STAT/BIOST 571: Homework 1

Philip Pham

January 22, 2019

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

In the context of random $\boldsymbol{X}$ linear regression, consider n independent observations (x_i, Y_i) generated as follows. First, the indepenent 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 approporiateness 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 β_1 to be the quantity for which $\hat{\beta}_1$ is consistent, in the limit as $n \to \infty$. Using the result on Slide 1.18 of the course lecture notes, calculate β_1 exactly.

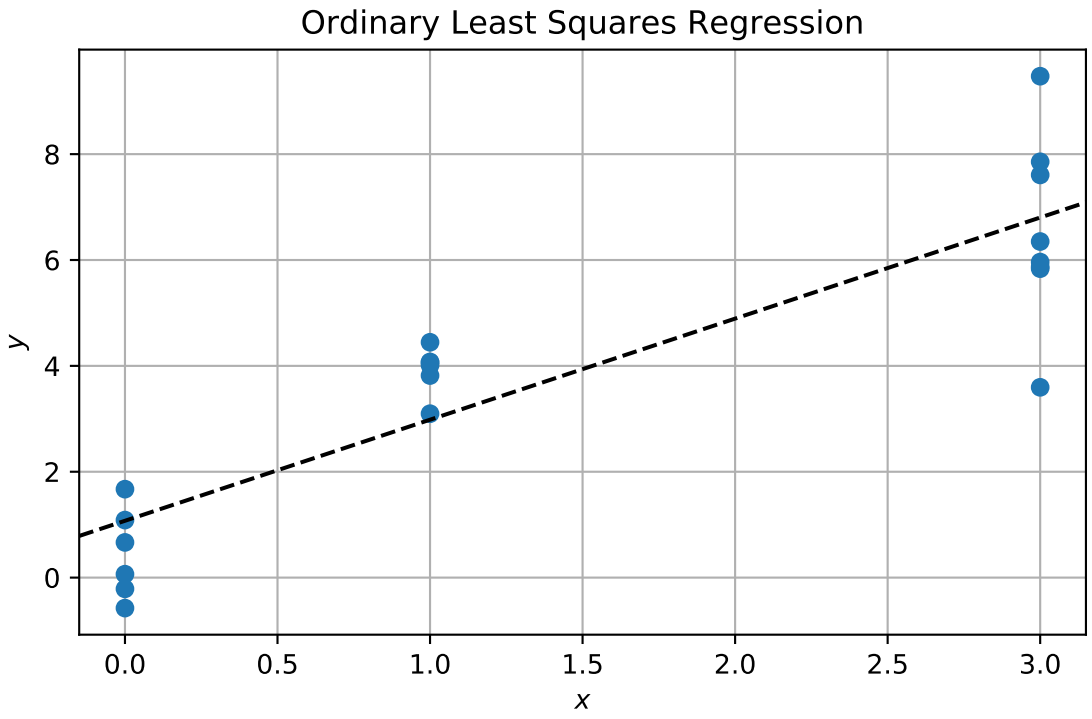

Figure 1: Ordinary least squares for data according to the model in Problem 1.

Solution: $\beta_0 = 1$ and $\boxed{\beta_1 = 2.}$

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

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

If we let $\beta = \arg\min_\gamma L(\gamma)$. Taking the derviative, 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 β_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 $\mathbb{E}\left[\epsilon_i^2\right] = 1$ for all i and the expected covariance matrix $= \bar{M}_n = n^{-1}\sum_{i=1}^n \mathbb{E}\left[X_i^2\right]$ is just a matrix 1×1 with a single positive entry, so it is certainly nonsingular is with a positive determinant.

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

Since $\hat{\beta}$ is also the maximum likelihood estimate which minimizes mean squared error, it converges almost surely to the value of β 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 $\hat{\beta} = (X^\intercal X)^{-1} X^\intercal Y$.

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

$$\mathbb{E}\left[\hat{\beta}\right] = \beta + \mathbb{E}\left[(X^\intercal X)^{-1} X^\intercal \epsilon\right] = \beta. \tag{1}$$

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 β_1. Plot the estimated bias as a function of n (for $n = 2, 4, 6, \ldots, 28, 30, 35, \ldots, 95, 100$), 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 β_1 surprising? Why or why not? How concerned should one be about this bias?

Solution: The bias is not suprising. 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 $\mathbb{E}\left[\hat{\beta}_1\right]$ was calculated numerically by enumerating over the possible draws of X.

