STAT 8670: Assignment 5

Fall 2026 | Computational Methods in Statistics

Chi-Kuang Yeh

STAT 8670 | Fall 2026
Due date: Monday, December 7, 2026
Total: 30 points
Coverage: Chapter 6: Cross Validation and Chapter 7: Jackknife and Bootstrap. Optional and Supplementary material is not required.

Question	Topic / format	Points
1	Cross-validation (6 multiple-choice items)	6
2	Jackknife and bootstrap (6 multiple-choice items)	6
3	Comparing prediction methods (3 short-answer parts)	6
4	Estimator uncertainty (4 short-answer parts)	8
5	Checking a CV implementation (2 errors)	4
Total		**30**

Assignment PDF | Response template (.qmd)

Submission instructions

Use **Quarto** or **R Markdown**. Upload your **PDF and matching source** to **Assignment 5** on **iCollege**. Name them `A5_LastName_FirstName.pdf` and `A5_LastName_FirstName.qmd` (or `.Rmd`). Include your name and course number. Label each question and subpart.

- **Questions 1-2:** Enter exactly one letter, A, B, C, or D, for each item. No explanation is required.
- **Questions 3-4:** Show the formula or substitution for each calculation. Give the requested short explanations.
- Round numerical answers to **three decimal places** only at the end. Report counts as integers.
- **Question 5:** Explain two errors and write corrected R lines.
- **No simulation or plots are required.** Put any R calculations in executable chunks. If you run a simulation, set `set.seed(8670)` once before its repetitions.
- Check that your source renders to your PDF. HTML or screenshots alone are not accepted.

Question 1: Cross-Validation [6]

Select **one best answer** per item (**1 point each**). Unless stated otherwise, observations are independent. The prediction target is a new observation from the same population.

(a) What is tuned? [1]

A polynomial regression uses least squares. CV compares degrees 1, 2, and 3. Which quantity is the **hyperparameter** being tuned?

A. The fitted intercept.
B. The estimated residual variance.
C. The polynomial degree.
D. The number of held-out predictions stored in a vector.

(b) Feature selection [1]

A procedure selects predictors using their associations with the outcome. It then fits a regression. How should CV evaluate this **whole procedure**?

A. Select predictors using all development outcomes. Then cross-validate only the regression fit.
B. Select predictors within each training set. Fit the model there and predict the held-out fold.
C. Select predictors using the final test outcomes.
D. Select predictors using each validation fold's outcomes.

(c) Predicting new people [1]

Each patient has four visits. We want to predict outcomes for **new patients**. How should the folds be formed?

A. Randomly assign visits to folds. One patient's visits may appear in different folds.
B. Use each patient's first visit for validation and their other visits for training.
C. Use patient ID as a predictor. Then split visits randomly.
D. Assign patients to folds. Keep each patient's visits together.

(d) Predicting the future [1]

We want to forecast next month's demand using past monthly records. Which validation design fits this goal?

A. Shuffle months into random folds. Allow later months to predict earlier ones.
B. Train on earlier months and validate on later months.
C. Train on later months and validate on earlier months.
D. Fit to every month and report the training error.

(e) Leave-one-out CV [1]

Use a fixed dataset and a deterministic fitting procedure. Which statement about LOOCV is correct?

A. It uses one fitted model to predict all observations.
B. Its prediction error must be zero because each training set is large.
C. It needs no random fold assignment. Collecting a new dataset can change its score.
D. It is guaranteed to have lower variability than 5-fold CV.

(f) Reusing the test set [1]

An analyst checks the final test error. They then change the polynomial degree to get the lowest error on that same test set. What has happened?

A. The test set now guides model selection. It no longer provides an untouched final assessment.
B. The test set is still untouched because it was not used to fit coefficients.
C. CV automatically adjusts for this extra model selection.
D. A final refit on the development data removes any bias in the selected test score.

Question 2: Jackknife and Bootstrap [6]

Select **one best answer** per item (**1 point each**). Use the ordinary delete-one jackknife and the nonparametric bootstrap. Assume independent observational units.

(a) Resampling paired measurements [1]

Each of 20 patients has a baseline and a follow-up measurement. We want to bootstrap their correlation. How should one bootstrap sample be formed?

A. Independently resample the baseline and follow-up columns.
B. Randomly reorder all 20 patient rows without replacement.
C. Sample 20 patient IDs with replacement. Keep both measurements from each selected patient.
D. Draw 20 baseline measurements and keep the follow-up column in its original order.

(b) What does the histogram describe? [1]

Compute one correlation from each of 5,000 bootstrap samples. What does a histogram of these correlations describe?

A. Variation of the statistic when resampling the observed data. This approximates its sampling distribution.
B. The population distribution of individual patient measurements.
C. The distribution of 5,000 newly recruited patients.
D. The exact sampling distribution under the unknown population, without approximation.

(c) The sign of estimated bias [1]

The observed estimate is $\hat{\theta} = 0.60$, and the bootstrap estimates average 0.57. What are the bootstrap bias estimate and the corresponding bias-corrected estimate?

A. Bias $= 0.03$; corrected estimate $= 0.57$.
B. Bias $= -0.03$; corrected estimate $= 0.63$.
C. Bias $= -0.03$; corrected estimate $= 0.57$.
D. Bias $= 0.57$; corrected estimate $= 0.03$.

