STA 35C – Homework 2

Submission due: Tue, April 14 at 11:59 PM PT

Instructor: Dogyoon Song

Instructions: Upload a single PDF file to Gradescope via Canvas ("Homework 2" under "Assignments"). Name the file using the prefix of your UC Davis email ID and the homework number (e.g., `dgsong_hw2.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 (20 points in total).

A manufacturing company ships products in boxes of **two** items each. We define two random variables:

$$X = \text{number of defective items in a box of size 2,}$$
$$Y = \text{time (hours) for the final inspection of the box.}$$

(a) (5 points) Suppose each of the 2 items has a $\frac{1}{3}$ chance of being defective, independently. Then

$$X \sim \text{Binomial}(2, \tfrac{1}{3}).$$

Compute $\mathbb{E}[X]$ and $\text{Var}(X)$.
(*Hint:* use the PMF $p_X(x) = \binom{2}{x}\left(\frac{1}{3}\right)^x\left(\frac{2}{3}\right)^{2-x}$, or let $X = X_1 + X_2$ where X_1, X_2 are i.i.d. Bernoulli($\frac{1}{3}$).)

(b) (5 points) Consider the total cost

$$W = X + 2\,Y + 2,$$

measured in cost units, where each defective item contributes cost 1, inspection is charged at rate 2 per hour, and 2 is a fixed operating cost. In reality, more defects might delay inspection. Suppose $\mathbb{E}[Y] = \text{Var}(Y) = 9$, and the correlation coefficient $\rho := \text{corr}(X, Y) = 0.3$.

Compute $\mathbb{E}[W]$ and $\text{Var}(W)$. (*Hint:* $\text{Cov}(X, Y) = \rho\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}$.)

(c) (10 points total) Each box (with 2 items) is produced by **Factory A** or **Factory B**, each with *equal probability*.

- If it is from **Factory A**, the two items are defective independently, each with probability $\frac{1}{3}$.
- If it is from **Factory B**, the two items are defective independently, each with probability $\frac{1}{10}$.

(i) (5 points) What is the probability that a randomly chosen box is from Factory A *and* has exactly one defective item ($X = 1$)?

(ii) (5 points) You observe $X = 1$ defective item in a newly received box. What is the posterior probability that this box came from **Factory A**?

Problem 2 (25 points in total).

Respond to each part in 2–4 sentences in your own words. When asked for an example, one well-chosen example is enough.

(a) **(5 points)** Explain supervised and unsupervised learning, highlighting what information is available in each setting and what each is used for. Give one example of each.

(b) **(5 points)** Explain regression and classification, highlighting their similarities and differences, including what type of response variable each involves. Give one example of each.

(c) **(5 points)** Explain prediction and inference, and give one example scenario/task for each.

(d) **(5 points)** Explain parametric and non-parametric methods, including their differences as well as one advantage and one disadvantage of each.

(e) **(5 points)** Explain what overfitting is and how it relates to the bias–variance tradeoff.

Problem 3 (30 points in total + 5 bonus points).

Consider the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, \dots, n,$$

where the errors $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. $\mathcal{N}(0, \sigma^2)$.

(a) (7 points) Assuming $\sum_{i=1}^n (x_i - \bar{x})^2 > 0$, prove that the minimizer of

$$\text{RSS} = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

is given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$.

(b) (8 points) Suppose we are given a dataset of (x, y) as follows:

$$\{(-3,-4), (-2,-2), (1,0), (2,3), (3,5), (4,6), (5,8)\}.$$

(i) Compute $\hat{\beta}_1$, $\text{SE}(\hat{\beta}_1)$, and a 95% confidence interval for β_1.

(ii) Compute the t-statistic for testing the null hypothesis $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$ at the 5% significance level, and decide whether to reject. How would you interpret the result?

(c) (15 points) Implement a short R script (using a fixed seed, e.g. `set.seed(2026)`) that:

(i) Generates $n = 100$ i.i.d. datapoints (X_i, Y_i), $i = 1, \dots, n$, via

$$Y_i = 2 + 3X_i + \varepsilon_i \qquad \text{where} \qquad X_i \sim \text{Uniform}(0, 10), \qquad \varepsilon_i \sim \mathcal{N}(0, 2^2).$$

(ii) Fits the linear model and reports $\hat{\beta}_0, \hat{\beta}_1$.

(iii) Computes the residual standard error (RSE) and compares it to the true $\sigma = 2$.

(iv) Constructs 95% confidence intervals for β_0 and β_1.

(v) Performs a two-sided hypothesis test of $H_0 : \beta_1 = 0$ at the 5% level.

(vi) Computes R^2 and interprets it.

Report the results you obtain. Repeat the entire experiment for a larger sample size $n = 1000$ (instead of $n = 100$) and discuss any differences you observe.

(d*) (5 bonus points) Repeat (**(c)**) but simulate $Y = 2 + 3X^2 + \varepsilon$. Fit a linear model anyway and compare results with the true quadratic form. Comment on the obtained results ($\hat{\beta}_0, \hat{\beta}_1$, R^2, etc.).

Problem 4 (25 points in total + 5 bonus points).

We collect $n = 100$ observations (x_i, y_i) and consider two regression models:

- A simple linear regression:
$$Y = \beta_0 + \beta_1 X + \varepsilon.$$
- A cubic polynomial regression:
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon.$$

(a) (5 points) Explain why the cubic model above can be viewed as a *multiple linear regression* model. In particular, define
$$Z_1 = X, \qquad Z_2 = X^2, \qquad Z_3 = X^3,$$
rewrite the cubic model using Z_1, Z_2, Z_3, and explain what is linear and what is nonlinear.

(b) (5 points) Compare the training RSS for the simple linear regression and the cubic regression fitted on the same training data. Which one must be smaller, or is there insufficient information to tell (so either can be smaller)? Justify your answer.

(c) (5 points) Now suppose the *true* relationship between X and Y is linear. Compare the *expected test RSS*, instead of the training RSS, of the simple linear and cubic models. Which would you typically expect to be smaller, and why?

(d) (10 points) Implement a short R script to do the following. To reduce simulation noise, repeat each experiment 100 times; in each repetition, generate a fresh i.i.d. training data (X_i, Y_i), $i = 1, \ldots, n$, from the stated model, and an independent test set of size 100 from the same model, then record the training and test RSS for both models.

(i) Generate data from
$$Y = 5 + 2X + \varepsilon, \qquad \text{where} \qquad X \sim \text{Uniform}(0, 10), \qquad \varepsilon \sim \mathcal{N}(0, 2^2).$$
Fit both models, and compare their *average* training RSS and *average* test RSS.

(ii) Generate data from
$$Y = 0.5X^2 + 2\sin X + \varepsilon, \qquad \text{where} \qquad X \sim \text{Uniform}(0, 10), \qquad \varepsilon \sim \mathcal{N}(0, 2^2).$$
Fit both models, compare their *average* training RSS and *average* test RSS, and discuss.

Finally, repeat the whole experiment after reducing the training sample size from 100 to 5, and then discuss any differences you observe.

(e*) (5 bonus points) Relate your findings from **(d)** to your answers in **(b)** and **(c)**. If the experimental results match your earlier reasoning, discuss why. If they do not match, speculate on possible reasons.