STAT/BIOST 571: Homework 7

Philip Pham

March 4, 2019

Problem 1: Relationship between fixed effects, random effects, GLS, and penalized regression; confounding; model misspecification (20 points)

This is an extension of problem 3 from homework 1. There are n subjects, indexed by $i = 1, \ldots, n$, each one of which is observed at m follow-up 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}. The design is fixed, meaning that the x_{ij} are all deterministic and known in advance, and the true data-generating mechanism can be written

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

with i.i.d. $\epsilon_{ij} \sim N(0, \sigma^2)$. We are interested in estimating β_L, without knowing the form of $f(x)$. For this problem, fix $n = 91$ and $m = 3$, set the initial follow-up age for subject i to be

$$x_{i1} = 1 + 0.1 * (i - 1),$$

have subsequent follow-ups evenly spaced at intervals of

$$\Delta x_i = \left(1 + \frac{10 - x_{i1}}{10}\right)^2,$$

and consider the fixed (but unknown) mean model implied by $\beta_L = 1$ and $f(x) = (10 - x)^2$. Initially fix the true variance at $\sigma^2 = 100$, but you will consider different values of σ^2 later in the problem.

(a) *Simulate a single realization of data from the model specified above and generate one or more plots that allows you to visualize the marginal and conditional trends in computer literacy.*

Solution: A single realization is plotted in Figure 1. By looking at the data from follow period 0 and comparing subjects, one can identify the marginal trend: literacy decreases with the base age, so younger subjects generally have higher literacy. To see the conditional trend, one looks within each subject: generally, as a subject ages, their literacy increases. This conditional trend isn't very strong, however, since the variance is rather large relative to the effect size.

Based on the plot you have generated (and experience from homework 1), it should be clear that trying to estimate β_L by fitting a partitioned model in which you treat $f(x)$ as linear will not work. You explain

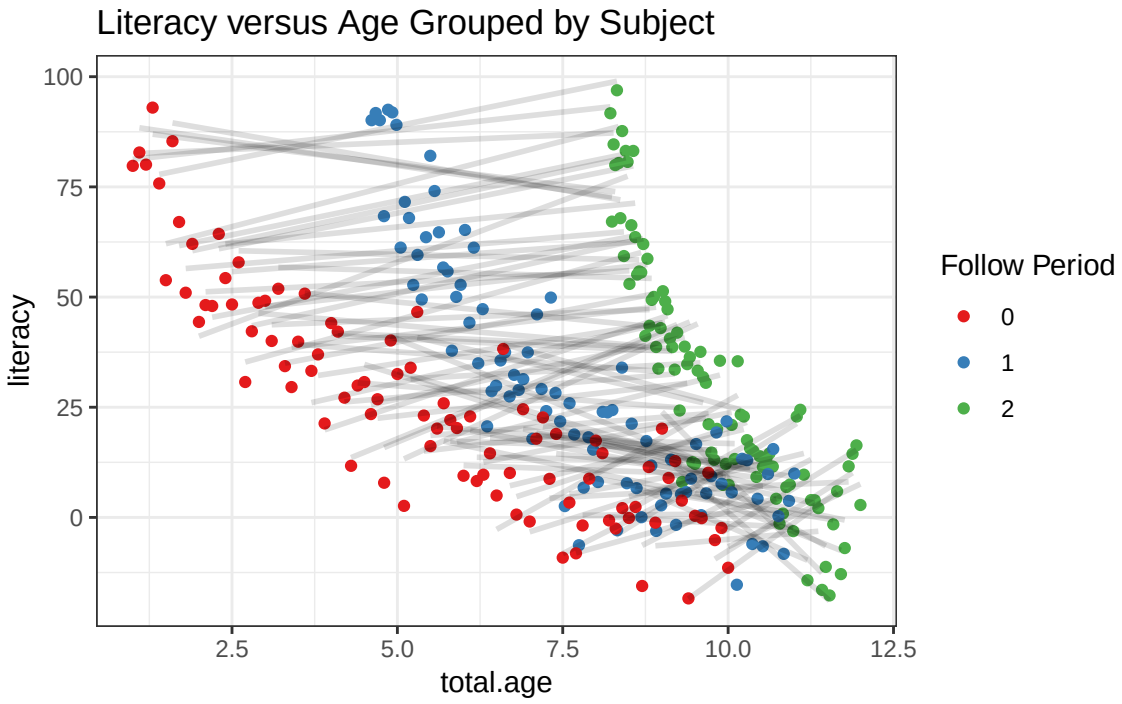

Figure 1: Plot of a single realization of the data with $\sigma^2 = 100$. The lines are fitted with OLS linear regression conditioned on each subject.

