Stat 200A, Fall 2018
A. Adhikari and Z. Bartha WEEK 13 EXERCISES

You are expected to do all these problems, but for **Homework 13** please turn in **only Problems 1, 3, 6, and 8** on **Thursday November 29 at the start of lecture**.

1. Regression Planes

Consider three random variables Y, X_1, and X_2 The goal of this exercise is to find the least squares linear predictor of Y based on the predictor variables X_1 and X_2.

For algebraic simplicity, let all the variables be measured in standard units. Then the random vector $\mathbf{W} = [Y \ X_1 \ X_2]^T$ has mean vector $\mathbf{0}$ and covariance matrix

$$\begin{bmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_3 \\ \rho_2 & \rho_3 & 1 \end{bmatrix}$$

where $\rho_1 = Cov(Y, X_1)$, $\rho_2 = Cov(Y, X_2)$, and $\rho_3 = Cov(X_1, X_2)$ are correlations.

(a) Suppose the two predictor variables are uncorrelated. Find the best linear predictor of Y based on X_1 and X_2.

With the usual notation, the best linear predictor is

$$\hat{Y} = \Sigma_{Y,X}\Sigma_X^{-1}(X_1, X_2)^T = \rho_1 X_1 + \rho_2 X_2.$$

(b) Suppose now that $-1 < \rho_3 < 1$. Find the best linear predictor of Y based on X_1 and X_2. You will have to recall the inverse of a 2×2 matrix.

$$\begin{aligned} \hat{Y} = \Sigma_{Y,X}\Sigma_X^{-1}(X_1, X_2)^T &= (\rho_1, \rho_2) \begin{pmatrix} 1 & \rho_3 \\ \rho_3 & 1 \end{pmatrix}^{-1} (X_1, X_2)^T \\ &= \frac{1}{1-\rho_3^2}(\rho_1, \rho_2) \begin{pmatrix} 1 & -\rho_3 \\ -\rho_3 & 1 \end{pmatrix} (X_1, X_2)^T \\ &= \frac{\rho_1 - \rho_2\rho_3}{1-\rho_3^2} \cdot X_1 + \frac{\rho_2 - \rho_1\rho_3}{1-\rho_3^2} \cdot X_2 \end{aligned}$$

(c) Find the mean squared error of the predictor in Part (b).

The mean squared error of $\hat{Y}$ is

$$\sigma_Y^2 - \Sigma_{Y,X}\Sigma_X^{-1}\Sigma_{X,Y} = 1 - \frac{1}{1-\rho_3^2}(\rho_1, \rho_2) \begin{pmatrix} 1 & -\rho_3 \\ -\rho_3 & 1 \end{pmatrix} (\rho_1, \rho_2)^T = \frac{1 - \rho_1^2 - \rho_2^2 - \rho_3^2 + 2\rho_1\rho_2\rho_3}{1-\rho_3^2}$$

(d) Suppose $\rho_3 = 1$. What is the relation between X_1 and X_2? Could you use the predictor you developed in Part (b)? If not, how would you use linear regression to predict Y based on the predictor variables?

If $\rho_3 = 1$, then the formula in part (b) does not apply, since Σ_X is not invertible. In fact, if $\rho_3 = 1$, then $X_2 = aX_1 + b$ for some constants $a > 0$ and b (with probability 1). But since X_1 and X_2 have the same

mean and variance, this relation has to be $X_2 = X_1$. Hence, predicting Y based on X_1 and X_2 is the same as predicting it based on just X_1, and in this case we use the simple linear regression:

$$\hat{Y} = \frac{Cov(Y, X_1)}{Var(X_1)} \cdot X_1 = \rho_1 X_1.$$

2. Multiple Regression Model

Consider the multiple regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\mathbf{Y}$ is an $n \times 1$ random vector, $\mathbf{X}$ is an $n \times p$ matrix of known constants, $\boldsymbol{\beta} = [\beta_0 \ \ \beta_1 \ \ \cdots \ \ \beta_{p-1}]^T$ is a $p \times 1$ vector of unknown coefficients, and $\boldsymbol{\epsilon}$ has the multivariate normal $(\mathbf{0}, \sigma^2\mathbf{I}_n)$ distribution for some $\sigma^2 > 0$ and $\mathbf{I}_n$ the $n \times n$ identity matrix.

As in class, assume that $n > p$, the first column of $\mathbf{X}$ is all 1's, and the rank of $\mathbf{X}$ is p.

(a) What is the distribution of $\mathbf{Y}$?

Since $\mathbf{X}\boldsymbol{\beta}$ is a constant vector, $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is multivariate normal $(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$.

(b) Let $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$. What is the distribution of $\bar{Y}$?

Normal $\left(\frac{1}{n}\sum_{i=1}^n (\mathbf{X}\boldsymbol{\beta})_i, \frac{\sigma^2}{n}\right)$.

