

# STAT/BIOST 571: Homework 1

Philip Pham

January 24, 2019

## Problem 1: Non-parametric linear regression (5 points)

In the context of random  $\mathbf{X}$  linear regression, consider  $n$  independent observations  $(x_i, Y_i)$  generated as follows. First, the independent variable  $x_i$  is selected according to

$$x_i = \begin{cases} 0 & \text{w.p. } 1/3 \\ 1 & \text{w.p. } 1/3 \\ 3 & \text{w.p. } 1/3 \end{cases},$$

then the systematic component of  $Y_i$  is set to

$$\mu_i = \begin{cases} 0.6 & \text{if } x_i = 0 \\ 3.6 & \text{if } x_i = 1 \\ 6.8 & \text{if } x_i = 3 \end{cases},$$

and finally  $Y_i \sim N(\mu_i, \sigma^2)$  with  $\sigma^2 = 1$ .

- (a) Simulate a single realization from the data-generating mechanism described above (with  $n = 20$ ) and present the data in a scatterplot, with the  $Y_i$  on the vertical axis and the  $x_i$  on the horizontal axis. Fit a simple linear regression model of the form

$$Y_i = \beta_0 + x_i\beta_1 + \epsilon_i$$

by ordinary least squares and overlay the fitted regression line on your scatterplot. Comment on the appropriateness of fitting a linear model to these data.

**Solution:** See Figure 1.

The line isn't a very good fit. It overestimates when  $x = 0$  and underestimates when  $x = 1$ . This isn't surprising since the linear model is incorrect.

- (b) Let  $\hat{\beta}_1$  be the estimated slope parameter from fitting the model above, and define  $\beta_1$  to be the quantity for which  $\hat{\beta}_1$  is consistent, in the limit as  $n \rightarrow \infty$ . Using the result on Slide 1.18 of the course lecture notes, calculate  $\beta_1$  exactly.

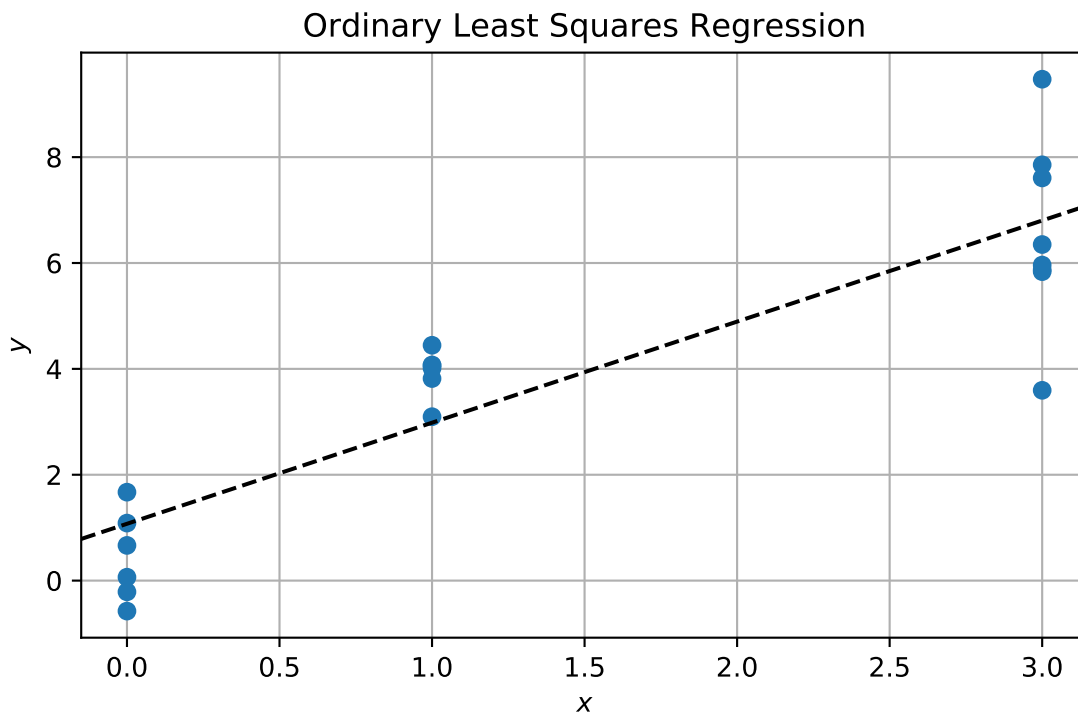


Figure 1: Ordinary least squares for data according to the model in Problem 1.