this to your collaborator and propose to fit a model where you include a fixed effect to stratify by subject. Your collaborator is sympathetic to the need to deal with confounding but is concerned that fitting a model with 92 *parameters and only* $91 \times 3 = 273$ *observations is not a very good idea. He suggests, as an alternative, adjusting for subject using a random intercept in a linear mixed effects model (fit using REML, of course).*

(b) *Explain to your collaborator, purely in words, why this is a bad idea.*

Solution: The random effects model assumes that the subject-specific adjustments to the intercept are normally distributed. This is not the case. See the histograms in Figure 2. In particular, look at the base literacy rate in follow period 0. This distribution is not normal and skews left. Thus, a random effects model would not properly account for the subject-specific variation.

Although you're sure that your explanation should have been persuausive, your collaborator is not yet convinced. So you conduct a simulation to further elucidate the differences between estimating β_L *from models with fixed effects and random effects for subject-specific intercepts.*

(c) *Write down the mathematical form of the two models you will be fitting and describe the simulation study.*

Solution: Let Y_i be the observed literacy rates for subject i. Let X_i be the covariates for subject i: an $m \times 2$ matrix of 1s in the first column and delta ages in the second. $\beta = \begin{pmatrix} \beta_0 & \beta_L \end{pmatrix}^\intercal$ are the fixed effect coefficients for the intercept and delta age.

