STA 35C – Homework 3

Submission due: Tue, April 21 at 11:59 PM PT

Instructor: Dogyoon Song

Instructions: Upload a single PDF file to Gradescope via Canvas ("Homework 3" under "Assignments"). Name the file using the prefix of your UC Davis email ID and the homework number (e.g., `dgsong_hw3.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.

There are three bonus problems in this problem set, each of which is worth 5 bonus points. At most two bonus subproblems will count, for a maximum of 10 bonus points.

Problem 1 (30 points in total + 5 bonus points).

You are analyzing an outcome measure Y (e.g., blood pressure or some clinical score) collected from n patients. Some patients received a *new treatment* ($D = 1$), while others received *standard* (control) treatment ($D = 0$). You also have an additional continuous predictor X (e.g., patient age or baseline risk). You suspect that Y may differ substantially between the two groups, possibly due to the different treatments introduced. Throughout, we assume that *smaller values of Y are better*.

Data: Your dataset consists of n tuples:

$$(y_1, d_1, x_1),\ \ldots,\ (y_n, d_n, x_n),$$

where y_i is the response, $d_i \in \{0, 1\}$ is treatment indicator, and x_i is the additional contextual predictor for patient $i \in [n]$. For concreteness, suppose you collected 10 data points as given below:

$(-9.3704, 1, 39.107)$, $(-9.2399, 1, 38.245)$, $(-8.1464, 0, 24.318)$, $(-7.1728, 1, 36.825)$, $(-6.3464, 1, 29.314)$,
$(-5.8743, 0, 23.497)$, $(-5.6763, 1, 31.137)$, $(-1.8406, 0, 14.919)$, $(-0.8549, 0, 13.691)$, $(1.0222, 0, 10.631)$.

Objective: As a working model, suppose

$$\begin{cases} \text{If } D = 1: & Y = a + b\,X + \varepsilon \\ \text{If } D = 0: & Y = c + d\,X + \varepsilon \end{cases}$$

where a, b, c, d are unknown constants and ε is random noise. We want to compare the treatment and control groups and assess whether the treatment group tends to have lower mean responses. A useful model-based treatment–control contrast over the observed x_i values is

$$\tau = \frac{1}{10} \sum_{i=1}^{10} [\,(a + bx_i) - (c + dx_i)\,],$$

which we want to estimate from data. In this problem, you will compare the two groups using methods learned so far. You may use R throughout.

(a) (5 points) Initially, compare the average outcome in each group:

$$\bar{y}_{\text{ctrl}} = \frac{1}{n_0} \sum_{i:\, d_i=0} y_i, \qquad \bar{y}_{\text{treat}} = \frac{1}{n_1} \sum_{i:\, d_i=1} y_i,$$

where n_0 is the number of patients with $D = 0$, and n_1 the number with $D = 1$ (here $n_0 = n_1 = 5$).

Compute $\bar{y}_{\text{ctrl}}$, $\bar{y}_{\text{treat}}$, and the raw difference

$$\bar{y}_{\text{treat}} - \bar{y}_{\text{ctrl}}.$$

Based only on these group means, which group appears to have lower outcomes? Then, at the 5% significance level, test $H_0 : \mu_{\text{ctrl}} = \mu_{\text{treat}}$ using the equal-variance two-sample t-test from STA 35B. (Note: Using Welch's t-test is also fine here, but it does not assume equal variances)

(b) (10 points) Now consider a simple linear model that only takes D into account, *ignoring* X:

$$Y \;=\; \beta_0 + \beta_1 D + \varepsilon.$$

(i) Interpret β_0 and β_1.

(ii) At the 5% significance level, test $H_0 : \beta_1 = 0$. What does your conclusion imply about the difference between treatment and control in this model?

(iii) Compare your conclusion about $\hat{\beta}_1$ drawn from this dummy model to your conclusion about the raw difference $\bar{y}_{\text{treat}} - \bar{y}_{\text{ctrl}}$ from part **(a)**. What do you notice?

(c) (10 points) Next, you include the additional predictor X into your model:

$$Y = \beta_0 + \beta_1 D + \beta_2 X + \varepsilon.$$

(i) Explain how β_2 is interpreted in this model.

(ii) Does including X change how we interpret β_1 relative to part **(b)**? If yes, how?