**Solution:**  $\beta_0 = 1$  and  $\beta_1 = 2$ .

Let  $\mathbb{P}$  be the probability measure over  $x$  and  $y$ . Let  $\phi(x | \mu, \sigma^2)$  be the density of the normal distribution with mean  $\mu$  and variance  $\sigma^2$ . Consider the integral,

$$\begin{aligned}
 L(\gamma) &= \int (y - (\gamma_0 + \gamma_1 x))^2 d\mathbb{P}(x, y) \\
 &= \frac{1}{3} \int_{-\infty}^{\infty} (y - \gamma_0)^2 \phi(y | 0.6, 1) dy + \\
 &\quad \frac{1}{3} \int_{-\infty}^{\infty} (y - \gamma_0 - \gamma_1)^2 \phi(y | 3.6, 1) dy + \\
 &\quad \frac{1}{3} \int_{-\infty}^{\infty} (y - \gamma_0 - 3\gamma_1)^2 \phi(y | 6.8, 1) dy \\
 &= \frac{1}{3} (62.56 - 22\gamma_0 - 48\gamma_1 + 3\gamma_0^2 + 10\gamma_1^2 + 8\gamma_0\gamma_1).
 \end{aligned}$$

If we let  $\beta = \arg \min_{\gamma} L(\gamma)$ . Taking the derivative, we setting  $\frac{\partial L}{\partial \gamma}(\beta) = 0$  to obtain  $\beta_0 = 1$  and  $\beta_1 = 2$ .

- (c) Download White's seminal paper from 1980 on sandwich estimation from the course website . Read (at least) the first few pages and explain how the value of  $\beta_1$  you derived in part (b) also follows from Lemma 1 of the paper.

**Solution:** In fitting the model, we are assuming that the linear model is correct and Assumption 1 of the paper holds.

Assumption 2 is satisfied so since  $\mathbb{E}[\epsilon_i^2] = 1$  for all  $i$  and the expected covariance matrix  $= \bar{M}_n = n^{-1} \sum_{i=1}^n \mathbb{E}[X_i^2]$  is just a matrix  $1 \times 1$  with a single positive entry, so it is certainly nonsingular with a positive determinant.

Thus, Lemma 1 tells us that  $\hat{\beta} \xrightarrow{\text{a.s.}} \beta$ , where  $\hat{\beta} = (X^\top X)^{-1} X^\top Y$ .

Since  $\hat{\beta}$  is also the maximum likelihood estimate which minimizes mean squared error, it converges almost surely to the value of  $\beta$  that minimizes expected mean squared error.

- (d) For finite  $n$ , we cannot discuss properties of the expectation of  $\hat{\beta}_1$ , at least not in the conventional sense. Explain why not.

**Solution:** Note that  $\hat{\beta} = (X^\top X)^{-1} X^\top Y$ .

If the linear model were correct, that is,  $Y = X\beta + \epsilon$ , taking the expectation would give

$$\mathbb{E}[\hat{\beta}] = \beta + \mathbb{E}[(X^\top X)^{-1} X^\top \epsilon] = \beta. \quad (1)$$

However, This calculation is incorrect when  $Y$  depends on  $X$  in a non-linear way as with these data.

For the remainder of this problem, condition on realizations of the data for which there exist  $i$  and  $i'$  such that  $x_i \neq x_{i'}$ .

- (e) Conduct a simulation study to assess the bias of  $\hat{\beta}_1$  for estimating  $\beta_1$ . Plot the estimated bias as a function of  $n$  (for  $n = 2, 4, 6, \dots, 28, 30, 35, \dots, 95, 100$ ), and overlay the standard error of  $\hat{\beta}_1$  as a function of  $n$ . Note that for this problem you should estimate the standard error of  $\hat{\beta}_1$  based on the standard deviation of  $\hat{\beta}_1$  across simulated realizations of the data, not using any further exact or asymptotic calculations.

**Solution:** See Table 1 for the slope estimates  $\hat{\beta}_1$  and standard error estimates  $\hat{\sigma}$  for different values of  $n$ .

These data are plotted in Figure 2.

- (f) Comment on your findings in part (e). Is the bias in estimating  $\beta_1$  surprising? Why or why not? How concerned should one be about this bias?

**Solution:** The bias is not surprising. As discussed in Part (d), the usual method of proving that the estimate is unbiased does not work since the true model is not linear.