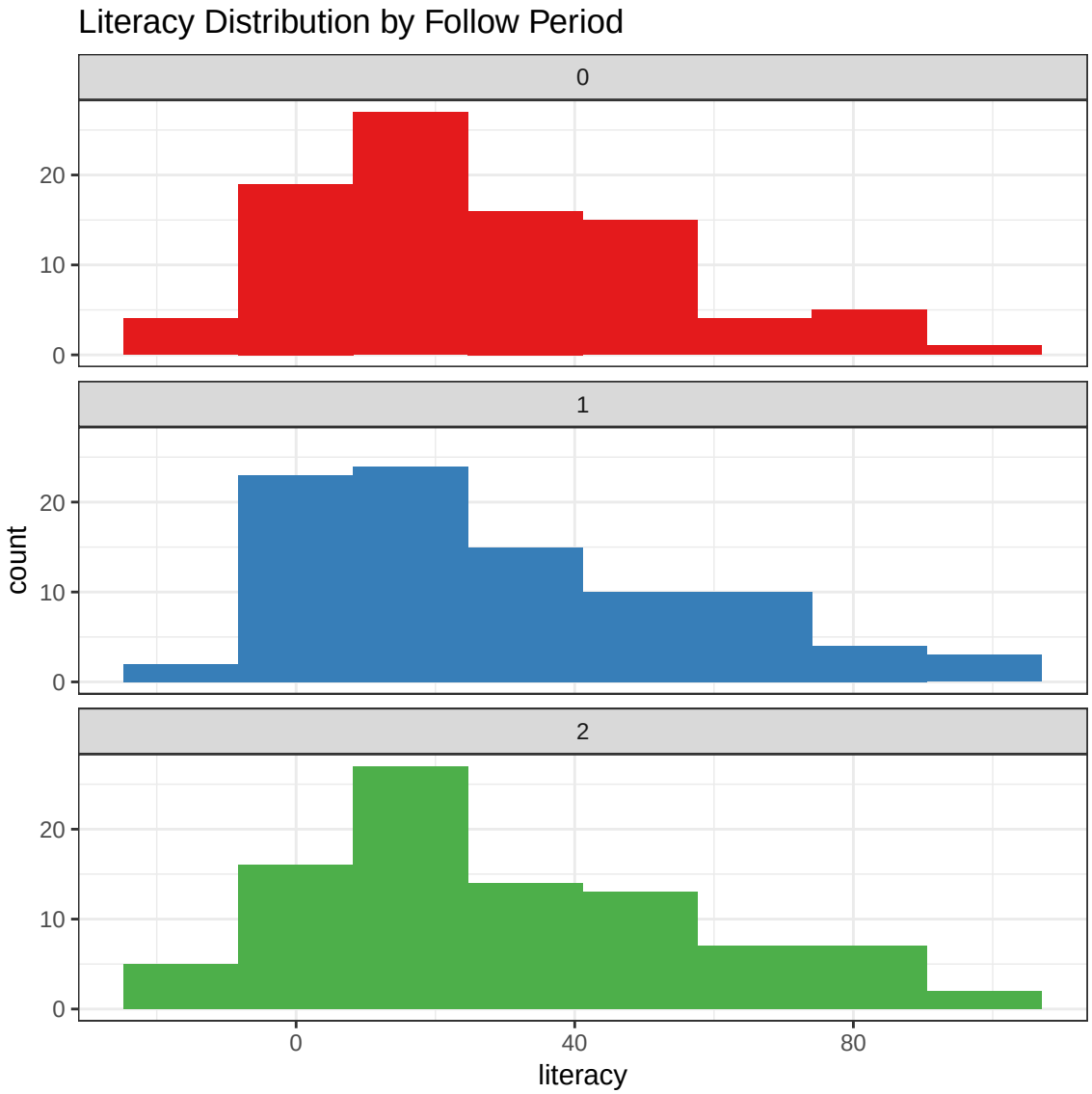

Figure 2: Distribution of literacy by the follow period. One sees that the distribution skews left and is not normal.

In the random effects model, we have that

$$\begin{aligned} Y_i &= X_i\beta + \mathbf{1}\gamma_i + \epsilon_i \\ \gamma_i &\sim \mathcal{N}\left(0, \sigma_\gamma^2\right) \\ \epsilon_{ij} &\sim \mathcal{N}\left(0, \sigma^2\right). \end{aligned} \tag{1}$$

The γ_i induces within-subject correlation. This model will be fit with a restricted maximum likelihood (REML) procedure.

In the fixed effects model, we have that

$$\begin{aligned} Y_i &= \begin{cases} X_i\beta + \epsilon_i, & i = 1 \\ X_i\beta + \mathbf{1}\alpha_i + \epsilon_i, & i \neq 1 \end{cases} \\ \epsilon_{ij} &\sim \mathcal{N}\left(0, \sigma^2\right). \end{aligned} \tag{2}$$

The α_i are subject-specific intercepts and considered fixed effects in this case. We can fit this model with ordinary least squares (OLS) regression that maximizes likelihood.

(d) *Present and explain the results of your simulation study to your collaborator, highlighting the simulation-based estimates of bias and standard errors of the $\hat{\beta}_L$ you obtain from each model and how these relate to his suggestion.*

	$\mathbb{E}\left[\hat{\beta}_L\right]$	$\mathbb{E}\left[\hat{\sigma}_{\hat{\beta}_L}\right]$	Sample $\hat{\sigma}_{\hat{\beta}_L}$	$\mathbb{E}\left[\hat{\sigma}\right]$	$\mathbb{E}\left[\hat{\sigma}_\gamma\right]$
Random Effects Intercept	1.276229	0.320826	0.328489	10.008001	24.350760
Fixed Effects Intercept	0.998420	0.321478	0.323965	9.983993	

Table 1: Results of a simulation study comparing modeling the subject-specific intercepts as a random effect or fixed effect. Parameter estimates were averaged over simulations. Standard errors for $\hat{\beta_L}$ are calculated two ways: (1) assuming the model is correct ($\mathbb{E}\left[\hat{\sigma}_{\hat{\beta}_L}\right]$), and (2) using the $\hat{\beta_L}$ samples (Sample $\hat{\sigma}_{\hat{\beta}_L}$).

Solution: 2^{12} simulations were done for each model. Equation 1 is used in the random effects intercept model, and Equation 2 is used in the fixed effects intercept model. The results can be seen in Table 1.

The random effects model has significant bias and overestimates β_L, which has true value $\beta_L = 1$, whereas the fixed effects model appears unbiased. The random effects model underestimates the subject-specific intercepts and compensates by overestimating β_L.

The standard errors of both models agree. The model-based standard errors also agree with the simulated standard errors in both models.

(e) *Summarize the estimated values of the residual and random effect variances ($\hat{\sigma}^2$ and $\hat{\sigma}_\gamma^2$, respectively) from your simulations.*

Solution: σ^2 measures the variance from the ϵ_{ij} term in the model, which is the error in each individual observation. The estimated value of the residual variance $\hat{\sigma}^2$ is very accurate in both models. See the square root in the $\mathbb{E}\left[\hat{\sigma}\right]$ column of Table 1. The actual value is 10.