3. Properties of Regression Estimates

Continue with the model in Exercise 2. As in class, let $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$.

(a) What is the distribution of $\hat{\boldsymbol{\beta}}$?

Using that $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ from 2 (a), $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ is multivariate normal with mean vector

$$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

and covariance matrix

$$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\sigma^2\mathbf{I}_n((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}.$$

(b) Let $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$. What is the distribution of $\hat{\mathbf{Y}}$? Compare with the answer to Exercise 2(a).

By part (a), $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ is multivariate normal with mean vector $\mathbf{X}\boldsymbol{\beta}$ and covariance matrix $\sigma^2\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$.

(c) Let $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$. True or false: $\mathbf{Y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{e}$. Compare with the regression model.

True: $\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{e} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{Y}$.

Hence, if we replace the true coefficient vector $\boldsymbol{\beta}$ with the estimated one $\hat{\boldsymbol{\beta}}$ in the equation $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, then the error of the estimation $\mathbf{e}$ will play the role of the noise.

(d) Classify each of the following as observable or unobservable, and also as random vector, random matrix, constant vector, or constant matrix: $\mathbf{Y}$ $\mathbf{X}$ $\boldsymbol{\beta}$ $\hat{\boldsymbol{\beta}}$ $\boldsymbol{\epsilon}$ $\mathbf{e}$

$\mathbf{Y}$: observable random vector

$\mathbf{X}$: observable contant matrix

$\boldsymbol{\beta}$: unobservable constant vector

$\hat{\boldsymbol{\beta}}$: observable random vector

$\boldsymbol{\epsilon}$: unobservable random vector

$\mathbf{e}$: observable random vector

4. Back to Simple Regression

For $1 \leq i \leq n$, let $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ where the ϵ_i are i.i.d. normal $(0, \sigma^2)$ errors. In terms of the model in Exercise 2, this is $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

Recall from class, or work out from $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$, that

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} \qquad \hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

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

(a) Show that the regression line $b(x) = \hat{\beta}_0 + \hat{\beta}_1 x$ passes through $(\bar{x}, \bar{Y})$.

$b(\bar{x}) = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} = \bar{Y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 \bar{x} = \bar{Y}$

(b) Show that

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} - \bar{Y}\frac{\sum_{i=1}^n (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

since $\sum_{i=1}^n (x_i - \bar{x}) = 0$.

(c) Use (b) to show that $Cov(\hat{\beta}_1, \bar{Y}) = 0$. Explain geometrically what the result means.

$$Cov(\hat{\beta}_1, \bar{Y}) = Cov\left(\frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}, \frac{\sum_{i=1}^n Y_i}{n}\right) = \frac{\sum_{i=1}^n (x_i - \bar{x}) Cov(Y_i, Y_i)}{n\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sigma^2 \sum_{i=1}^n (x_i - \bar{x})}{n\sum_{i=1}^n (x_i - \bar{x})^2} = 0,$$

using that $Cov(Y_i, Y_j) = 0$ if $i \neq j$, and $Cov(Y_i, Y_i) = Var(Y_i) = \sigma^2$ for every i.

Geometrically, it means that a shift of the data points (x_i, Y_i) along the Y axis doesn't change the slope of the regression line.

(d) Find $Var(\hat{\beta}_1)$ and then $Var(\hat{\beta}_0)$.

$$Var(\hat{\beta}_1) = Var\left(\frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 Var(Y_i)}{\left[\sum_{i=1}^n (x_i - \bar{x})^2\right]^2} = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$Var(\hat{\beta}_0) = Var(\bar{Y} - \hat{\beta}_1 \bar{x}) = Var(\bar{Y}) + \bar{x}^2 Var(\hat{\beta}_1) - 2\bar{x} Cov(\bar{Y}, \hat{\beta}_1) = \frac{\sigma^2}{n} + \frac{\sigma^2 \bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

(e) Find $Cov(\hat{\beta}_0, \hat{\beta}_1)$. Explain why the sign makes sense. Keep in mind that $\bar{x}$ can be negative.

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = Cov(\bar{Y} - \hat{\beta}_1\bar{x}, \hat{\beta}_1) = Cov(\bar{Y}, \hat{\beta}_1) - Cov(\hat{\beta}_1\bar{x}, \hat{\beta}_1) = -\bar{x}Var(\hat{\beta}_1) = \frac{-\bar{x}\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

The sign makes sense, since the regression line has to pass through $(\bar{x}, \bar{Y})$. This means that, in the $\bar{x} > 0$ case, as we vary the Y_i's in a way that keeps $\bar{Y}$ fixed and increases the slope, the intercept of the line has to decrease. The reverse holds in the $\bar{x} < 0$ case.

