In-Class Test 1

36–402

March 17, 2022

SOLUTIONS

Exam Instructions:

- We will distribute exams promptly at 8:35am. The exam is due at **9:55am**.
- You may not discuss the contents of this exam with other students **until grades are released**.
- Write your answers in the boxes provided. If your answer doesn't fit, you may be writing too much—check if you can simplify your answer.
- Please write legibly; we will not give credit for what we cannot read.
- You may use the back of each page for scratch work, but your answers must be in the boxes provided. If you need additional scratch paper, ask us.
- You may use **one** normal-size (letter or A4) piece of paper with notes.
- No calculators, computers, or other electronic devices.
- Definitions and formulas are provided when needed by a specific question.
- Your exam will be returned to you via Gradescope.

Useful Formulas

You may use these results anywhere you feel they are necessary.

Expectation

- When $a \in \mathbb{R}$ is a constant, $\mathbb{E}[aX] = a\mathbb{E}[X]$
- When $a \in \mathbb{R}^p$ is a constant vector and X is a p-dimensional random vector, $\mathbb{E}[a^T X] = a^T \mathbb{E}[X]$
- When $A \in \mathbb{R}^{n \times p}$ is a constant matrix and X is a p-dimensional random vector, $\mathbb{E}[AX] = A\mathbb{E}[X]$
- When X and Y are random variables, $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
- When X and Y are independent, $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$
- $\mathbb{E}[X] = \mathbb{E}\left[\mathbb{E}[X \mid Y]\right]$, where the outer expectation averages over Y

Variance and Covariance

- When $a \in \mathbb{R}$ is a constant, $\operatorname{Var}(aX) = a^2 \operatorname{Var}(X)$
- When $a \in \mathbb{R}^p$ is a constant vector and X is a p-dimensional random vector, $\operatorname{Var}(a^T X) = a^T \operatorname{Var}(X) a$
- When $A \in \mathbb{R}^{n \times p}$ is a constant matrix and X is a p-dimensional random vector, $\operatorname{Var}[AX] = A \operatorname{Var}[X] A^T$
- When $a_i \in \mathbb{R}$ are constants and X_i are random variables,

$$\operatorname{Var}\left(\sum_i a_i X_i\right) = \sum_i a_i^2 \operatorname{Var}(X_i) + 2 \sum_j \sum_{i<j} a_i a_j \operatorname{Cov}(X_i, X_j)$$

- When $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ are random variables and $a_1, \ldots, a_m$ and $b_1, \ldots, b_n$ are constants, then

$$\operatorname{Cov}\left(\sum_{i=1}^m a_i X_i, \sum_{j=1}^n b_j Y_j\right) = \sum_{i=1}^m \sum_{j=1}^n a_i b_j \operatorname{Cov}(X_i, Y_j)$$

- The sample variance of the observations $x_1, x_2, \ldots, x_n$ is

$$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2,$$

  where $\bar{x}$ is the sample mean of the observations.

- The sample standard deviation of the observations $x_1, x_2, \ldots, x_n$ is the square root of its sample variance.

Multiple-Choice Questions

Select **only one** answer to these questions. Mark your choice by clearly circling the letter. You do not need to write any derivations or justification, and there is no partial credit.

1. We observe $(X_1, Y_1), \ldots, (X_n, Y_n)$ and consider fitting different regression models. The effective degrees of freedom of a model $\hat{r}(x)$:

   A. Measures the effective number of parameters used by an estimator

   B. Is random, i.e., it depends on the data

   C. Is always an integer

   D. Cannot be negative

   E. Both (A) and (C) are true

   F. (A), (C), and (D) are all true

   G. (A) through (D) are true

2. Given a sample of size n, leave-one-out cross-validation is the same as

   A. 1-fold cross-validation

   B. 2-fold cross-validation

   C. $(n-1)$-fold cross-validation

   D. n-fold cross-validation

   E. None of the above

3. We have done several examples of K-fold cross-validation. Each instance produces an estimate of prediction MSE. As we increase K from 2 to n (the size of the sample), the K-fold cross-validation estimate of prediction MSE

   A. has more bias but less variance

   B. has less bias but more variance

   C. could have bias and variance that increase or decrease, depending on the data

4. Consider fitting a regression model $\hat{r}(x)$ to training data $(X_1, Y_1), \ldots, (X_n, Y_n)$. We also have an independent test sample $(X'_1, Y'_1), \ldots, (X'_m, Y'_m)$ drawn from the same population.

