STA 35C: Statistical Data Science III

Lecture 19: Review for Midterm 2

Dogyoon Song

Spring 2025, UC Davis

Announcement

Midterm 2 on Fri, May 15 (1:10 pm–2:00 pm in class)

- **Arrive early**: The exam starts at 1:10 pm and ends at 2:00 pm sharp
- **One hand-written cheat sheet:** Letter-size (8.5"×11"), double-sided, brief formulas/notes
- **Calculator**: A simple (non-graphing) scientific calculator is allowed
- **No other materials** beyond the single cheat sheet (no textbooks, etc.)
- **SDC accommodations:** Confirm scheduling with AES online ASAP

Office hour today (Wed, May 13) 3:30–4:30 PM

Agenda: Midterm 2 review

Goal: Review the major tools and the decisions they support.

- **Resampling:** cross-validation and bootstrap
- **Model selection:** subset selection and regularization
- **Multiple testing:** FWER, FDR, and adjusted decision rules

Guiding question: Given a problem, can you identify the right tool and explain what it estimates or controls?

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.

Main exam advice: state 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 train/test; 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:

- 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 desired statistic (mean, regression coefficient, etc.) on the 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:

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

- Use percentiles or normal approximation to construct confidence intervals
- Interpreting the coverage of confidence intervals requires care

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:

- Tries all 2^p subsets (exhaustive); picks the best model for each size k, then chooses among them by 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
- Complexity $\mathcal{O}(p^2)$ vs. 2^p for best subset
- Might miss the absolute best subset but often works well in practice

Pros/Cons:

- Direct variable selection (some coefficients set to zero)
- Can be unstable for large p; small data changes may change the selected subset substantially

Review: Regularization

Motivation: Least squares can have high variance, especially when predictors are highly correlated, $p \approx n$, or $p > n$

Ridge regression:

- Add penalty $\lambda \sum_j \beta_j^2$
- Typically shrinks coefficients toward 0, but generally does not set them exactly to 0
- More stable under collinearity

Lasso:

- Add penalty $\lambda \sum_j |\beta_j|$
- 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—depends on data and interpretability needs

Review: Multiple hypotheses testing

Problem: Testing many hypotheses inflates chance of false positives

- If all m nulls are true and tests are independent,

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

- 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
- Bonferroni, Holm's step-down keep FWER $\leq \alpha$
- Often conservative, can reduce power when m is large

FDR (False Discovery Rate):

- Expected fraction of false positives among 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

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

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

Pop-up quiz: 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 lasso because it sets some coefficients exactly to zero.

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

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

Answer: B. Exact zeros are characteristic of lasso and correspond to variable selection.

Pop-up quiz: Multiple testing

Suppose we test $m = 20$ independent null hypotheses at level $\alpha = 0.05$, and suppose all null hypotheses are actually true.

Question: Which statement is most accurate?

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

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

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

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

Answer: B. 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.
- Mark which steps failed: identifying the tool, setting up the model, or doing the calculation.
- 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, tune λ, or control false positives.
- State 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 vs. original statistic, FP vs. TP.

Advice: Be fluent in matching problems to the right tools and explaining what they do.