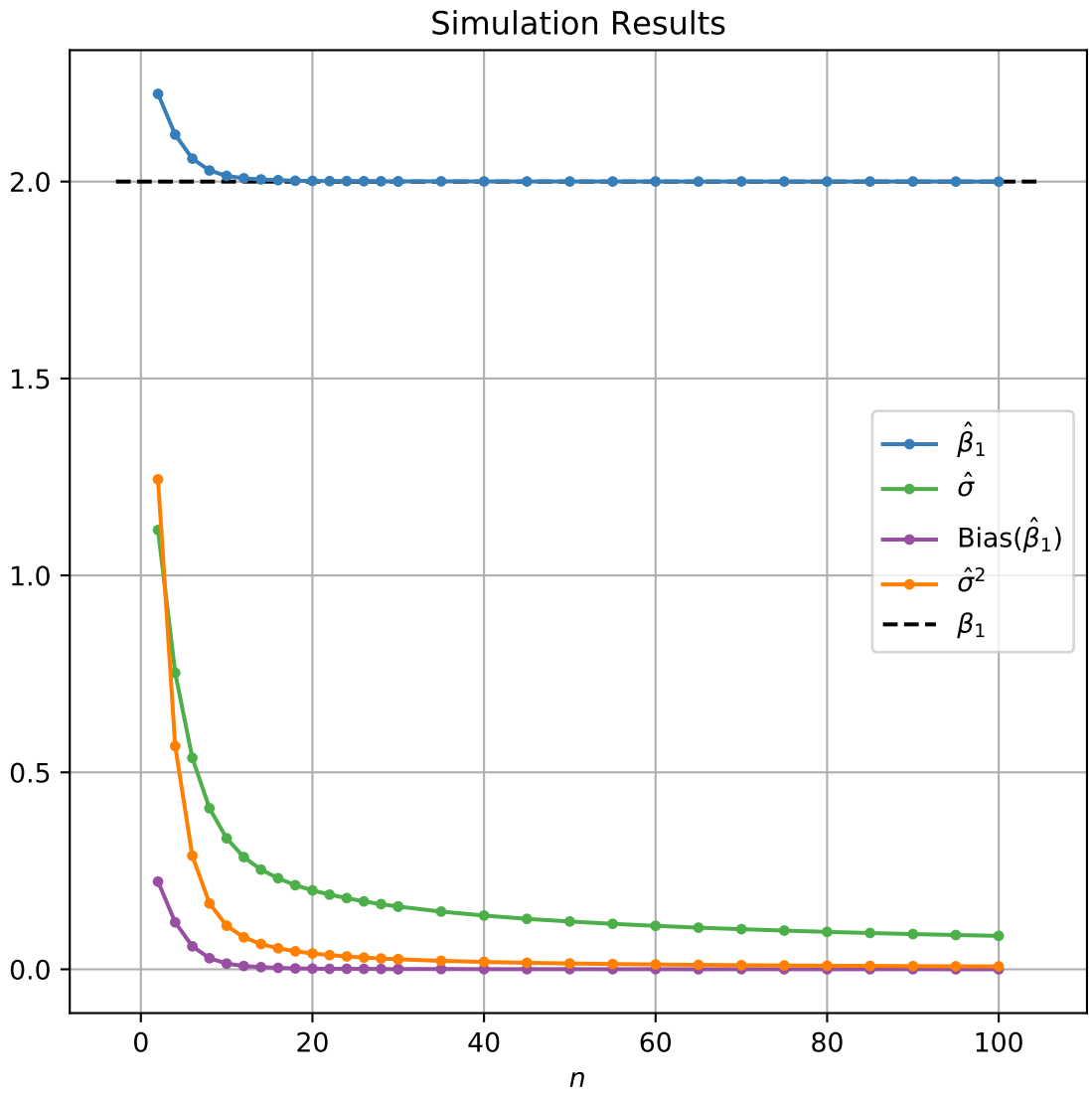

Figure 2: Data from Table 1 along with asymptotic slope (β_1), the bias ($\hat{\beta}_1 - \beta_1$), and variance ($\hat{\sigma}^2$).

decomposition,

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

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 $x_1, \ldots, x_n$. 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.

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 $Y_1, \ldots, Y_n$. As we saw on slide 1.18, the OLS estimator of β_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 $\boldsymbol{\Sigma}$, the covariance of $\boldsymbol{Y}$, is a diagonal matrix with known entries $\sigma_1^2, \ldots, \sigma_n^2$. Derive the weights for an "average slope" representation for the optimal generalized least squares (GLS) estimate of β_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 σ_i^2 are all rational numbers, i.e., they can be expressed as fractions.

(b) Now consider what the weights would be in an "average slope" representation for the optimal GLS estimate of β_1, for a general covariance matrix $\boldsymbol{\Sigma}$. 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 $\boldsymbol{\Sigma}$. 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.

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 $i = 1, \ldots, n$, each one of which is observed at m followup times, indexed by $j = 1, \ldots, m$. 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

$$Y_{ij} = f(x_{i1}) + \beta_L(x_{ij} - x_{i1}) + \epsilon_{ij} \tag{2}$$

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

$$E(Y_{ij}) = \beta_0 + \beta_C x_{i1} + \beta_L(x_{ij} - x_{i1}). \tag{3}$$

(a) Give a sufficient condition on the data-generating mechanism for Y_{ij} in equation (2) to guarantee unbiased estimation of β_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 β_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 β_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 (3) to the data.

(e) Design and conduct a simulation study to illustrate the following

- *(a) Model (3) can fail to give unbiased estimates of β_L without at least one of the additional conditions from parts (a) and (b).*
- *(b) Model (3) 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 (3) 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.*