STA 35C – Homework 6

Submission due: Thu, November 19 at 11:59 PM PT

Instructor: Dogyoon Song

Instructions: Upload a **single PDF file** to Gradescope via Canvas ("Homework 6" under "Assignments"). Name it using your UC Davis email ID prefix and homework number, e.g., `dgsong_hw6.pdf`. Include "STA 35C," your name, and the last four digits of your student ID on the front page. For coding problems, include your R code, relevant output, and requested plots/tables in the PDF. Late submissions receive 0 points. See the syllabus for full submission and late-policy details.

Problem 1: Midterm 2 review (25 points total).

(a) Bootstrap (10 points). You observe seven independent draws from an integer-valued population with unknown median m:

$$2, \quad 0, \quad 0, \quad 1, \quad 2, \quad 3, \quad 2.$$

Let $\hat{m}$ be the sample median. Each column below is a bootstrap sample of size seven drawn with replacement from these observations.

Draw	Bootstrap 1	Bootstrap 2	Bootstrap 3	Bootstrap 4	Bootstrap 5
1	2	0	2	3	0
2	2	1	2	0	1
3	0	0	2	2	0
4	0	0	3	2	3
5	2	2	2	2	1
6	3	2	1	1	2
7	3	2	3	0	2

(i) **(3 points)** What is the probability that a bootstrap sample reproduces the sequence of values $(2, 0, 0, 1, 2, 3, 2)$ in exactly this order?

(ii) **(2 points)** Compute $\hat{m}$ and the median of each bootstrap sample.

(iii) **(2 points)** Estimate the bootstrap standard error using the sample standard deviation of the five bootstrap medians. Calculate the nominal 95% normal-approximation interval $\hat{m} \pm 1.96\, \widehat{\mathrm{SE}}_{\mathrm{boot}}$.

(iv) **(3 points)** What repeated-sampling probability is intended to be approximately 95% for a confidence-interval procedure?

(b) Cross-validation (5 points). Two models, A and B, are compared by three-fold cross-validation on ten observations. Each model is fitted on the other two folds before evaluating the held-out fold. The held-out sums of squared errors (SSE) are:

Fold	Size	Model A SSE	Model B SSE
1	2	1	4
2	3	3	6
3	5	15	5

(i) **(3 points)** Compute each model's overall CV MSE, weighting every held-out observation equally. Which model would you select?

(ii) **(2 points)** Would fitting each model once to all ten observations and then computing its errors on the three folds give a valid CV estimate? Explain briefly.

(c) Regularization (5 points). Recall the lasso objective

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|.$$

(i) On a fixed training dataset, can the training RSS of the lasso fit decrease as λ increases? Explain briefly.

(ii) Explain why test MSE need not change monotonically with λ.

(d) Multiple testing (5 points). For M simultaneous tests, let the random counts N_1, N_2, N_3, N_4 record:

	Null is true	Null is false
Reject	N_1	N_2
Do not reject	N_3	N_4

(i) Write the false discovery proportion (FDP), including the case of no rejections.

(ii) Express FWER control at level 0.05 and FDR control at level 0.10 as inequalities involving these counts.

Problem 2: Basis functions and spline basics (25 points total + 5 bonus points).

Work by hand. Write $(u)_+ = \max\{u, 0\}$.

(a) Basis functions (6 points). Consider

$$f(x) = \beta_0 + \beta_1 x + \beta_2 (x-2)_+, \qquad (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) = (1, 2, -3).$$

(i) For $x = 0, 1, 2, 3, 4$, make a table with columns x, 1, $(x-2)_+$, and $\hat{f}(x)$.

(ii) Explain why the model is linear in its coefficients even though $\hat{f}(x)$ need not be a straight line.

(b) A linear spline (6 points). Use the fitted function from part **(a)**.

(i) Write $\hat{f}(x)$ separately for $x \le 2$ and $x > 2$.

(ii) Sketch $\hat{f}$ on $[0, 4]$. Label the knot and the slope on each side.

(c) A cubic spline (7 points). [JWHT21, Chapter 7, Exercise 1], organized into the two tasks below. Consider

$$h(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x-\xi)_+^3.$$

(i) Write the polynomial pieces for $x \le \xi$ and $x > \xi$ in the form $a + bx + cx^2 + dx^3$. Express their coefficients in terms of $\beta_0, \ldots, \beta_4$ and ξ.

(ii) Show that the two pieces, their first derivatives, and their second derivatives agree at $x = \xi$.

(d) A smoothing penalty (6 points). A smoothing spline minimizes

$$Q_\lambda(g) = \sum_{i=1}^{n} \big(y_i - g(x_i)\big)^2 + \lambda \int_0^1 [g''(x)]^2 \, dx,$$

where $n \ge 2$ and the x_i are distinct and lie in $[0, 1]$. On one dataset, $g_A(x) = 1 + x$ and $g_B(x) = 1 + x + x^2$ have RSS values 6 and 2, respectively.

(i) Compute $\int_0^1 [g''(x)]^2 \, dx$ for each candidate.