(iii) Why might controlling for X (i.e., conditioning on a value of X) alter the apparent difference between groups? (*Hint:* It may be helpful to try a scatter plot to see how X is distributed by group.)

(d) (5 points) Finally, fit the model with an interaction term

$$Y = \beta_0 + \beta_1 D \;+\; \beta_2 X \;+\; \beta_3 (D \times X) + \varepsilon$$

which is equivalent to

$$Y = \begin{cases} \beta_0 + \beta_1 \;+\; (\beta_2 + \beta_3) X + \varepsilon & \text{if } D = 1, \\ \beta_0 \;+\; \beta_2 X + \varepsilon & \text{if } D = 0. \end{cases}$$

(i) Explain how β_3 captures a possible difference in *slopes* across the two groups.

(ii) Using this fitted model, test $H_0 : \beta_3 = 0$ at the 5% significance level. Would you conclude that the two groups have the same slope?

(e*) (5 bonus points) Using the interaction model from part (d), provide estimates of a, b, c, d and τ. What do your fitted values suggest about τ?

Problem 2 (30 points in total + 5 bonus points).

Scenario: A company administers a mandatory test to new hires in a training program. Each candidate either *Passes* ($Y = 1$) or *Fails* ($Y = 0$). The HR department believes that prior work experience (X, in months) influences the probability of passing.

(a) (8 points) Assume a logistic regression model

$$\Pr(Y = 1 \mid X = x) = \frac{1}{1 + e^{-(\alpha+\beta x)}}.$$

(i) Interpret α and β in this context. (*Hint:* Think about how log-odds is affected.)

(ii) How does a 1-month increase in X affect the *odds* of passing?

(b) (12 points) Suppose HR policy says if $\hat{p}(x) \geq 0.4$, the candidate is predicted to pass and placed in an "advanced" track; otherwise, they go to a "basic" track.

(i) If $\alpha = -3.0$ and $\beta = 0.07$, compute $\hat{p}(x)$ for $x = 20$ months.

(ii) Which track would that candidate be assigned to?

(iii) If a candidate actually passes ($Y = 1$) but is assigned to the "basic" track under the 0.4 cutoff, is this a false positive or a false negative? Briefly explain.

(iv) Discuss the pros and cons of using 0.4 rather than 0.5 as a threshold. In particular, consider which type of misclassification might be costlier for the company.

(c) (10 points) Implement the following in an R script and report your results.

(i) Using `set.seed(2026)`, simulate a dataset of size $n = 50$ by letting $X \sim \text{Uniform}(0, 30)$ and

$$\Pr(Y = 1 \mid X = x) = \frac{1}{1 + e^{-(-5+0.25\,x)}}.$$

Generate Y accordingly (*Hint:* Use `rbinom()` in R).

(ii) Fit a logistic model in R (via `glm(..., family=binomial)`).

(iii) Print the estimated coefficients. Compare them to ($\alpha = -5$, $\beta = 0.25$). Are they close?

(iv) Plot the fitted logistic curve and the data points.

(v) Using the simulated sample itself, produce predictions at thresholds 0.4 and 0.5, treating $Y = 1$ (Pass) as the positive class, and create a confusion matrix for each. Which threshold yields more false negatives versus false positives?

(vi) Briefly discuss how a larger sample size might improve the parameter estimates.

(d*) (5 bonus points) Suppose HR also records whether a new hire has a professional certificate, say $C \in \{0, 1\}$.

(i) Propose how to include C in your logistic regression (along with X).

(ii) Discuss how you would interpret the coefficient of C and, potentially, any interaction $C \times X$.

(iii) Briefly outline how you might compare the performance of this extended model to your original logistic regression model.

Problem 3 (20 points in total).

Consider the logistic model for binary classification:

$$\log \frac{p(X)}{1-p(X)} \;=\; \beta_0 + \beta_1 X_1 + \beta_2 X_2, \qquad \text{where} \qquad p(X) = \Pr[Y = 1 \mid X].$$

Given $X = (x_1, x_2)$, we classify $\hat{Y} = 1$ if $p(x_1, x_2) \geq p^*$ and $\hat{Y} = 0$ otherwise. Suppose the estimated coefficients are:

$$\hat{\beta}_0 = -1.2, \qquad \hat{\beta}_1 = 2.0, \qquad \hat{\beta}_2 = -0.5.$$

(a) (5 points) You observe a new test point $x_{\text{test}} = (x_1, x_2) = (2, 3)$. Decide whether $\hat{y}_{\text{test}} = 1$ or $\hat{y}_{\text{test}} = 0$ at $p^* = 0.5$.