σ_L^2 measures the variance associated from the γ_i term in the model, which accounts for subject-level differences. It's only estimated in the random effects model by $\hat{\sigma}_\gamma^2$.

(f) *Although the subject-specific intercepts in this problem are not really random, there is a natural quantity that you can think of $\hat{\sigma}_\gamma^2$ as trying to estimate. Calculate this number and call it σ_γ^2. Explain how your estimates $\hat{\sigma}_\gamma^2$ compare to σ_γ^2.*

Solution: σ_γ^2 can be thought of as the variance in the base literacy score. In this case, the base literacy score for subject i is $f(x_{i1}) = f(1 + 0.1(i-1))$.
We have that the standard error in base literacy is

$$\sigma_\gamma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(f(x_{i1}) - \frac{1}{n}\sum_{j=1}^{n} f(x_1)\right)^2} \approx 24.433055478. \tag{3}$$

Indeed, this number is very close to our estimate in the $\mathbb{E}[\hat{\sigma}_\gamma]$ column of Table 1.

For the remainder of the problem, consider the idealized situation where REML always estimates the variances exactly, meaning $\sigma^2 = \hat{\sigma}^2 = 100$ and $\sigma_\gamma^2 = \hat{\sigma}_\gamma^2$.

(g) *Calculate the exact (to your computer's numerical precision) bias of $\hat{\beta}_L$ from the random effect model as an estimator for β_L. Report your answer to five significant digits. Do this calculation two ways, first by interpreting the estimation procedure as GLS regression and then by interpreting it as penalized OLS regression. You cannot answer this question with a simulation study, but you are free to use either your own R code or any publicly available R packages for the calculations. You may find the `penalized` package in R helpful. Note that although we have seen a theorem that implies these two approaches will give precisely the same answer, in this problem you are asked to find the solution two separate ways, without using your knowledge that they are equivalent.*

Solution: The random effects model implies a subject-specific covariance structure of

$$\Sigma_i = \sigma_\gamma^2 \mathbf{1}\mathbf{1}^\intercal + \sigma^2 I, \tag{4}$$

where we use Equation 3 to compute σ_γ^2. Using Equation 4, we can compute $\hat{\beta}$ with the generalized least squares (GLS) estimator:

$$\hat{\beta}_{\text{GLS}} = \left(\sum_{i=1}^{n} X_i^\intercal \Sigma_i^{-1} X_i\right)^{-1} \sum_{i=1}^{n} X_i \Sigma_i^{-1} Y_i. \tag{5}$$

Let Y be the concatenation of the Y_i. Let X be the $nm \times (n+2)$ design matrix for an overspecified OLS problem. The first two columns are a result of stacking the X_i and the last $j+2$ column represents an indicator for subject j. The first 2 coefficients are the fixed effects, and the remaining n are random effects, which we want to penalize. In penalized OLS regression, we are minimizing the mean squared error with an additional penalty term:

$$\hat{\beta}_{\text{OLS}} = \arg\min_{\beta} \left[(Y - X\beta)^\intercal (Y - X\beta) + \beta^\intercal Q \beta\right] \tag{6}$$

$$\text{where } Q = \begin{pmatrix} 0 \cdot I_2 & \mathbf{0} \\ \mathbf{0} & \frac{\sigma^2}{\sigma_\gamma^2} \cdot I_n \end{pmatrix},$$

which means the fixed effects are not penalized, and the random effects are penalized with weight $\frac{\sigma^2}{\sigma_\gamma^2}$ as in slide 3.70 of the course notes.

(h) *Using the calculation from part (g) of this problem, plot the bias of $\hat{\beta}_L$ as a function of σ^2, over an interesting range of choices for σ^2 (continue to use the same fixed value of σ^2_γ you estimated from the data).*

(i) *Explain your plot from part (h) of this problem to your collaborator. Offer two explanations for the pattern in the plot, one based on the clustered WLS (equivalently, GLS) intepretation and the other based on the penalized OLS interpretation of the linear mixed effects model estimation algorithm.*

(j) *Your collaborator is nearly convinced, but not quite. He reminds you that you once told him mixed model effect estimates can be thought of as solutions to a GLS problem and that any old GLS estimator will do (regardless of correlation structure) to consistently estimate the coefficients in a linear regression model. Explain why this argument doesn't apply in the present situation.*