstrap resampling; compute standard errors and construct and interpret confidence intervals.
- **Problem 4: Subset selection** Apply subset-selection methods and distinguish model fitting, selection, and evaluation.
- **Problem 5: Regularization and multiple testing** Interpret regularized fits, compare candidate models, and apply multiple-testing corrections.

Midterm 2 review: What to know cold

Core skills:

- **Estimate test performance:** distinguish training, validation, CV, and test error.
- **Quantify uncertainty:** describe bootstrap sampling and compute bootstrap SEs/CIs from replicates.
- **Select models:** compare subset selection, stepwise selection, ridge, and lasso.
- **Tune complexity:** explain how λ affects bias, variance, training error, and test error.
- **Control false positives:** distinguish FWER vs. FDR and apply Bonferroni, Holm, and Benjamini–Hochberg.

Advice: Consider what the method estimates or controls before doing the calculation.

Review: Cross-validation

Goal: Estimate test performance from training data alone

Key ideas:

- Single split (validation set): random partition into training/validation sets; simple but high variance
- LOOCV (leave-one-out): train on $n-1$ points, validate on 1 point, repeat for all points
- k-fold CV: partition data into k folds, systematically rotate which fold is the validation set

Trade-offs:

- Single split is simple, but can vary substantially with the random split.
- LOOCV uses $n-1$ observations for training each time, but requires n model fits.
- k-fold CV, often $k = 5$ or 10, is a practical compromise.

Usage:

- CV MSE: average held-out squared errors over observations to approximate test MSE
- Model selection: pick model that yields lowest CV error
- Tuning parameters (e.g. λ in ridge/lasso)

Review: Bootstrap

Goal: Approximate the sampling distribution (e.g. standard errors) using just one dataset

Method:

- Sample n points *with replacement* from the original dataset of size n (a "bootstrap sample")
- Compute the statistic of interest (mean, regression coefficient, etc.) on each bootstrap sample
- Repeat B times, forming a distribution of the statistic estimates $\{\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*\}$

Bootstrap SE/CI:

- Bootstrap standard error:

$$\widehat{\mathrm{SE}}_{\mathrm{boot}} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_b^* - \bar{\theta}^*\right)^2}, \qquad \text{where} \qquad \bar{\theta}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*.$$

- Construct a confidence interval using the specified rule: bootstrap percentiles or a normal approximation
- A 95% CI has a repeated-sampling interpretation: over many datasets, about 95% of intervals constructed this way would contain the true parameter

Key premise: The observed sample is representative of the population

Review: Subset selection

Goal: Identify a relevant subset of predictors among many

Best subset selection:

- Fits all 2^p subsets (exhaustive); picks the best model for each size k, then chooses among them by using adjusted R^2, CV, etc.
- Feasible only if p is small or moderate (can be very expensive for large p)

Forward/backward stepwise:

- Greedy approximations: add/remove one predictor at a time
- Number of model fits: $\mathcal{O}(p^2)$ vs. 2^p for best subset
- Might miss the absolute best subset but often works well in practice

Pros/Cons:

- Direct variable selection: excluded predictors are treated as having coefficient 0
- Can be unstable for large p; small data changes may change the selected subset substantially

Review: Regularization

Motivation: Least squares can have high variance or nonunique coefficients

Ridge regression:

- Add penalty $\lambda \sum_{j=1}^{p} \beta_j^2$; typically do not penalize the intercept
- Typically shrinks coefficients toward 0, but generally does not set them exactly to 0
- More stable under collinearity

Lasso:

- Add penalty $\lambda \sum_j |\beta_j|$; typically do not penalize the intercept
- Can zero out some coefficients, enabling variable selection
- Can be less stable than ridge especially when choosing among highly correlated predictors

Tuning λ: Usually chosen by cross-validation; neither ridge nor lasso always wins

- **As λ increases:** on the same data and scaling, training RSS cannot decrease; test error need not be monotone
- Shrinkage often trades increased bias for reduced variance