The bias decreases quickly with increasing  $n$ , so it is not too concerning with reasonable sample sizes. Moreover, the standard error is larger than the bias, so in the bias-variance

| $n$ | $\hat{\beta}_1$ | $\hat{\sigma}$ | $\mathbb{E}[\hat{\beta}_1]$ |
|-----|-----------------|----------------|-----------------------------|
| 2   | 2.222531        | 1.115256       | 2.222222                    |
| 4   | 2.119261        | 0.752701       | 2.119358                    |
| 6   | 2.058301        | 0.536735       | 2.057653                    |
| 8   | 2.028134        | 0.408784       | 2.028077                    |
| 10  | 2.014462        | 0.332531       | 2.014561                    |
| 12  | 2.008274        | 0.284690       | 2.008200                    |
| 14  | 2.005376        | 0.253067       | 2.005023                    |
| 16  | 2.003600        | 0.231102       | 2.003320                    |
| 18  | 2.002070        | 0.213511       | 2.002338                    |
| 20  | 2.001587        | 0.200356       | 2.001732                    |
| 22  | 2.001091        | 0.189500       | 2.001336                    |
| 24  | 2.001169        | 0.180668       | 2.001063                    |
| 26  | 2.000937        | 0.172259       | 2.000867                    |
| 28  | 2.000748        | 0.165419       | 2.000721                    |
| 30  | 2.000182        | 0.159203       | 2.000610                    |
| 30  | 2.000505        | 0.159288       | 2.000610                    |
| 35  | 2.000618        | 0.146462       | 2.000424                    |
| 40  | 2.000300        | 0.136427       | 2.000312                    |
| 45  | 2.000245        | 0.127994       | 2.000239                    |
| 50  | 2.000259        | 0.121297       | 2.000189                    |
| 55  | 2.000184        | 0.115340       | 2.000154                    |
| 60  | 2.000254        | 0.110308       | 2.000127                    |
| 65  | 2.000014        | 0.105664       | 2.000107                    |
| 70  | 2.000246        | 0.101802       | 2.000091                    |
| 75  | 2.000068        | 0.098277       | 2.000079                    |
| 80  | 2.000033        | 0.094885       | 2.000069                    |
| 85  | 2.000091        | 0.092020       | 2.000060                    |
| 90  | 2.000172        | 0.089272       | 2.000054                    |
| 95  | 1.999983        | 0.087027       | 2.000048                    |
| 100 | 1.999924        | 0.084749       | 2.000043                    |

Table 1: The estimate for the slope ( $\hat{\beta}_1$ ) and standard error for the estimate ( $\hat{\sigma}$ ) were each calculated with  $10^6$  trials. The last column  $\mathbb{E}[\hat{\beta}_1]$  was calculated numerically by enumerating over the possible draws of  $X$ , which is detailed in Part (g).

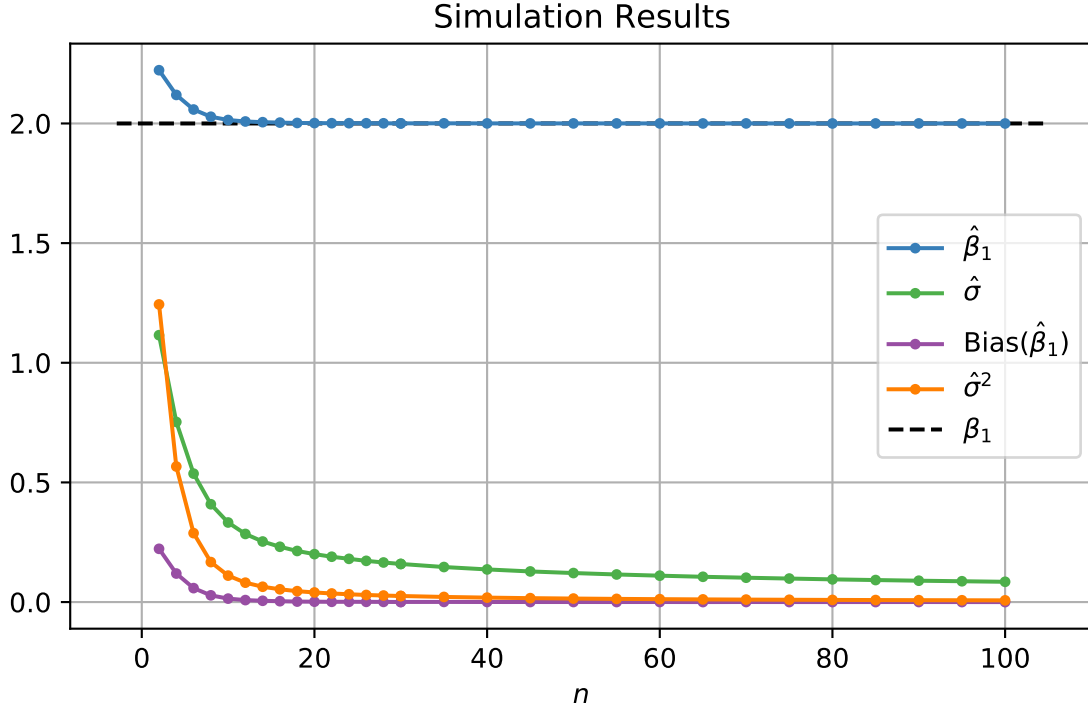


Figure 2: Data from Table 1 along with asymptotic slope ( $\beta_1$ ), the bias ( $\hat{\beta}_1 - \beta_1$ ), and variance ( $\hat{\sigma}^2$ ).

