STA 35C – Homework 1

Submission due: Tue, April 15 at 11:59 PM PT

Instructor: Dogyoon Song

Instructions: Upload a PDF file, named with your UC Davis email ID and homework number (e.g., `dgsong_hw1.pdf`) to Gradescope (accessible through Canvas). Please make sure to include "STA 35C," your name, and the last four digits of your student ID on the front page. For more details on submission requirements and the late submission policy, see the syllabus.

Problem 1 (20 pt).

(a) Let X take values in $\{1, 2, 3\}$ with probabilities proportional to

$$P(X = k) \propto \theta^k, \quad k = 1, 2, 3,$$

for some constant $\theta > 0$.

(i) Find the normalizing constant so that $P(X = k)$ is a valid probability distribution, and write $p_X(k)$ in closed form.

(ii) Using your expression for $p_X(k)$, compute the cumulative distribution function $F_X(k)$.

(iii) Compute $\mathbb{E}[X]$ and $\mathrm{Var}(X)$.

(b) Let $\alpha > 0$, and suppose that Y is a continuous random variable with density

$$f_Y(y) = c\, e^{-\alpha y}, \quad y \geq 0,$$

for some constant $c > 0$.

(i) Find the value of c so that $f_Y(y)$ integrates to 1 over $[0, \infty)$.

(ii) Derive $F_y(y)$ from $f_Y(y)$.

(iii) Compute $\mathbb{E}[Y]$ and $\mathrm{Var}(Y)$.

(c) Compute $\mathbb{E}[\,2X + Y\,]$ and $\mathrm{Var}(2X + Y)$ under each of the following assumptions:

(i) X and Y are independent.

(ii) The correlation coefficient between X and Y is 0.4.

Problem 2 (25 pt).

Suppose that we have a coin with unknown head probability $p_{\text{true}} = \Pr(\text{Head}) \in [0, 1]$, which *we believe* can only take one of three possible values:

$$p \in \{0.2,\ 0.5,\ 0.8\}.$$

Let us denote this model parameter by θ, and place a *uniform prior* on these three values:

$$P(\theta = 0.2) \ = \ P(\theta = 0.5) \ = \ P(\theta = 0.8) \ = \ \frac{1}{3}.$$

This means we have no prior information about θ, and subjectively believe each of the three values are equally probable.

(a) Suppose we flip the coin once and observe $X \in \{0, 1\}$, where 1 denotes a Head and 0 denotes a Tail. Use the Law of Total Probability to compute the marginal probability of observing a head:

$$P(X = 1) \ = \sum_{k \in \{0.2,\, 0.5,\, 0.8\}} P(X = 1 \mid \theta = k)\, P(\theta = k).$$

Then compute $P(X = 0)$ similarly. Give numerical values with the uniform prior.

(b) Derive the *posterior* probability for each $\theta \in \{0.2, 0.5, 0.8\}$ if the observed flip is a head ($X = 1$):

$$\underbrace{P(\theta = k \ \mid \ X = 1)}_{\text{posterior}} \ = \ \frac{P(X = 1 \mid \theta = k)\ \overbrace{P(\theta = k)}^{\text{prior}}}{P(X = 1)}.$$

Compute these three posterior probabilities for $k = 0.2$, 0.5, 0.8. Then do the same for $X = 0$. Check that each set of posteriors sums to 1.

(c) Now consider a *stream of n flips* from a coin whose *true* probability might be $p_{\text{true}} = 0.3$ or $p_{\text{true}} = 0.5$. Implement the following in a R script, and report the resulting plots for $p_{\text{true}} = 0.3$ and $p_{\text{true}} = 0.5$.

(i) Generate n independent flips from Bernoulli(p_{true}).

(ii) Initialize your prior to $(1/3,\ 1/3,\ 1/3)$ and update it *sequentially* after each flip. Specifically, for a head ($X = 1$):

$$\begin{aligned}
\text{Posterior}(0.2) &\ \propto\ 0.2 \times \text{Prior}(0.2),\\
\text{Posterior}(0.5) &\ \propto\ 0.5 \times \text{Prior}(0.5),\\
\text{Posterior}(0.8) &\ \propto\ 0.8 \times \text{Prior}(0.8),
\end{aligned}$$

and similarly for a tail with $(1 - p)$. Then normalize so these three probabilities sum to 1. Use the posterior after the i-th flip as the prior for the $(i + 1)$-th flip.

(iii) Store and plot how Posterior(θ) changes over time for each θ.

Problem 3 (30pt).

Suppose a model $Y = \beta_0 + \beta_1 X + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

(a) Prove that the minimizer of the RSS $\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$ takes the form

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

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) Given a dataset

$$(x, y) \in \{(1,0), (2,3), (3,5), (4,6), (5,8)\},$$

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

(ii) Compute the t-statistic for $H_0 : \beta_1 = 0$ and decide whether to reject. How would you interpret the result?

(c) Implement a short R script that (1) generates $n = 50$ points via

$$Y = 2 + 3X + \varepsilon, \quad X \sim \text{Uniform}(0, 10), \quad \varepsilon \sim \mathcal{N}(0, 2^2),$$

and fits the linear model and report $\hat{\beta}_0$, $\hat{\beta}_1$. Then report your answers to the following questions:

(i) Compute the residual standard error (RSE). Compare it to the true $\sigma = 2$.

(ii) Construct confidence intervals for β_0 and β_1.

(iii) Perform a hypothesis test for $H : \beta_1 = 0$.

(iv) Compute R^2 and interpret.

(v) Repeat for $n = 500$ or $n = 1000$; discuss any differences.

(d) 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 (25pt).

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

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

(a) Suppose the *true* relationship between X and Y is actually linear. Compare the training RSS for the linear regression and for the cubic regression. Which do we expect to be lower, or is there insufficient information to tell? Explain.

(b) Answer (**(a)**) again, but using the *test* RSS instead of the training RSS.

(c) Implement a short R script to address the following:

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

(ii) Generate data from $Y = 0.5X^2 + 2 \sin X + \varepsilon$, where $X \sim \text{Uniform}(0, 10)$ and $\varepsilon \sim \mathcal{N}(0, 2^2)$. Fit both models, compare the results, and discuss.

Relate your findings to your answers in **(a)** and **(b)**. Finally, comment on any differences that arise if you reduce the training sample size to $n = 5$ or $n = 10$.