   (a) We define the training and testing error as:

$$\text{training error} = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \hat{r}(X_i)\right)^2$$

$$\text{test error} = \frac{1}{m} \sum_{i=1}^{m} \left(Y'_i - \hat{r}(X'_i)\right)^2$$

   How do these errors relate to each other? Choose one.

   A. Which error is larger or smaller depends on the data and the model, and whether we have correctly chosen tuning parameters such as the bandwidth

   B. The training error is always smaller than the test error

   C. The training error is always larger than the test error

   (b) We can define the expected training and test errors by taking the expectations of the definitions above.

   For a nontrivial model, such as a linear model or kernel smoother, how do the expected training and test errors relate to each other?

   A. The expected test error is always smaller than the expected training error

   B. The expected test error is never smaller than the expected training error

   C. Which expected error is larger or smaller depends on the data and the model, and whether we have correctly chosen tuning parameters such as the bandwidth

   (c) The expectation in the expected test error averages over:

   A. All possible test datasets from the population, for a specific type of model and specific training data set

   B. All possible training datasets from the population, for a specific type of model and specific training data set

   C. All possible training and test datasets from the population, for a specific type of model

Short Answer Questions

Write your answers inside the boxes provided. For questions that ask for explanation, justify each answer with **only a few sentences**.

5. Suppose

$$Y = \beta_1 \log(X) + \beta_2 Z^2 + \varepsilon,$$

where $\varepsilon \sim N(0, 1)$, and X and Z are normally distributed random variables.

If you regress Y on $\log(X)$ only (with Z omitted), will the slope be an unbiased estimator of the coefficient β_1 in the equation above? Briefly justify why or why not.

Solution: In general, no, because omitting a variable means the residuals are now correlated with X. It's okay to omit Z in a regression **if** Y and Z are independent conditional on X, because then $\mathbb{E}(Y \mid X, Z) = \mathbb{E}(Y \mid X)$.

Student answers should say no, and indicate this is because $\log(X)$ and Z may be dependent. They can also simply say X and Z can be dependent.

6. Bootstrap methods rely on fewer assumptions than the built-in *R* functions (such as `lm`). Do they always lead to more accurate results but at a higher computational cost? Briefly explain why or why not.

Solution: No, not necessarily. If the parametric assumptions are correct, then parametric models generally have lower variance. Recall also that the bootstrap involves two steps of approximation, and so may not be exact. The main consideration is that parametric models can have non-vanishing bias due to model misspecifications.

7. Consider fitting a regression model $\hat{r}(x)$ to training data, with the goal of achieving the best prediction risk (measured by the expected squared error) on new observations. You have training data $(X_1, Y_1), \ldots, (X_n, Y_n)$, and you fit two models:

   Model 1: A smoothing spline that minimizes

   $$\sum_{j=1}^{n} (Y_j - \hat{r}(X_j))^2 + \lambda \int [\hat{r}''(x)]^2 \, \mathrm{d}x.$$

   The penalty parameter λ is chosen to minimize the prediction risk of the model.

   Model 2: An ordinary linear regression model.

   You run leave-one-out cross-validation (LOOCV) to estimate the prediction risk of each model. For Model 1, you estimate a prediction risk of 104.7; for Model 2, you estimate a prediction risk of 108.9.

   (a) Explain why these results do not definitively prove that Model 1 is superior to Model 2. State where any uncertainty in the results comes from. Be brief; a couple sentences is enough.

   Solution: LOOCV *estimates* the generalization error using the training data. Because it is an estimate from data, it is a random variable; if we had a different training set, we will get a slightly different LOOCV estimate of error. Depending on the sampling distribution of the error, it is possible that Model 1 actually performs worse than Model 2 in general.

   Note that unlike k-fold cross-validation, running LOOCV twice on the *same* dataset will give the same answer; so the randomness does not come from LOOCV itself being random.

   Student answers must say the errors are estimates or that they are random variables, and indicate that the randomness is because the errors are estimated from a (finite) sample. Simply saying that CV is random is not sufficient.