decomposition,

$$\begin{aligned}
 \mathbb{E} \left[ \left( \hat{\beta}_1 - \beta_1 \right)^2 \right] &= \mathbb{E} \left[ \left( \hat{\beta}_1 - \beta_1 \right)^2 \right] = \mathbb{E} \left[ \hat{\beta}_1^2 \right] - 2\mathbb{E} \left[ \hat{\beta}_1 \right] \beta_1 + \beta_1^2 \\
 &= \text{var} \left( \hat{\beta}_1 \right) + \mathbb{E} \left[ \hat{\beta}_1 \right]^2 - 2\mathbb{E} \left[ \hat{\beta}_1 \right] \beta_1 + \beta_1^2 \\
 &= \left( \mathbb{E} \left[ \hat{\beta}_1 \right] - \beta_1 \right)^2 + \text{var} \left( \hat{\beta}_1 \right) \\
 &= \left[ \text{Bias} \left( \hat{\beta}_1 \right) \right]^2 + \text{var} \left( \hat{\beta}_1 \right),
 \end{aligned}$$

the variance term which is the standard error squared will dominate. Indeed, according to Figure 2, the variance is much larger than the bias when  $n$  is small. When  $n$  is large, both the bias and variance are small, so neither is too concerning.

- (g) It is possible to derive exact values for the biases estimated in part (d) by enumerating all possible draws of  $x_1, \dots, x_n$ . Do this calculation for  $n = 2, 4, 6, 8$ . Report your numerical findings and overlay them on your plot from part (e). Show how you can do the calculation entirely by hand for  $n = 2$  (you may use a computer for matrix algebra). Although it is highly advisable to use a computer for this calculation for  $n > 2$ , simulation results are not sufficient. Your calculation of the bias must be exactly correct up to the computer's floating point precision.

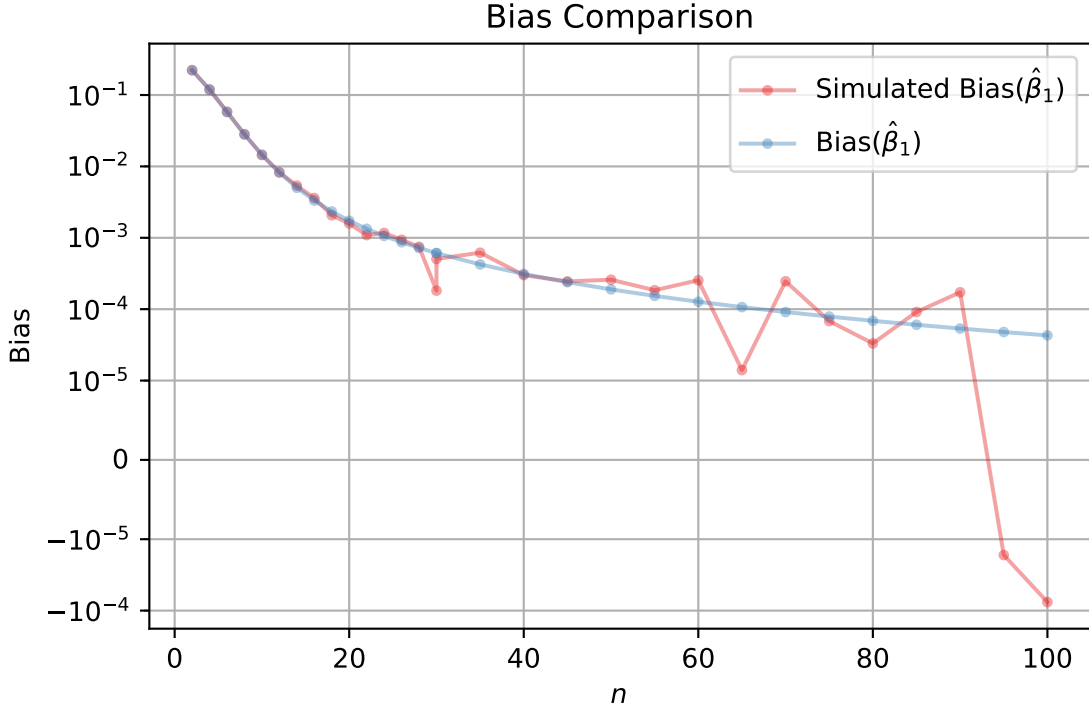


Figure 3: Comparison of the simulated bias from Part (e) against the exact calculation in Part (g).

