Homework 7: Mathematical Statistics (MATH-UA 234)

Due 12/08 at the beginning of class on Gradescope. The quiz will still be 12/06, and will cover content from problems 1-3 (i.e. Bayesian inference). No solutions will be posted prior to the quiz.

Reminder. Remember than the project presentations are on December 14th!

Problem 1. *Suppose $X_1, \dots, X_n \sim \text{Ber}(p)$ (with 1 representing heads and zero representing tails) and that we use the prior distribution $p \sim \text{Beta}(\alpha, \beta)$.*

(a) *Compute the posteriori distribution for $p | X_1 = x_1, \dots, X_n = x_n$.*

(b) *For each of the coins below, find values of α and β so that your prior distribution represents your belief about the parameter p of the coin. Plot and label these 6 prior distributions. Note that the head side is the side marked with the number.*

(c) *Suppose you flipped coin zero and got 53 heads and 47 tails. Make a plot showing the prior and posterior densities for p.*

(d) *Suppose you flipped coin 4 and got 39 heads and 61 tails. Make a plot showing the prior and the posterior densities for p.*

(e) *Suppose you flipped coin 6 and got 0 heads and 100 tails. Make a plot showing the prior and the posterior densities for p.*

(f) *For the coin 6 example, is the probability that $p = 0$ under your posterior 100%? Does this make sense? Why or why not?*

This image was taken from this site: `https://izbicki.me/blog/how-to-create-an-unfair-coin-and-prove-it-with-math.html`

Solution.

(a) This is almost identical to the book problem, and the posterior is $\text{Beta}(\alpha + s, \beta + n - s)$ where $s = x_1 + \cdots + x_n$.

(b) The exact values of α and β you pick are not that important. Indeed, in Bayesian statistics the prior comes down to belief. However, the prior should have some basic properties. For instance, coin zero should probably be symmetric about $p = 1/2$, coin 6 should heavily favor tails, etc.

problems with a textbook reference are based on, but not identical to, the given reference

(c)

(d)

(e)

(f) The probability is zero unless your prior had $\mathbb{P}[p = 0] = 1$. This makes sense because it could have been chance that we got 100 tails. This is particularly true if $p \approx 0$.

Problem 2 (Wasserman 11.1). *Suppose $X_1, \dots, X_n \sim N(\theta, \sigma^2)$, and that we use the prior distribution $\theta \sim N(a, b^2)$. Show that $\theta | X_1 = x_1, \dots, X_n = x_n \sim N(\bar{\theta}, \tau^2)$ where*

$$\bar{\theta} = w\frac{x_1 + \cdots + x_n}{n} + (1-w)a, \qquad w = \frac{1/\mathsf{se}^2}{1/\mathsf{se}^2 + 1/b^2}, \qquad \tau = 1/\sqrt{1/\mathsf{se}^2 + 1/b^2}, \qquad \mathsf{se} = \sigma/\sqrt{n}.$$

Solution. The prior is

$$f_\Theta(\theta) \propto \exp\left(-\frac{1}{2}\left(\frac{\theta - a}{b}\right)^2\right)$$

and the likilehiid function is

$$L_n(\theta) = \prod_{i=1}^n f_{X_i|\Theta=\theta} \propto \prod_{i=1}^n \exp\left(-\frac{1}{2}\left(\frac{X_i - \theta}{\sigma}\right)^2\right).$$

Thus, the posterior is proportional to

$$\exp\left(-\frac{1}{2}\left(\frac{\theta - a}{b}\right)^2\right)\prod_{i=1}^n \exp\left(-\frac{1}{2}\left(\frac{x_i - \theta}{\sigma}\right)^2\right) = \exp\left(-\frac{1}{2}\left[\left(\frac{\theta - a}{b}\right)^2 + \sum_{i=1}^n \left(\frac{x_i - \theta}{\sigma}\right)^2\right]\right).$$

The problem also gives us that the posterior is proportional to

$$\exp\left(-\frac{1}{2}\left(\frac{\theta - \bar{\theta}}{\tau}\right)^2\right).$$

Thus, we just need to match these up. In particular, for some c, we have

$$\exp\left(-\frac{1}{2}\left[\left(\frac{\theta - a}{b}\right)^2 + \sum_{i=1}^n \left(\frac{x_i - \theta}{\sigma}\right)^2\right]\right) = c\exp\left(-\frac{1}{2}\left(\frac{\theta - \bar{\theta}}{\tau}\right)^2\right).$$

Thus,

$$\left(\frac{\theta - a}{b}\right)^2 + \sum_{i=1}^n \left(\frac{x_i - \theta}{\sigma}\right)^2 = -2\ln(c) + \left(\frac{\theta - \bar{\theta}}{\tau}\right)^2.$$

Quadratics are equal if and only if each of their coefficients are equal. The θ^2 term gives

$$\frac{1}{b^2} + \sum_{i=1}^n \frac{1}{\sigma^2} = \frac{1}{\tau^2}$$

so $\tau = 1/\sqrt{1/\mathsf{se}^2 + 1/b^2}$ where $\mathsf{se} = \sigma/\sqrt{n}$. The θ term gives

$$-\frac{2a}{b^2} + \sum_{i=1}^n -\frac{2x_i}{\sigma^2} = -\frac{2\bar{\theta}}{\tau^2}$$

so $\bar{\theta} = \tau^2(a/b^2 + (x_1 + \cdots + x_n)/\sigma^2)$.