(b) Because of this uncertainty, you decide to define the statistic

$$T = |(\text{LOOCV error of model 2}) - (\text{LOOCV error of model 1})|$$

for the absolute difference between the two LOOCV errors. Give a step-by-step description of how you can use the bootstrap to estimate the sampling distribution of T, and the standard deviation of T. Suppose you are unwilling to make any assumptions about the distribution of the residuals of both models.

Your steps should be specific: if you randomly draw some values, indicate how many must be drawn; if you repeat some steps several times, indicate how many times they should be repeated; if you refit Model 1, indicate what to do about λ. You don't need to elaborate on or explain your steps.

Solution:

1. Draw n pairs (X_i, Y_i) from the training data **with replacement**. Call this the bootstrap sample.
2. Refit Model 1 to the bootstrap sample. Choose λ using the same procedure you used to choose it on the training data.
3. Compute the LOOCV estimate of the prediction error of model 1.
4. Refit Model 2 to the bootstrap sample. Compute the LOOCV estimate of the prediction error of model 2.
5. Calculate T^*, the absolute difference between the two LOOCV estimates.
6. Repeat steps 1–5 as many times as is practical; 1000 is a good number, but more is better.

The resulting T^* values give our estimate of the sampling distribution of T.

Student answers can express this in different ways, but key features to look for are:

- The bootstrap sampling must be with replacement
- The bootstrap sample must be of size n
- They must reselect λ on the bootstrap sample
- They must calculate the difference between the LOOCV estimates on the bootstrap, and use the difference to make the distribution
- They should indicate the bootstrapping should be repeated a large number of times

(c) Suppose that from the results of part (b), you decide that T is approximately normally distributed. Describe how you can use your bootstrap results to construct a normal-based $(1 - \alpha)$ confidence interval for the population parameter $\theta = \mathbb{E}[T]$.

Solution: Take $T \pm a \times$ standard deviation of T, where the SD is of the bootstrap Ts and a is the appropriate quantile of the standard normal distribution.

Students may write ± 1.96 or ± 2 for a 95% CI, but we asked for $1 - \alpha$, so this should not receive full credit. They also should not divide the SD by $\sqrt{B}$ or $\sqrt{n}$.

Pivotal or quantile confidence intervals do not get credit; the answer must be normal-based.

(d) How can you use your estimated confidence interval for θ to decide which regression model to choose? Suppose you decide that the two models have similar prediction performance; which model should you choose then and why?

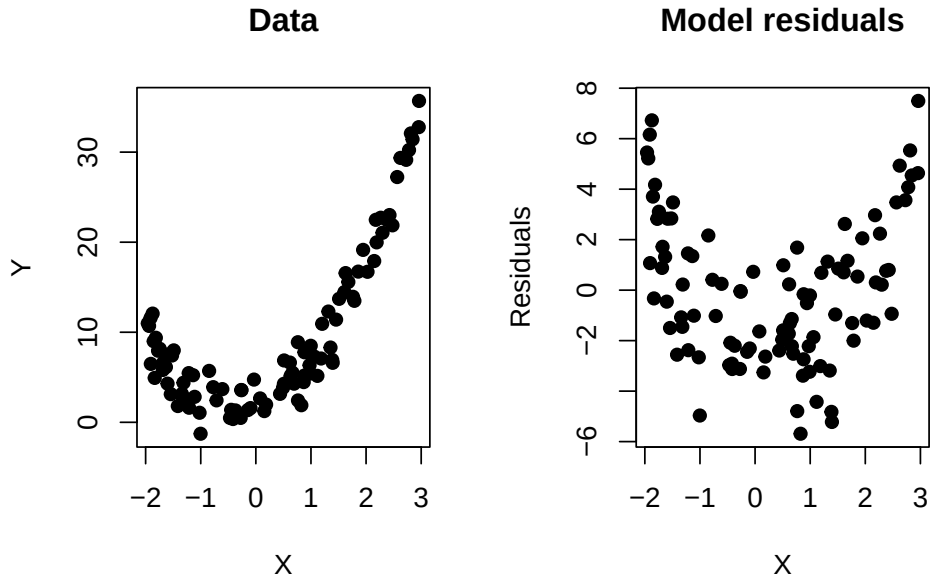
Solution: If the confidence interval contains 0, there is no statistically significant difference in the prediction risks. If the models have similar prediction performance, we should pick the simpler one, Model 2. We generally prefer simpler models over more complicated ones that don't work better.

