

STA 35C – Homework 4

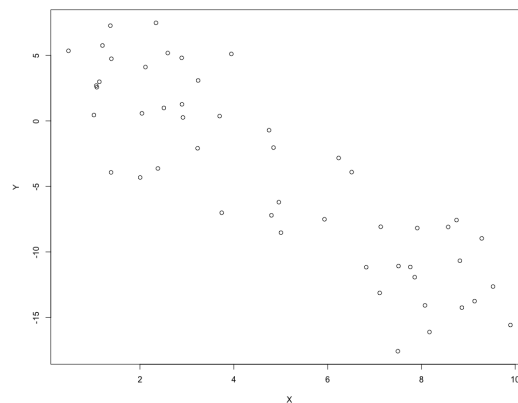
Submission due: Thu, October 29 at 11:59 PM PT

Instructor: Dogyoon Song

Instructions: Upload a **single PDF file** to Gradescope via Canvas (“Homework 4” under “Assignments”). Name the file using the prefix of your UC Davis email ID and the homework number (e.g., `dgsong_hw4.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 any requested plots/tables in the PDF. No late submissions will be accepted; any submission received after the deadline will receive 0 points. For full information about submission requirements and the late submission policy, see the syllabus.

Problem 1: Regression review (15 points in total).

- (a) (5 points) On the scatterplot below, sketch a plausible least-squares regression line and indicate its prediction at $X = 6$. An exact fit is not required.



- (b) (6 points) A simple linear regression fitted to a separate sample of $n = 22$ observations gives:

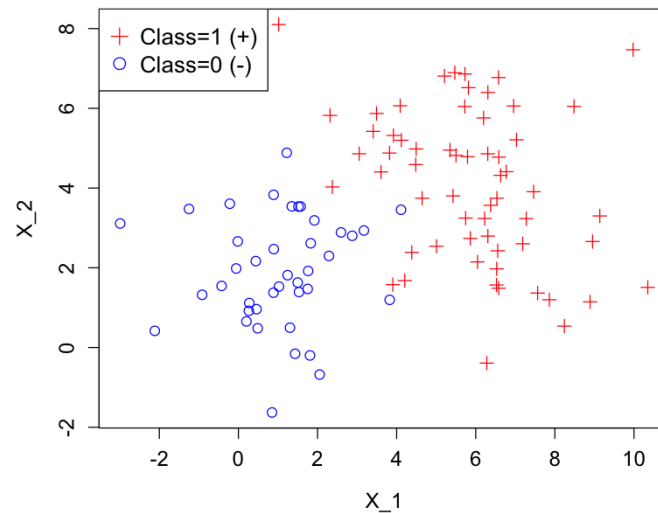
Coefficient	Estimate	Std. Error
Intercept	−2.0	0.5
X	1.5	0.4

Assume independent Gaussian errors with mean zero and constant variance, conditional on the predictors. Test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ at the 5% level. Report the t -statistic, degrees of freedom, two-sided p -value, and conclusion. Briefly interpret the slope. (*Hint:* use `2 * pt(-abs(t), df)` in R.)

- (c) (4 points) Adding X_2 to the model in (b) changes the estimated coefficient of X from 1.5 to -0.2 . Explain how adjustment can reverse this association, and interpret -0.2 while holding X_2 fixed. Do not make a causal claim.

Problem 2: Classification review (15 points in total).

Website visitors are labeled $Y = 1$ (subscriber) or $Y = 0$ (non-subscriber). Predictors X_1 and X_2 describe user behavior. Treat $Y = 1$ as the positive class. Consider a logistic model with linear predictor $\beta_0 + \beta_1 X_1 + \beta_2 X_2$. Predict 1 when $\hat{p}(X_1, X_2) \geq p^*$, and 0 otherwise.



- (a) (4 points) Sketch a plausible straight decision boundary and label the two predicted regions. Exact fitting is not required. Mark $A = (6, 2)$ and $B = (1, 3)$ and classify them using *your* boundary.
- (b) (3 points) Define a false positive and a false negative in this context.
- (c) (4 points) Separately from your sketch, a fitted model at $p^* = 0.5$ yields:

	Pred = 1	Pred = 0
$Y = 1$	54	6
$Y = 0$	5	35

Compute the error rate, TPR, and FPR. For this fixed model and evaluation set, can either TPR or FPR increase when the cutoff rises to 0.9? Explain briefly.

- (d) (4 points) Suppose false negatives are especially costly. Using a threshold-labeled ROC curve computed on *validation data*, describe how to choose a cutoff that minimizes FPR subject to FNR being *at most* 0.1 (equivalently, $\text{TPR} \geq 0.9$) on those data. Explain why lowering the cutoff can help meet the constraint. No numerical cutoff is required.

Problem 3: Bias–variance and prediction (25 points in total).

For a fixed input x_0 , let a new response be $Y_0 = f(x_0) + \varepsilon_0$, with $\mathbb{E}[\varepsilon_0] = 0$ and $\text{Var}(\varepsilon_0) = 1$. The new noise is independent of the training data used to fit $\hat{f}$. You may use

$$\mathbb{E}[(Y_0 - \hat{f}(x_0))^2] = (\mathbb{E}[\hat{f}(x_0)] - f(x_0))^2 + \text{Var}(\hat{f}(x_0)) + 1.$$

The expectation averages over both the training data and the new noise.