**Solution:**  $\mathbb{E} [\hat{\beta}]$  can be efficiently calculated by noticing that

$$\mathbb{E} [\hat{\beta} | X] = \mathbb{E} [(X^\top X)^{-1} X^\top Y | X] = (X^\top X)^{-1} X^\top \mu. \quad (2)$$

is the same for different permutations of  $X$ .

Assume that each  $X_{(j_0, j_1, j_3)} \in \mathbb{R}^n$  is a vector of  $j_0$  0s,  $j_1$  1s, and  $j_3$  3s, where  $j_0 + j_1 + j_3 = n$  and  $j_0, j_1, j_3 \in \{0, 1, \dots, n-1\}$ . The triplets  $(j_0, j_1, j_3)$  can be seen as coming from the multinomial distribution, so

$$\mathbb{P}((j_0, j_1, j_3)) = \frac{n!}{j_0! j_1! j_3!} \left(\frac{1}{3}\right)^n. \quad (3)$$

Therefore, we have that

$$\begin{aligned} \mathbb{E} [\hat{\beta}] &= \mathbb{E} [\mathbb{E} [\hat{\beta} | X]] \\ &= \sum_{\{(j_0, j_1, j_3) \in \{0, 1, \dots, n-1\}^3 | j_0 + j_1 + j_3 = n\}} \mathbb{P}((j_0, j_1, j_3)) (X^\top X)^{-1} X^\top \mu \\ &= \sum_{\{(j_0, j_1, j_3) \in \{0, 1, \dots, n-1\}^3 | j_0 + j_1 + j_3 = n\}} \frac{n!}{j_0! j_1! j_3!} \left(\frac{1}{3}\right)^n (X^\top X)^{-1} X^\top \mu. \end{aligned}$$

The results of this calculation can be seen in the last column of Table 1. The biases from the simulation are plotted against the exact calculation in Figure 3.

Both the table and figure show that the simulation and calculations agree. There appears to be more noise with larger  $n$ , however.

## Problem 2: Average slope interpretation of generalized least squares (5 points)

Consider simple linear regression with a single covariate and an intercept, such that we are fitting

$$E(Y_i) = \beta_0 + \beta_1 x_i$$

with scalar observations  $Y_1, \dots, Y_n$ . As we saw on slide 1.18, the OLS estimator of  $\beta_1$  can be written as a weighted average of the pairwise slopes  $(Y_i - Y_j)/(x_i - x_j)$ . This is called the “average slope” representation.

- (a) Suppose that  $\Sigma$ , the covariance of  $\mathbf{Y}$ , is a diagonal matrix with known entries  $\sigma_1^2, \dots, \sigma_n^2$ . Derive the weights for an “average slope” representation for the optimal generalized least squares (GLS) estimate of  $\beta_1$  (see exercise 8.1 in the Wakefield text if you are not sure of the form of the optimal GLS estimate). A formal proof is optional, but you should provide a convincing justification for the specific form of your answer. HINT: One way to approach this problem is to think about what happens when the  $\sigma_i^2$  are all rational numbers, i.e., they can be expressed as fractions.

**Solution:** The likelihood function is

$$L_{X,y,\Sigma}(\beta) = f_\beta(y) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp\left(-\frac{1}{2}(y - X\beta)^\top \Sigma^{-1}(y - X\beta)\right). \quad (4)$$

Maximizing this likelihood by taking the log and derivative, we estimate  $\beta$  to obtain the maximum likelihood estimate  $\hat{\beta} = (X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1} y$ .