(ii) Compute their objective values at $\lambda = 1/2$ and $\lambda = 2$. Which candidate is preferred at each value?

(iii) For the unrestricted smoothing-spline fit, explain why $\lambda \to \infty$ leads to the least-squares straight line.

(e*) (5 bonus points) Complete [JWHT21, Chapter 7, Exercise 2], parts (a)–(e). Give a schematic sketch and a one-sentence explanation for each case. Assume at least three distinct predictor values. Interpret $\lambda = \infty$ as imposing a zero-penalty constraint. For $\lambda = 0$, note whether the curve between observations is uniquely determined.

Problem 3: Regression and smoothing splines in R (30 points total).

This adapts [JWHT21, Chapter 7, Exercise 9] and adds a smoothing-spline comparison. Use `dis` as the predictor and `nox` as the response in `Boston`:

```
data("Boston", package = "ISLR2")
library(splines)
boston <- Boston[, c("dis", "nox")]
n <- nrow(boston)
grid <- data.frame(dis = seq(min(boston$dis), max(boston$dis),
                             length.out = 200))
```

All regression models include an intercept. In `bs(dis, df = d)`, d counts basis columns, so the regression has $d + 1$ coefficients.

Outputs: Build one comparison table with four model rows and columns for training RSS and five-fold CV MSE. Submit the two labeled plots requested in (a) and (d).

(a) Initial fits (5 points).

(i) Fit a cubic polynomial and a cubic regression spline with `df = 4`. Record their training RSS in the first two rows of your table.

(ii) Plot both fitted curves with the data on one scatterplot. Evaluate the curves on `grid`.

(*Hints:* use `poly(dis, 3)` or `bs(dis, df = 4)` in `lm`. Obtain predictions with `predict(fit, newdata = grid)`.)

(b) More flexible regression splines (5 points).

(i) Fit cubic regression splines with `df = 7` and `df = 10`. Add their training RSS as two more rows in the table.

(ii) Explain why the smallest training RSS need not imply the best predictions on new observations.

No additional plot is required.

(c) Five-fold cross-validation (10 points). Use the same folds for all four models:

```
set.seed(2026)
fold <- sample(rep(1:5, length.out = n))
```

(i) For each fold k, refit each model using only `boston[fold != k, ]` and predict on `boston[fold == k, ]`. Construct each basis from its training fold and use `predict` for the held-out observations.

(ii) For each model, divide the sum of all held-out squared errors by n to obtain its CV MSE. Add this column to your table and identify the model with the smallest CV MSE.

Do not construct a new basis separately for the held-out observations.

(d) Smoothing splines (10 points).

(i) Using all observations, fit smoothing splines with effective degrees of freedom 4 and 10 and `all.knots = TRUE`.

(ii) On a second scatterplot of the data, overlay the two smoothing-spline curves and the CV-selected model from part **(c)**, refitted to all observations.

(iii) Compare the smoothness of the two smoothing-spline fits in 1–2 sentences.

(*Hints:* use `smooth.spline(boston$dis, boston$nox, df = d, all.knots = TRUE)` with $d = 4, 10$. Obtain predictions with `predict(fit, x = grid$dis)$y`.) Here `df` is the effective degrees of freedom of the entire fit, not a count of basis columns.

Problem 4: Principal component analysis (20 points total + 5 bonus points).

We will explore PCA conceptually, numerically, and in R, without heavy linear algebra.

(a) (6 points) In your own words, explain what principal components are and why PCA is useful for dimension reduction. Briefly describe how PCA finds directions of maximum variance, verbally or graphically.

(b) (7 points) Consider the eight points:

$$(1,2),\ (2,1.5),\ (3,2.5),\ (4,3),\ (2,3.5),\ (3,4.5),\ (4,5),\ (5,6).$$

(i) Compute the coordinate-wise means and center the points.

(ii) Draw the centered points with equal axis scales. Sketch a line through the origin that appears to maximize the variance of their projections, and briefly explain your choice.

(iii) Describe how to represent each centered point by one number using the first principal component.

You may use R for arithmetic and the scatterplot.

(c) (7 points) Using R:

(i) Create an 8×2 matrix `X` whose rows are the eight original points in the given order.

(ii) Run `pca_res <- prcomp(X, center = TRUE, scale. = FALSE)`.

(iii) Report the first PC direction (up to sign) and the percentage of total variance it explains.

(iv) Compare this direction with your sketch in part **(b)** in one sentence.

(*Hints:* use `pca_res$rotation[, 1]` for the direction. The component variances are `pca_res$sdev^2`.)

(d*) (5 bonus points) Let u be the unit first-PC direction from part **(c)** and $\bar{x}$ the vector of coordinate-wise means. For the original point $x = (1, 2)$:

(i) Compute its first-PC score $z = u^\top(x - \bar{x})$ and its one-PC reconstruction $\tilde{x} = \bar{x} + zu$.

(ii) Compute $\|x - \tilde{x}\|^2$ and explain why replacing u by $-u$ does not change the reconstruction.

References

[JWHT21] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R*, volume 112 of *Springer Texts in Statistics*. Springer, New York, NY, 2nd edition, 2021.