STA 35C – Homework 5

Submission due: Tue, May 27 at 11:59 PM PT

Instructor: Dogyoon Song

Instructions: Upload a PDF file, named with your UC Davis email ID and homework number (e.g., `dgsong_hw5.pdf`) to Canvas ("Homework 5" under "Assignment"). Please make sure to include "STA 35C," your name, and the last four digits of your student ID on the front page. No late homework will be accepted for any reason; submissions after the deadline will receive 0 points. For more details on submission requirements and the late submission policy, see the syllabus.

Problem 1: Bootstrap review (20 points total).

You have an integer-valued random variable with an unknown distribution, whose median m you want to estimate. You collect a sample of size 7 from this population, obtaining the following data:

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

Let $\hat{m}$ be the sample median computed from these 7 observations.

(a) (6 points) If you resample *with replacement* from this original sample of size 7 to generate a new bootstrap sample of size 7, what is the probability of drawing the *exact same* sequence $(2, 0, 0, 1, 2, 3, 2)$ in that bootstrap sample?

(b) (8 points total) Suppose we generated 5 bootstrap samples (each of size 7) as follows. Each column below corresponds to one bootstrap sample, and each row gives the value in that "slot" of the sample:

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

(i) **(4 points)** Compute the bootstrap estimates of the median for each bootstrap sample.

(ii) **(4 points)** Construct a 95% confidence interval for the unknown median m, using:

- $\hat{m}$ estimated from the original sample $(2, 0, 0, 1, 2, 3, 2)$, and
- the *normal approximation*, with the standard deviation estimated from the 5 bootstrapped median estimates. (*Hint*: $z_{0.975} \approx 1.96$.)

(c) (6 points) In this context, how should we interpret "95%" in the 95% confidence interval for m? Specifically, describe succinctly which probability is intended to be approximately 95%.

Problem 2: Regularization & multiple testing review (20 points total).

(a) (10 points) Recall the lasso regression estimates for a linear model is obtained by minimizing

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

As we increase λ from 0 to ∞, how do you expect each of the following to behave?

Pick among these five options: (1) "Remain constant," (2) "Steadily increase," (3) "Steadily decrease," (4) "Decrease initially, and then eventually start increasing in a U shape," or (5) "Increase initially, and then eventually start decreasing in an inverted U shape."

Each question is worth 2 points. Briefly justify your choice in one sentence.

(i) Training RSS (=training MSE)

(ii) Test RSS (=test MSE)

(iii) (Squared) bias

(iv) Variance

(v) Irreducible error

(b) (10 points total) Consider a single null hypothesis H_0; the table on the *left* below shows the probabilities of each outcome ($p_1+p_2+p_3+p_4=1$). Now suppose we have m (e.g. 100) hypotheses tested simultaneously; let N_1, N_2, N_3, N_4 count each outcome, so $N_1+N_2+N_3+N_4=m$.

Single	H_0 is true	H_0 is not true
Reject H_0	p_1	p_2
Not reject H_0	p_3	p_4

Multiple	H_0 is true	H_0 is not true
Reject H_0	N_1	N_2
Not reject H_0	N_3	N_4

(i) **(4 points total)**

- **(2 points)** What is the probability that H_0 is rejected (whether true or not)?
- **(2 points)** Among m null hypotheses, how many are not true (i.e. should ideally be rejected), and how many are actually rejected?

(ii) **(3 points)** Often we reject H_0 at significance level α (e.g. 0.05). Write this requirement as an inequality involving p_1, p_2, p_3, p_4 and α.

(iii) **(3 points)** Suppose instead we aim to control the *false discovery rate (FDR)* at level q (e.g. 0.10). Express this goal as an inequality involving N_1, N_2, N_3, N_4 and q.

Problem 3: Regression & smoothing splines (40 points total + 15 bonus points).

1) (20 points) [JWHT21, Chapter 7, Exercise 1].

2) (20 points) [JWHT21, Chapter 7, Exercise 9].
(*Note:* The `Boston` dataset is contained in the `ISLR2` library.)

3*) (*15 bonus points) [JWHT21, Chapter 7, Exercise 2].

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

We will explore the basics of PCA conceptually, numerically, and through a brief coding demonstration in **R**, all without heavy linear algebra.

(a) (5 points). In your own words, explain what principal components are and why PCA can be useful for dimension reduction. Provide a short conceptual explanation of how PCA "finds" directions of maximum variance, e.g., verbally or graphically.

(b) (5 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 sample mean of the x-coordinates and y-coordinates, then subtract these means from each point to center the data.

(ii) By drawing a scatter plot of these centered points, identify a plausible "best line" of maximum spread. Explain briefly why it seems correct.

(c) (5 points). Using **R**, do the following:

(i) Create an 8×2 matrix `X` from the eight original data points.
(e.g. `X <- rbind(c(1,2), c(2,1.5), ..., c(5,6))`).

(ii) Run PCA using `prcomp(X, center=TRUE, scale=FALSE)` (ensuring you do not scale).

(iii) Report the first principal component vector (from `pca_res$rotation`) and the percentage of variance explained by it.

(iv) In a sentence or two, compare these PCA results to your "best line" direction from part **(b)**.

(d) (5 points). Describe how to reduce these 2D points to 1D using only the first principal component. Explain briefly the benefits of this dimension reduction approach, and what may be lost by discarding the second principal component.

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.