(d) Interpreting a confidence procedure [1]

Consider a nominal 95% bootstrap confidence procedure for a population correlation. Assume its approximation is appropriate. What is its frequentist interpretation?

A. About 95% of individual measurements fall inside each interval.
B. It is a 95% prediction interval for the next patient's measurement.
C. Across repeated original samples, about 95% of the intervals contain the fixed population correlation.
D. For the one computed interval, the fixed parameter has probability 0.95 of lying inside it.

(e) More resamples [1]

For the same dataset, increase B from 2,000 to 8,000. The bootstrap draws are independent. Which statement is correct?

A. The conditional Monte Carlo SE of the bootstrap mean is approximately halved. The original sample size stays the same.
B. The original estimator's sampling SE must be halved.
C. The dataset now contains four times as many independent observations.
D. Selection bias in the original sample is removed.

(f) Jackknife versus LOOCV [1]

Both jackknife and LOOCV omit one observation at a time. What is the usual goal of each method?

A. Both estimate prediction error by averaging the retained-sample correlations.
B. Jackknife predicts the omitted outcome; LOOCV estimates bias from parameter replicates.
C. Jackknife requires random fold assignments, whereas LOOCV samples with replacement.
D. Jackknife uses parameter estimates to estimate SE and bias. LOOCV averages held-out losses to assess prediction.

Question 3: Comparing Prediction Methods [6]

Two regression procedures, A and B, use the **same three folds** of 11 independent development observations. For each fold, fit each procedure on the other two folds. Then predict the held-out fold. Each observation receives exactly one held-out prediction.

Fold	Validation observations	Held-out MSE: A	Held-out MSE: B
1	4	1	4
2	4	1	4
3	3	9	2

Each MSE averages squared prediction errors **within that fold**. Treat the values as exact. Use the definition in Section 6.3: How K-Fold CV Works.

(a) Overall CV error [2]

Calculate the overall CV MSE for **A** and **B** (**1 point each**). Give each held-out observation equal weight. Show your weights or your total squared-error calculation.

(b) Model comparison [2]

Which procedure should CV select [**1**]? Why is an unweighted average of the three fold MSEs inappropriate here [**1**]? Explain in one or two sentences.

(c) Final assessment [2]

Refit the selected procedure on **all 11 development observations**. Its predictions on a separate test set are shown below. This test set was not used for preprocessing, fitting, or model selection.

Test observation	Observed outcome	Prediction
1	1	1
2	2	1
3	3	5
4	4	4

Calculate the test MSE [**1**]. Give one reason why it may differ from the selected procedure's CV MSE [**1**]. See Section 6.5: Select, Refit, and Test.

Question 4: Estimator Uncertainty [8]

There are $n = 10$ **independent units**. Each unit has two paired measurements. The statistic is their sample correlation.

The table summarizes the full sample, all 10 jackknife replicates, and $B = 5000$ bootstrap replicates. **Treat the displayed values as exact**. You do not need raw data or programming.

Quantity	Value
Full-sample correlation $\hat{\theta}$	0.72287
Mean of jackknife replicates $\bar{\theta}_{(\cdot)}$	0.72483
Raw sample SD of jackknife replicates	0.04992
Mean of bootstrap replicates $\bar{\theta}^*$	0.72744
Sample SD of bootstrap replicates	0.14941
Bootstrap 2.5th percentile	0.36387
Bootstrap 97.5th percentile	0.92512

The raw jackknife SD uses denominator $n-1$, as R's `sd()` does. The bootstrap SD uses denominator $B-1$. Use the formulas in Section 7.5: Jackknife, Section 7.4: Bootstrap, and Section 7.6: Bootstrap Confidence Interval.

(a) Jackknife SE and bias [2]

Calculate the jackknife SE [**1**] and jackknife bias estimate [**1**]. Show both calculations.

(b) Bias correction [2]

Calculate the jackknife bias-corrected estimate, $\hat{\theta} - \widehat{\text{bias}}_{\text{jack}}$ [**1**]. Does bias correction always reduce mean squared error? Explain in one sentence [**1**].

(c) Bootstrap SE and bias [2]

Report the bootstrap SE [**1**] and calculate the bootstrap bias estimate [**1**].

(d) Percentile interval [2]

Report the nominal 95% bootstrap percentile interval (**1 point for both endpoints**). The interval concerns the **population correlation**. Explain why it does not predict a new unit's measurement [**1**]. Use one sentence.

Question 5: Checking a CV Implementation [4]

Suppose `dat` has 20 independent observations. It contains numeric columns `x` and `y`, with no missing values. The rows are in arbitrary order. We want the **5-fold CV MSE** for the regression model `y ~ x`. Use the five folds given below.

The following R code runs, but contains **two statistical errors**:

```r
n <- nrow(dat)
fold_id <- rep(1:5, each = 4)
pred <- numeric(n)

fit <- lm(y ~ x, data = dat)
for (k in 1:5) {
  held_out <- fold_id == k
  pred[held_out] <- predict(fit, newdata = dat[held_out, ])
}

cv_mse <- mean(dat$y - pred)^2
```

(a) First error [2]

Identify one error. Explain why it gives an incorrect CV MSE [**1**]. Write corrected R line(s) and state where they belong [**1**].

(b) Second error [2]

Identify a second error. Explain the problem [**1**]. Write corrected R line(s) and state where they belong [**1**].

List the errors in either order. Code fragments are enough. **No code execution or numerical CV score is required.** See Section 6.3: How K-Fold CV Works.