Expanding this out, we find that

$$\hat{\beta} = \frac{\left( \begin{pmatrix} \sum_{i=1}^n \frac{x_i^2}{\sigma_i^2} & \sum_{i=1}^n \frac{y_i}{\sigma_i^2} \\ \sum_{i=1}^n \frac{1}{\sigma_i^2} & \sum_{i=1}^n \frac{x_i y_i}{\sigma_i^2} \end{pmatrix} - \begin{pmatrix} \sum_{i=1}^n \frac{x_i}{\sigma_i^2} & \sum_{i=1}^n \frac{x_i y_i}{\sigma_i^2} \\ \sum_{i=1}^n \frac{x_i}{\sigma_i^2} & \sum_{i=1}^n \frac{y_i}{\sigma_i^2} \end{pmatrix} \right)}{\left( \begin{pmatrix} \sum_{i=1}^n \frac{1}{\sigma_i^2} & \sum_{i=1}^n \frac{x_i}{\sigma_i^2} \\ \sum_{i=1}^n \frac{x_i}{\sigma_i^2} & \sum_{i=1}^n \frac{x_i^2}{\sigma_i^2} \end{pmatrix} - \begin{pmatrix} \sum_{i=1}^n \frac{1}{\sigma_i^2} & \sum_{i=1}^n \frac{x_i}{\sigma_i^2} \\ \sum_{i=1}^n \frac{x_i}{\sigma_i^2} & \sum_{i=1}^n \frac{x_i^2}{\sigma_i^2} \end{pmatrix} \right)}.$$

Let  $\bar{x} = \sum_{i=1}^n \frac{1}{\sigma_i^2} x_i / \sum_{i=1}^n \frac{1}{\sigma_i^2}$  denote the weighted mean. We can rewrite our estimate for the slope

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n \frac{1}{\sigma_i^2} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n \frac{1}{\sigma_i^2} (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n \frac{1}{\sigma_i^2} \sum_{j=1}^n \frac{1}{\sigma_j^2} \sum_{k=1}^n \frac{1}{\sigma_k^2} (x_i - x_j)(y_i - y_k)}{\sum_{i=1}^n \frac{1}{\sigma_i^2} \sum_{j=1}^n \frac{1}{\sigma_j^2} \sum_{k=1}^n \frac{1}{\sigma_k^2} (x_i - x_j)(x_i - x_k)}.$$

Now with some algebra,

$$\begin{aligned}
\sum_{i=1}^n \frac{1}{\sigma_i^2} \sum_{j=1}^n \frac{1}{\sigma_j^2} \sum_{k=1}^n \frac{1}{\sigma_k^2} (x_i - x_j) (y_i - y_k) &= \sum_{j=1}^n \frac{1}{\sigma_j^2} \sum_{i=1}^n \frac{1}{\sigma_i^2} \sum_{k=1}^n \frac{1}{\sigma_k^2} (x_i - x_j) (y_i - y_k) \\
&= \left( \sum_{k=1}^n \frac{1}{\sigma_k^2} \right) \sum_{i=1}^n \frac{1}{\sigma_i^2} \sum_{j=1}^n \frac{1}{\sigma_j^2} (x_i - x_j) y_i \\
&= \frac{1}{2} \left( \sum_{k=1}^n \frac{1}{\sigma_k^2} \right) \sum_{i=1}^n \frac{1}{\sigma_i^2} \sum_{j=1}^n \frac{1}{\sigma_j^2} (x_i - x_j) (y_i - y_j),
\end{aligned}$$

where the last two steps take advantage of the symmetry of the  $i$  and  $j$  indices. Using this result, we have can write our slope estimate

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^n \frac{1}{\sigma_i^2} \sum_{j=1}^n \frac{1}{\sigma_j^2} (x_i - x_j) (y_i - y_j)}{\sum_{i=1}^n \frac{1}{\sigma_i^2} \sum_{j=1}^n \frac{1}{\sigma_j^2} (x_i - x_j)^2} \\
&= \boxed{\sum_{i=1}^n \sum_{\substack{j=1 \\ x_i \neq x_j}}^n \left( \frac{(x_i - x_j)^2}{\sigma_i^2 \sigma_j^2} \right) \left( \frac{y_i - y_j}{x_i - x_j} \right) / \sum_{i=1}^n \sum_{j=1}^n \frac{(x_i - x_j)^2}{\sigma_i^2 \sigma_j^2}}.
\end{aligned}$$

Thus, the unnormalized weight for the slope between  $i$  and  $j$  is  $\boxed{\frac{(x_i - x_j)^2}{\sigma_i^2 \sigma_j^2}}$ . The normalized

weights are  $\boxed{w_{ij} = \frac{(x_i - x_j)^2}{\sigma_i^2 \sigma_j^2} / \sum_{i=1}^n \sum_{j=1}^n \frac{(x_i - x_j)^2}{\sigma_i^2 \sigma_j^2}}.$

Intuitively, this make sense. As  $\sigma_i^2$  and  $\sigma_j^2$  increase, the weight decreases since our estimated slope is noisier. On the other hand, as  $(x_i - x_j)^2$  increases the weight increases because the slope estimate becomes more robust to noise since the effect size is larger. To see this mathematically, note that  $\frac{y_i - y_j}{x_i - x_j} = \beta_1 + \frac{\sigma_i \epsilon_i - \sigma_j \epsilon_j}{x_i - x_j}$ , where  $\epsilon_i \sim \mathcal{N}(0, 1)$ . The error term increases in magnitude as a function of  $\sigma_i^2$  and  $\sigma_j^2$  and decreases in magnitude as a function of  $(x_i - x_j)^2$ .