- (a) **(10 points)** Let x_0 denote a (fixed) test point. Suppose $f(x_0) = 2$. Three hypothetical methods have the following predictions across four equally likely training-data scenarios:

Scenario	Method A	Method B	Method C
1	0.5	1	-1
2	1.0	2	1
3	1.0	2	3
4	1.5	3	5

Treat these as the only possible scenarios, each with probability $1/4$. For each method, report the mean prediction, squared bias, variance, and expected squared prediction error in one table. Which method has the smallest prediction error, and why does zero bias alone not ensure the smallest error?

- (b) **(10 points)** Now use the following independent training and test samples:

```
set.seed(2026)
f <- function(x) 2 + x + x^2
make_data <- function(n) {
  x <- runif(n, -2, 2)
  data.frame(x = x, y = f(x) + rnorm(n))
}
train <- make_data(40)
test <- make_data(1000)
```

Fit polynomial regressions of degrees 1, 2, 4, and 16, each with an intercept, to the *same training set*. Report training and test MSE for each degree in one four-row table. Here $\text{MSE} = m^{-1} \sum_{i=1}^m (y_i - \hat{y}_i)^2$ on an evaluation set of size m . No repeated simulation or plots are required. (*Hints:* use `lm(y ~ poly(x, d), data = train)` and `predict` with `newdata = test`.)

- (c) **(5 points)** In 3–4 sentences, explain why the training MSE cannot increase as the degree rises, but the test MSE need not decrease. Relate this distinction to your table. Do the two MSEs from this single run separately estimate the method’s squared bias and variance over training samples? Explain.

Problem 4: Cross-validation (20 points in total + 5 bonus points).

Use `Auto` from `ISLR2` and polynomial regressions of `mpg` on `horsepower` of degrees 1, 2, 3, 4, 5, each with an intercept. This exercise adapts the polynomial-model comparison in *An Introduction to Statistical Learning with Applications in R*, second edition, Chapter 5, Exercise 8. Load the data with:

```
data("Auto", package = "ISLR2")
auto <- Auto[, c("mpg", "horsepower")]
n <- nrow(auto)
```

Use `R`. Fit only on the training observations for each split or fold.

- (a) **(8 points)** Make a random 50/50 split using `set.seed(2026)` followed by `train_id <- sample.int(n, n/2)`; use the remaining observations for validation. Compute validation MSE for all five degrees using this same split. Repeat with seed 35. Report both columns of MSEs in one table and identify the minimizing degree for each split. (*Hint*: fit `mpg ~ poly(horsepower, d)` using `data = auto[train_id, ]`, and predict on `auto[-train_id, ]`.)
- (b) **(7 points)** Compute the LOOCV MSE for each degree and add it as a third column to your table. Identify the minimizing degree. You may use `boot::cv.glm(auto, fit)$delta[1]`, where `fit` is fitted by `glm(mpg ~ poly(horsepower, d), data = auto)` with its default Gaussian family.
- (c) **(5 points)** Compare one validation split with LOOCV. Report the training sample size and the number of fits per degree under direct refitting. In 2–3 sentences, explain how the methods differ in split randomness and whether eliminating split randomness eliminates uncertainty across datasets.
- (d*) **(5 bonus points)** Add a 10-fold CV column to the same table using

```
set.seed(2026)
fold <- sample(rep(1:10, length.out = n))
```

Use these same folds for all degrees. Compute each CV MSE as the sum of all held-out squared errors divided by n , so that unequal fold sizes are handled correctly. Identify the minimizing degree and briefly compare with LOOCV.

Problem 5: The bootstrap (25 points in total + 5 bonus points).

- (a) **(5 points)** Ten observed tosses of a fair coin contain six heads and four tails. Consider a bootstrap sample of size ten drawn independently with replacement from these observations. Conditional on the observed sample, give the distribution and expectation of the number of bootstrap heads. Compare with ten fresh independent flips of the fair coin, and explain why the two distributions differ.
- (b) **(5 points)** For the remaining required parts, generate *one* dataset:

```
set.seed(2026)
x <- rnorm(100)
y <- -1 + 2*x + rnorm(100)
dat <- data.frame(x, y)
```

Fit $Y = \beta_0 + \beta_1 X + \varepsilon$ by least squares. Report the fitted slope, its usual model-based standard error, and its 95% t confidence interval. (*Hints:* `summary` and `confint`. The true slope is 2.)

- (c) **(10 points)** Set `set.seed(35)` once before generating $B = 500$ bootstrap samples from `dat`. Each sample must contain 100 *rows* drawn with replacement, keeping each (x_i, y_i) pair together. Refit the regression on each resample and store the fitted slope. Estimate the slope's standard error by the standard deviation of these 500 slopes, and report the normal-approximation bootstrap interval

$$\hat{\beta}_1 \pm 1.96 \widehat{\text{SE}}_{\text{boot}},$$

centered at the original fitted slope. (*Hints:* use `sample.int(100, 100, replace = TRUE)` for row indices and `sd` for the bootstrap standard error.)

- (d) **(5 points)** In 3–4 sentences, compare the two standard errors and intervals. Explain why resampling x and y independently would be inappropriate. Does increasing B reduce the sampling uncertainty of the original slope estimate or only the simulation error in the bootstrap approximation? Explain briefly.
- (e*) **(5 bonus points)** Using `set.seed(100)` once, simulate 200 fresh independent datasets of size 100 from the model in part (b). Fit a regression to each and compute its model-based 95% t confidence interval for the slope. Report the fraction containing 2 and explain why this repeated-sampling check differs from checking whether a single interval contains 2. The fraction need not equal 0.95 exactly. Do not bootstrap these 200 datasets.