Problem 3 (Wasserman 11.2)**.** *Let $X_1, \dots, X_n \sim N(\mu, 1)$.*

(a) *Simulate a dataset (using $\mu = 5$) consisting of $n = 100$ observations*

(b) *Take $f(\mu) = 1$ as the prior density, and find the posterior density given the observed data. Plot this density*

Problem 4. *Consider a model of the form $f(x) = \hat{\beta}_0 + \hat{\beta}_1 x$ and, given data*

$$(X_1, Y_1), \dots, (X_n, Y_n),$$

define the loss function

$$L(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} (Y_i - f(X_i))^2.$$

(a) *Compute the partial derivatives $\partial L(\hat{\beta}_0, \hat{\beta}_1)/\partial \hat{\beta}_0$ and $\partial L(\hat{\beta}_0, \hat{\beta}_1)/\partial \hat{\beta}_1$*

(b) *Find the minimizers $\hat{\beta}_0$ and $\hat{\beta}_1$ for $L(\hat{\beta}_0, \hat{\beta}_1)$.*

(c) *Show that you can write the loss function in the form $\|\vec{b} - \vec{A}\vec{x}\|_2^2$, where $\vec{b}$ is a particular vector of length n, $\vec{A}$ is a $n \times 2$ matrix, and $\vec{x}$ is a length 2 vector.*

Solution.

(a) We have

$$L(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2.$$

Therefore,

$$\frac{\partial}{\partial \hat{\beta}_0} L(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} \frac{\partial}{\partial \hat{\beta}_0} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = \sum_{i=1}^{n} (-2)(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = -2n(\bar{Y}_n - \hat{\beta}_0 - \hat{\beta}_1 \bar{X}_n).$$

and

$$\frac{\partial}{\partial \hat{\beta}_1} L(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} \frac{\partial}{\partial \hat{\beta}_1} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = \sum_{i=1}^{n} (-2X_i)(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i).$$

(b) Setting the first equation to zero clearly gives $\hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_1 \bar{X}_n$.
Plugging this into the second equation we find

$$0 = \sum_{i=1}^{n} (-2X_i)(Y_i - (\bar{Y}_n - \hat{\beta}_1 \bar{X}_n) - \hat{\beta}_1 X_i) = \sum_{i=1}^{n} (-2X_i)(Y_i - \bar{Y}_n - \hat{\beta}_1 (X_i - \bar{X}_n)).$$

so

$$\sum_{i=1}^{n} X_i (Y_i - \bar{Y}_n) = \sum_{i=1}^{n} X_i \hat{\beta}_1 (X_i - \bar{X}_n)).$$

which gives

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i (Y_i - \bar{Y}_n)}{\sum_{i=1}^{n} X_i (X_i - \bar{X}_n)}.$$

Note that

$$\sum_{i=1}^{n} \bar{X}_n (Y_i - \bar{Y}_n) = n\bar{X}_n \bar{Y}_n - n\bar{X}_n \bar{Y}_n = 0, \qquad \sum_{i=1}^{n} \bar{X}_n (X_n - \bar{X}_n) = n\bar{X}_n \bar{X}_n - n\bar{X}_n \bar{X}_n = 0.$$

Thus, we also have

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n X_i(Y_i - \bar{Y}_n) - \sum_{i=1}^n \bar{X}_n(Y_i - \bar{Y}_n)}{\sum_{i=1}^n X_i(X_i - \bar{X}_n) - \sum_{i=1}^n \bar{X}_n(X_i - \bar{X}_n)} = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sum_{i=1}^n (X_i - \bar{X}_n)(X_i - \bar{X}_n)}$$

which is the formula in the book.

Problem 5. *Consider the following four data sets:*

```
x1 = [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5]
y1 = [8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68]

x2 = [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5]
y2 = [9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26, 4.74]

x3 = [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5]
y3 = [7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73]

x4 = [8, 8, 8, 8, 8, 8, 8, 19, 8, 8, 8]
y4 = [6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 5.56, 7.91, 6.89]
```

(a) *Find the sample mean and sample variance of each datasets' X and Y values. Compute the sample correlation between the X and Y values for each dataset.*

(b) *Find the linear regression line and compute the R^2 value for each dataset.*

(c) *Now, plot the datasets and the linear regression lines. Explain what happened.*

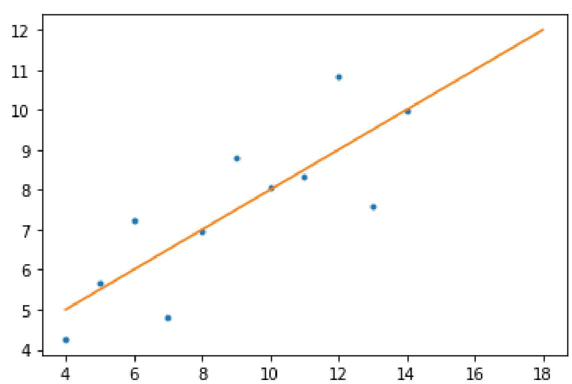
Solution.