- (b) Now consider what the weights would be in an “average slope” representation for the optimal GLS estimate of  $\beta_1$ , for a general covariance matrix  $\Sigma$ . Give some properties of the weights that you would expect to hold. In answering this question, it might be helpful to think about how the weights would change as you vary individual entries in  $\Sigma$ . Also discuss any situations where you would expect the weight for a particular pair of observations to be infinite or zero. Note that you are **not** asked to derive an average slope representation for GLS.

**Solution:** Let  $\mathbf{1}$  be a vector a in  $\mathbb{R}^n$  of all 1s. Denote the weighted mean,  $\bar{x} = \mathbf{1}^\top \Sigma^{-1} x / \mathbf{1}^\top \Sigma^{-1} \mathbf{1}$ . Denote the centered vector,  $\tilde{x} = x - \bar{x} \mathbf{1}$ .

Then, we have that

$$\hat{\beta}_1 = \frac{\tilde{x}^\top \Sigma^{-1} \tilde{y}}{\tilde{x}^\top \Sigma^{-1} \tilde{x}} = \sum_{i=1}^n \sum_{j=1}^n \Sigma_{ij}^{-1} \tilde{x}_i \tilde{y}_j / \sum_{i=1}^n \sum_{j=1}^n \Sigma_{ij}^{-1} \tilde{x}_i \tilde{x}_j. \quad (5)$$



Let  $\Delta \in \mathbb{R}^{n \times n}$  be the skew-symmetric matrix where  $\Delta_{ij} = x_i - x_j$ . Let  $\Sigma_i$  denote the  $i$ th row or column of the covariance matrix (the matrix is symmetric). We can rewrite our slope estimate as

$$\hat{\beta}_1 = \sum_{i=1}^n \sum_{\substack{j=1 \\ x_i \neq x_j}}^n \frac{y_i - y_j}{x_i - x_j} (x_i - x_j) \operatorname{tr} \left( \Delta \Sigma_i^{-1} \left( \Sigma_j^{-1} \right)^\top \right) / \sum_{i=1}^n \sum_{j=1}^n (x_i - x_j) \operatorname{tr} \left( \Delta \Sigma_i^{-1} \left( \Sigma_j^{-1} \right)^\top \right),$$

where  $\operatorname{tr}$  is the trace operator. Since

$$\begin{aligned} - (x_i - x_j) \operatorname{tr} \left( \Delta \Sigma_i^{-1} \left( \Sigma_j^{-1} \right)^\top \right) &= (x_i - x_j) \sum_{k=1}^n \sum_{l=1}^n (x_k - x_l) \Sigma_{ik}^{-1} \Sigma_{jl}^{-1} \\ &= (x_i - x_j) \sum_{k=1}^n \sum_{l=k+1}^n (x_k - x_l) \det \begin{pmatrix} \Sigma_{ik}^{-1} & \Sigma_{il}^{-1} \\ \Sigma_{jk}^{-1} & \Sigma_{jl}^{-1} \end{pmatrix}, \end{aligned} \quad (6)$$

we can see the pairwise slopes are weighted according to the entries from the precision matrix. Moreover, the situation differs from the independent case in that the weight of each slope depends on all possible pairwise differences  $x_j - x_l$ , which intuitively makes sense since there may be some correlation with the other observations.

Let  $\tilde{\rho}_{ij}$  be the partial correlation coefficient between  $Y_i$  and  $Y_j$  given the remaining  $Y_k$ . The partial correlation measures the correlation between observations controlling for cofounders. Entries of the precision matrix are the scaled partial correlations,  $\Sigma_{ij}^{-1} = \tilde{\rho}_{ij} \sqrt{\Sigma_{ii}^{-1} \Sigma_{jj}^{-1}}$ .

With this identity, we can rewrite unnormalized weights for the slope estimate

$$- (x_i - x_j) \operatorname{tr} \left( \Delta \Sigma_i^{-1} \left( \Sigma_j^{-1} \right)^\top \right) = (x_i - x_j) \Sigma_{ii}^{-1} \Sigma_{jj}^{-1} \sum_{k=1}^n \sum_{l=k+1}^n (x_k - x_l) \det \begin{pmatrix} \tilde{\rho}_{ik} & \tilde{\rho}_{il} \\ \tilde{\rho}_{jk} & \tilde{\rho}_{jl} \end{pmatrix}, \quad (7)$$

Equation 7 gives us insight into how the covariance matrix  $\Sigma$  affects the weights for the pairwise slope between observations  $i$  and  $j$ .