If the confidence interval does not contain 0, the difference between prediction risks is statistically significant, so we pick the model with the smallest estimated prediction error, Model 1.

This is like performing a hypothesis test for the difference between the two model errors.

8. You have some data $(X_1, Y_1), \ldots, (X_n, Y_n)$. After making a scatterplot, you conclude the relationship is not linear, so you decide to use the `npreg` to fit a kernel smoother. You choose a bandwidth of $h = 1.5$.

   After fitting the model, you make a plot of the residuals $\hat{\epsilon}_i$ versus X_i. Here are plots of the data and the residuals:



   (a) What do the residuals suggest about the model fit, and what could you change to improve the fit? (Consider only changes to the kernel smoother, rather than switching to a different kind of model entirely.)

   Solution: The residuals suggest the kernel smoother is not fitting the mean response well. Perhaps the bandwidth is too large, so it is oversmoothing (or underfitting) the data. We could try using a smaller bandwidth h.

   (b) For kernel smoothers, the bandwidth h controls the flexibility or effective degrees of freedom (edf) of the model. For smoothing splines, the penalty λ has a similar role.

   Briefly state what happens to a kernel smoother if we let h be nearly zero, and what happens if we let h be extremely large. Explain what is similar and what is different if we have a smoothing spline and let λ be nearly zero or let λ be extremely large.

   Solution: With h nearly zero, a kernel smoother interpolates between the data and is very flexible. This is similar to a smoothing spline if λ is nearly zero.

   When h is large, a kernel smoother gets increasingly flat; in the limit, it will become a horizontal line and return the sample mean of the observed response values. When λ

is large, a smoothing spline becomes a straight line (more specifically the least squares line), but that line can still have a nonzero slope, unlike for a kernel smoother.

(c) For the data plotted above, which do you expect would fit better: A kernel smoother with 2–4 effective degrees of freedom, or a smoothing spline with similar effective degrees of freedom? Give a brief justification for your choice. *Hint:* Consider your answer to part (b), when each model has low effective degrees of freedom.

Solution: Smoothing spline. Splines will for decreasing EDF converge to the least squares line; a kernel smoother with too low effective degrees of freedom will just average everything out, and as the edf gets too low, the fit will get closer to a flat line (the average of all Y's). The trend in the plot has a slope, so a smoothing spline with low edf would be better able to fit that slope.

9. For each subject $i = 1, \ldots, n$ in a study, we observe a treatment $X_i \in \{0, 1\}$ and a response $Y_i \in \{0, 1\}$. We are interested in the causal effect of X on Y.

   When we discussed causal inference, we introduced the counterfactual function $C_i(\cdot)$ for each subject. We defined the average treatment effect as $\theta = \mathbb{E}[C_i(1)] - \mathbb{E}[C_i(0)]$.

   (a) Explain why we cannot use the observed data to estimate θ by taking the average of $C_1(1), \ldots, C_n(1)$ minus the average of $C_1(0), \ldots, C_n(0)$.

   Solution: We only observe $C_i(X_i)$. That is, we only observe the outcome for the treatment the subject received, not for the treatment it did not receive; and so we do not know the value of $C_i(1)$ for all i, or of $C_i(0)$ for all i. We can't take the averages.

   (b) Suppose the study was conducted by randomly assigning each subject to treatment 0 or 1 with equal probability. At the time of assignment, the scientist also recorded the variable Z_i for each subject; Z is known to be related to both X and Y in the population. Unfortunately, however, the scientist's dog ate the part of the spreadsheet containing the values of Z_i, so the scientist only has $(X_1, Y_1), \ldots, (X_n, Y_n)$.

   Explain if it is possible to estimate the average treatment effect θ. If it is, say how to do so; if it is not, explain why not.

   Solution: It **is** possible. Because the treatment was randomly assigned, Z is not a confounding variable—it can have no causal effect on X. We can simply compute

   $$\hat{\theta} = \hat{\mathbb{E}}[Y_i \mid X_i = 1] - \hat{\mathbb{E}}[Y_i \mid X_i = 0].$$

   In words, we can take the average Y for subjects with $X = 1$, and subtract the average Y for subjects with $X = 0$.