5. Switching Chain

Consider a Markov Chain $X_0, X_1, \ldots$ with the transition matrix given below, for some $0 < p < 1$ and $q = 1 - p$.

	0	**1**
0	p	q
1	q	p

(a) For $n \geq 1$, let C_n be the number of *switches* up to time n. That is, C_n is the number of times the chain changes state up to and including time n. For example, if the path is 0 0 0 1 0 0 0 1 1, then $C_8 = 3$ (remember that the path starts at X_0). What is the distribution of C_n, and why?

(b) For brevity, let $P_n(0,0) = P(X_n = 0 \mid X_0 = 0)$. Fill in the blank with a word:

For $n \geq 1$,

$$P_n(0,0) \;=\; P(C_n \text{ is } \_\_\_\_\_ \mid X_0 = 0)$$

(c) Now find $P_n(0,0)$ using Part (b).

[Hint: Compare the expansions of $(p+q)^n$ and $(p-q)^n$. How can you use both of them to get just the terms that you need?]

(d) Use Part (c) (not the balance equations) to find the stationary distribution of the chain.

(e) In the long run, what proportion of time is the chain expected to spend at 0?

6. Switching Particles

For N a fixed positive integer, N blue particles and N red particles are distributed in two containers so that there are N particles in each container. A process evolves as follows.

At every step, one particle is selected uniformly at random from each container, independently of the other choice; then these two particles are switched. That is, each chosen particle is taken out of its container and placed in the other one.

For $n \geq 0$, let X_n be the number of blue particles in Container 1.

(a) What is the state space of the chain $\{X_0, X_1, X_2, \ldots\}$? Find the one-step transition probabilities.

(b) Are the detailed balance equations satisfied? Why or why not?

(c) Find the stationary distribution of the chain. Recognize this as one of the famous ones and provide its

name and parameters. It might help to remember that $\binom{n}{k} = \binom{n}{n-k}$.

7. Wet Professor

I don't know why this problem has become a classic. But versions of it appear in many texts and exercise sets with diverse people getting wet.

It's essentially an exercise in setting up a useful chain, based on which you can easily find the desired proportion.

A professor has two umbrellas, each of which could either be in her office or in her car. The professor walks from her car to her office; she also walks from her office to her car. Assume that on each of these walks:

- It rains with probability 0.7, independently of all other walks.
- If it is not raining, the professor ignores the umbrellas.
- If it is raining, she uses an umbrella if there is one, and gets wet if there isn't.

In the long run, what is the expected proportion of walks on which the professor gets wet?

8. "Move to Front" Permutations

Consider an alphabet of length N. A *move-to-front* permutation of the N letters consists of picking one of the letters (randomly or otherwise) and moving it to the front of the list. For example, if the alphabet consists of the three letters A, B, and C, and you start with the permutation ABC, then CAB is the result of one move-to-front (the chosen letter is C), as is ABC (the chosen letter is A), but not CBA.

(a) As a preliminary, recall that all transition matrices are stochastic, that is, each row sums to 1. Suppose the transition matrix of a finite-state irreducible aperiodic chain is *doubly stochastic*, that is, each of its columns also sums to 1. Explain why the stationary distribution must be uniform.

(b) A standard deck consists of 52 cards. A "random to front" shuffle is defined as follows: Pick one of the 52 cards uniformly at random and move it to the front of the deck (which you are welcome to think of as the top of the deck, if you prefer). Explain why if you perform this move over and over again, in the long run the deck will become well shuffled; that is, all permutations will be equally likely. [Set up an appropriate chain and use Part (a). You might want to try it out first with an alphabet of just the letters A, B, and C.]

(c) The *Tsetlin Library* imagines that a finite number of books are being permuted according to a move-to-front scheme in which the book to be moved is chosen not uniformly but based on its "importance" or "popularity". More popular books are therefore more likely to be selected and moved to the front of the row, that is, in the leftmost position. Applications of this model to computer science include dynamic file management and cache management.

Consider a library of just three books: A, B, and C. Let the "weight" or "importance" of each book be reflected in the probability with which it is selected to be moved to the left. Call these probabilities p_A, p_B, and p_C and note that $p_A + p_B + p_C = 1$. Set up the Tsetlin Library as a Markov chain on the space of permutations of the books. Explain why the chain is aperiodic.

All subsequent parts are about the chain in Part (c).

(d) Set up all the balance equations. Use the equations to find the expected long run proportion of time when A is the leftmost book. Your answer should be in terms of the weights. Explain why the answer makes sense.

(e) Find the expected long run proportion of time that B is the leftmost book. Also find the expected long run proportion of time that C is the leftmost book.

(f) Now find the ratio of the following quantities, in terms of the weights:

Numerator: expected long run proportion of time when A is leftmost and B is second from left

Denominator: expected long run proportion of time when A is leftmost and C is second from left

Simplify your answer as much as possible. What do you notice?