Review: Multiple hypotheses testing

Problem: Testing many hypotheses inflates chance of false positives

- If all m nulls are true and independent tests each reject with probability exactly α,

$$P(\geq 1 \text{ false positive}) = 1 - (1 - \alpha)^m.$$

- p-hacking/data dredging: repeatedly searching for small p-values can lead to spurious "discoveries"

FWER (Family-Wise Error Rate):

- Probability of any (=at least 1) false positive
- With valid p-values, Bonferroni/Holm control FWER $\leq \alpha$
- Often conservative, can reduce power when m is large

FDR (False Discovery Rate):

-
- Expected false-positive proportion among total rejections (=FP + TP): $\mathrm{FDR} = \mathbb{E}\left[\frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TP}}\right]$.
- Benjamini–Hochberg procedure can control FDR
- Less conservative, typically yields more rejections, tolerating some false positives

Midterm 2 Checkpoint: Cross-validation vs. bootstrap

Suppose you have one dataset and two goals:

- Goal 1: Choose between a linear and a quadratic model for prediction.
- Goal 2: Estimate the standard error of a sample median.

Question: Which pairing is most appropriate?

A) Goal 1: cross-validation; Goal 2: bootstrap.
B) Goal 1: bootstrap; Goal 2: cross-validation.
C) Both goals should be handled by training error.
D) Neither goal can be addressed using resampling.

Answer: A. Cross-validation estimates predictive performance; bootstrap approximates a sampling distribution for uncertainty.

Midterm 2 Checkpoint: Best subset vs. forward stepwise

Suppose X_1 is the best one-predictor model, but $\{X_2, X_3\}$ is the best two-predictor model, obtained via best subset selection method.

Question: What can happen under forward stepwise selection?

A) It must choose $\{X_2, X_3\}$ as the two-predictor model.

B) It cannot choose $\{X_2, X_3\}$ at the two-predictor step if it already chose X_1 first.

C) It checks all 2^p subsets, so it always matches best subset selection.

D) It always chooses the model with the largest training RSS.

Answer: B. Forward stepwise is greedy: once it chooses X_1, its two-predictor candidates must include X_1.

Midterm 2 Checkpoint: Model selection and regularization

Suppose two methods produce the following coefficient patterns:

Method	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
A	2.1	0.8	0.3	0.1
B	2.0	0	0.7	0

Question: Which statement is most reasonable?

A) Method A is more likely lasso because it keeps all coefficients nonzero.

B) Method B is more likely ridge because ridge performs variable selection.

C) Method B is more likely lasso because it sets some coefficients exactly to zero.

D) Neither method can be regularized because some coefficients are nonzero.

Answer: C. Exact zeros are characteristic of lasso-style variable selection, but the coefficient pattern alone does not identify the fitting method.

Midterm 2 Checkpoint: Multiple testing

Suppose all $m = 20$ null hypotheses are true and the tests are independent, each with false-positive probability exactly 0.05.

Question: Which statement is most accurate?

A) The probability of at least one false positive is exactly 0.05.

B) The probability of at least one false positive decreases as m increases.

C) Multiple testing is not a concern if each individual test uses $\alpha = 0.05$.

D) The expected number of false positives is $20 \cdot 0.05 = 1$.

Answer: D. If all nulls are true, each test has false-positive probability 0.05, so the expected number of false positives is $m\alpha = 1$.

Wrap-up: How to prepare

Use the practice midterms for review

- First attempt them without notes and under time pressure.
- Identify weak steps in your problem solving process: method choice, setup, calculation, or interpretation.
- Then review and prepare your cheat sheet around those weak points.

When preparing your cheat sheet

- Consider including definitions and assumptions, not only formulas.
- Ensure you are fluent with the tools you have written.

During the exam

- First identify the task: estimate test error, quantify uncertainty, select predictors, etc.
- Write the relevant criterion before calculating: CV error, bootstrap SE, adjusted R^2, training/CV MSE, FWER, or FDR.
- Keep notation clear: training vs. validation/test, bootstrap replicate vs. original statistic, FP vs. TP.

Advice: Identify the right tool first, then compute carefully and interpret in words.