(b) (5 points) Derive the equation of the decision boundary for $p^* = 0.5$, and sketch this line in the (x_1, x_2) plane, indicating where $\hat{Y} = 1$ vs. $\hat{Y} = 0$.

(c) (10 points total) Using R, fit the logistic model `Label` $\sim X1 + X2$ on the following dataset of 100 points. For parts **(i)** and **(ii)**, use the fitted probabilities from this model on the same dataset, classify $\hat{Y} = 1$ when $\hat{p} \geq p^*$, and treat $Y = 1$ as the positive class.

```
set.seed(1)  # for reproducibility
X1_1 <- rnorm(40, mean = 2, sd = 1)
X2_1 <- rnorm(40, mean = 2, sd = 1)

# Generate (X1, X2) for label=0
X1_0 <- rnorm(60, mean = 0, sd = 1.2)
X2_0 <- rnorm(60, mean = 0, sd = 1.2)

# Combine into a single dataset
X1 <- c(X1_1, X1_0)
X2 <- c(X2_1, X2_0)
Y <- c(rep(1, 40), rep(0, 60))

# Create a data frame
df <- data.frame(X1, X2, Label)
```

(i) (5 points) Create a confusion matrix, and compute the true positive rate (TPR = sensitivity), and the false positive rate (FPR = 1 - specificity) at $p^* = 0.5$.

(ii) (5 points) For $p^* \in \{0, 0.05, 0.1, \ldots, 0.95, 1\}$, compute TPR and FPR, then plot the ROC curve (FPR, TPR).

Problem 4 (20 points total + 5 bonus points).

You catch crabs of two species, A and B, recording these weights (pounds):

$$\text{Species A: } x \in \{1.0,\ 2.0,\ 3.0\}, \quad \text{Species B: } x \in \{3.0,\ 4.0,\ 5.0,\ 6.0\}.$$

Assume:

- Species A: Gaussian with mean μ_A, variance σ^2.
- Species B: Gaussian with mean μ_B, variance σ^2.

For all parts below, use class prior probabilities estimated by the sample proportions:

$$\hat{\pi}_A = \frac{3}{7}, \qquad \hat{\pi}_B = \frac{4}{7}.$$

(a) (4 points) Compute the sample means $\bar{x}_A$, $\bar{x}_B$ and the pooled sample variance

$$s^2 = \frac{1}{n_A + n_B - 2}\left(\sum_{i \in A}(x_i - \bar{x}_A)^2 + \sum_{i \in B}(x_i - \bar{x}_B)^2\right).$$

(b) (4 points) Write the linear discriminant functions $\delta_A(x)$ and $\delta_B(x)$, using $\bar{x}_A$, $\bar{x}_B$, s^2 obtained above.

(c) (4 points) If $x_{\text{new}} = 3.2$, which species do you predict? Show your reasoning.

(d) (4 points) If you add three more data points for Species A, and then re-estimate the class priors using the new sample proportions while keeping the class mean and pooled variance roughly the same, would your prediction at $x_{\text{new}} = 3.2$ change? Explain briefly.

(e) (4 points) Suppose you do *not* want to miss any crabs of Species B. You decide to predict A only if $\Pr(Y = A \mid X = x) \geq p^*$ with $p^* > 0.5$ (e.g. $p^* = 0.9$). How does this modify the LDA decision rule in terms of $\delta_A(x)$ and $\delta_B(x)$? State your new decision boundary and apply it to $x_{\text{new}} = 3.2$.

(f*) (*5 bonus points) Suppose that Species A and B are not Gaussian-distributed but Laplace-distributed with class means μ_A, μ_B, a *common* scale parameter b, and the same class priors as above. The Laplace probability density function with mean μ and variance $2b^2$ is

$$f(x; \mu, b) = \frac{1}{2b}\exp\Big(-\frac{|\,x - \mu|}{b}\Big).$$

(i) Draw the graph of $f(x; 0, 1)$ and compare it with the graph of Gaussian density.

(ii) Derive the discriminant function under Laplace density and compare it with the linear discriminant function under Gaussian. (Hint: consider $\log(\pi_k f_k(x))$.)