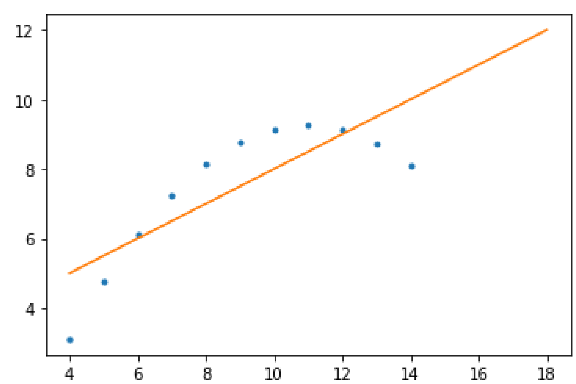
(a) Each of the datasets have $\mu_X = 9$, $\mu_Y = 7.5$, $\sigma_X^2 = 10$, $\sigma_Y^2 = 3.75$, and $\sigma_{X,Y}^2 = 7.79$. The exact values you get will depend on if you use the formula for the sample mean and variance using $n-1$ or n, but no matter what they will be very close for all of the datasets.

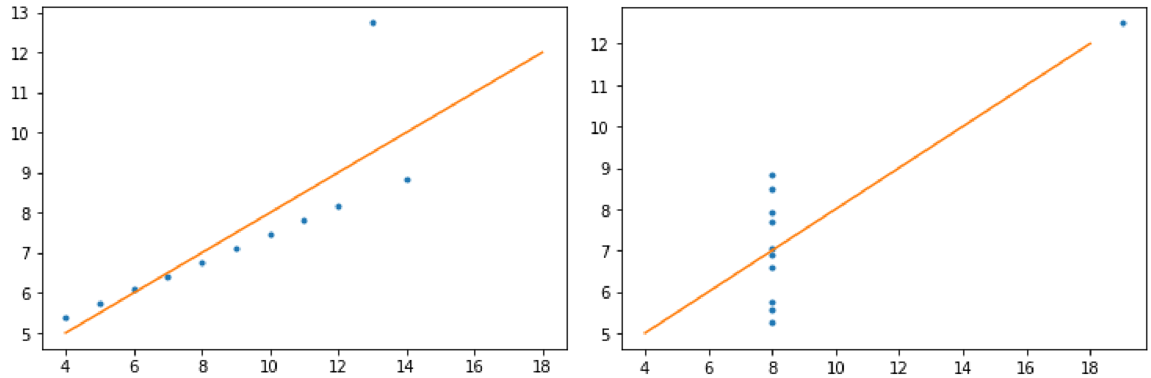
(b) All of the data have a regression line of roughly $y = 3 + 0.5x$ and a R^2 values of roughly 0.66.

(c) The data all look very different. While the statistics are the same, the second dataset looks like a quadratic, the third like a line with an outlier, and the fourth like a single outlying point.

Without plotting the data, we might not have realize such patterns were there.







Problem 6 (Wasserman 13.2). *Suppose $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, where $\mathbb{E}[\epsilon_i | X_i] = 0$ and $\mathbb{V}[\epsilon_i | X_i] = \sigma^2$.*

Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be the least squares estimates given in Theorem 13.4. Show that $\mathbb{E}[\hat{\beta}_0 | X_1, \dots, X_n] = \beta_0$ and $\mathbb{E}[\hat{\beta}_1 | X_1, \dots, X_n] = \beta_1$. You should regard $X_1, \dots, X_n$ as constant.

Solution. Recall that

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sum_{i=1}^n (X_i - \bar{X}_n)^2}$$

Since $\mathbb{E}[Y_i] = \beta_0 + \beta_1 X_i$, $\mathbb{E}[\bar{Y}_n] = \beta_0 + \beta_1 \bar{X}_n$ and $\mathbb{E}[Y_i - \bar{Y}_n] = \beta_1 (X_i - \bar{X}_n)$. Therefore,

$$\mathbb{E}[\hat{\beta}_1] = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)\mathbb{E}[Y_i - \bar{Y}_n]}{\sum_{i=1}^n (X_i - \bar{X}_n)^2} = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)\beta_1 (X_i - \bar{X}_n)}{\sum_{i=1}^n (X_i - \bar{X}_n)^2} = \beta_1.$$

This implies $\mathbb{E}[\hat{\beta}_0] = \mathbb{E}[\bar{Y}_n] - \mathbb{E}[\hat{\beta}_1]\bar{X}_n = \beta_0 + \beta_1 \bar{X}_n - \beta_1 \bar{X}_n = \beta_0$.

Problem 7. *Pick at least one of the following articles to read. Provide a one paragraph summary of what you think the most important points of the article were. Discuss how this is relevant to what we are learning in class.*

- *Why algorithms can be racist and sexist*
- *All the Ways Hiring Algorithms Can Introduce Bias*
- *Racial Discrimination in Face Recognition Technology*