As before, large differences  $x_i - x_j$  increase the magnitude of the weight, and large variances, manifested as small entries of  $\Sigma_{ii}^{-1}$  and  $\Sigma_{jj}^{-1}$ , decrease the magnitude of the weight. Now, the differences  $x_j - x_l$  also affect the weight in a similar manner to  $x_i - x_j$ . The partial correlation coefficients govern the degree of this effect. If there is no partial correlation, that is,  $\tilde{\rho}_{ik} = 0$  and  $\tilde{\rho}_{il} = 0$  or  $\tilde{\rho}_{jk} = 0$  and  $\tilde{\rho}_{jl} = 0$ , then the term involving  $x_k - x_l$  is zeroed out. Of course, when the observations are independent only the terms with  $(k, l) = (i, j)$  or  $(k, l) = (j, i)$  are nonzero, and we recover the same weights as the previous part.

### Problem 3: Cross-sectional/longitudinal effects, partitioning the exposure, mean model misspecification (10 points)

Consider a made-up computer literacy trial similar to the one discussed in lecture (slides 1.35–1.48). There are  $n$  subjects, indexed by  $i = 1, \dots, n$ , each one of which is observed at  $m$  followup times, indexed by  $j = 1, \dots, m$ . Let  $Y_{ij}$  be the literacy score of subject  $i$  at follow-up time  $j$ , at which time the subject's

age is  $x_{ij}$ . Suppose the design is fixed, meaning that the  $x_{ij}$  are all deterministic and known in advance, and assume the true data-generating mechanism can be written

$$Y_{ij} = f(x_{i1}) + \beta_L(x_{ij} - x_{i1}) + \epsilon_{ij} \quad (8)$$

with i.i.d.  $\epsilon_{ij} \sim N(0, \sigma^2)$ . The unknown function  $f(\cdot)$  represents a potentially nonlinear cohort effect and  $\beta_L$  is the coefficient for the linear longitudinal effect. Suppose we try to estimate  $\beta_L$  using the exposure partitioning method proposed on slide 1.44 (see, also, Diggle et al. page 16), based on the linear regression estimating equations with the mean model

$$E(Y_{ij}) = \beta_0 + \beta_C x_{i1} + \beta_L(x_{ij} - x_{i1}). \quad (9)$$

- (a) Give a sufficient condition on the data-generating mechanism for  $Y_{ij}$  in equation (8) to guarantee unbiased estimation of  $\beta_L$  and valid model-based standard errors.
- (b) Give a sufficient condition on the choice of follow-up times in the study design to guarantee unbiased estimation of  $\beta_L$ , even if the condition in part (a) is violated. Thinking about when confounding bias is a problem will be helpful here.
- (c) Download the modified made-up literacy data from the course web site (you can find these data with the homework files, not in the general dataset section), and generate one or more exploratory plots. Describe how you can graphically see evidence that neither of the above conditions is satisfied.
- (d) Propose an alternative regression model that can still be used to unbiasedly estimate  $\beta_L$  and give valid model-based standard errors. Fit this model to the downloaded data using the `lm()` function in R and report your findings (point estimate and standard error). Compare to what you obtain by fitting model (9) to the data.
- (e) Design and conduct a simulation study to illustrate the following
  - (a) Model (9) can fail to give unbiased estimates of  $\beta_L$  without at least one of the additional conditions from parts (a) and (b).
  - (b) Model (9) does give unbiased estimates if either of the additional conditions from parts (a) and (b) is satisfied.
  - (c) The model you propose in part (d) works, regardless.

You will need to consider a few choice of  $f(\cdot)$  and study designs  $\{x_{ij}\}$  to illustrate points (i) through (iii). Note that in earlier parts of the problem, you should have argued theoretically that each of these points is true; the purpose of the simulation study is to verify and illustrate your conclusions.

- (f) In your simulation study, what do you notice about the standard error estimates from the model in equation (9) when the condition from part (b) is satisfied but the condition from part (a) is not satisfied? Do sandwich standard errors fix the problem?
- (g) Explain, theoretically, what is going wrong in part (f) of this problem. This involves some fairly careful thinking about misspecified mean models and what is required for sandwich estimation to be valid.

## Appendix

Code for fitting models and generating plots is attached.