36-617: Applied Linear Models
Fall 2019
HW09 – Due Fri Nov 15, 11:59pm
Grace until Sun Nov 17, 11:59pm as usual

- Please turn the homework in online to the gradescope link under HW09 in our course webspace at canvas.cmu.edu, under Assignments.
- There is no reading and no quiz scheduled for Monday Nov 18. I will let you know when there is another reading/quiz assignment.
- There are two main exercises below.
- Since this assignment is coming out so late, I will hold an additional office hour 4–5 on Thu this week.

Exercises

1. Please list the people you collaborated with on this assignment.
2. **This is a math problem, not a data analysis problem.** Consider the following multilevel model for data y_i, $i = 1, \ldots, n$, arranged into J groups, $j = 1, \ldots, J$, where each group j has n_j observations:

$$\left.\begin{array}{rcl} y_i & = & \alpha_{j[i]} + \epsilon_i,\ \epsilon_i \overset{iid}{\sim} N(0, \sigma^2) \\ \alpha_j & = & \beta_0 + \eta_j,\ \eta_j \overset{iid}{\sim} N(0, \tau^2) \end{array}\right\} . \qquad (*)$$

   Prove the following assertions:

   (a) If $i \neq i'$ and $j[i] \neq j[i']$, then $\text{Corr}\,(y_i, y_{i'}) = 0$.
   (b) If $i \neq i'$ but $j[i] = j[i']$, then $\text{Corr}\,(y_i, y_{i'}) = \frac{\tau^2}{\tau^2+\sigma^2}$.
   (c) Let $\overline{y}_{j.} = \frac{1}{n_j} \sum_{i:j[i]=j} y_i$, the average of all observations in group j. Then $\text{Var}\,(\overline{y}_{j.}) = \tau^2 + \sigma^2/n_j$
   (d) Suppose we exactly replicate the experiment generating new data y_i^* following the model

$$\left.\begin{array}{rcl} y_i^* & = & \alpha_{j[i]} + \epsilon_i^*,\ \epsilon_i^* \overset{iid}{\sim} N(0, \sigma^2) \\ \alpha_j & = & \beta_0 + \eta_j,\ \eta_j \overset{iid}{\sim} N(0, \tau^2) \end{array}\right\} , \qquad (**)$$

   so that the group level α's and η's (and β_0) are the same between $(*)$ and $(**)$ [the conditions we are measuring didn't change] but the new set of ϵ^*'s are independent of η's and ϵ's [we re-measured, and so we have new measurement error on each observation]. Form the group averages $\overline{y}_{j.}^*$, analogous to $\overline{y}_{j.}$. Then

$$\text{Corr}\,(\overline{y}_{j.}, \overline{y}_{j.}^*) = \frac{\tau^2}{\tau^2 + \sigma^2/n_j} .$$

   In all four parts, carefully state any assumptions that you need.

[next problem on the next page]

3. **Looking at multilevel models.** Consider again the `cdi.dat` data from Project 02. Construct the variable `pct.hs.grad` $<-$ (`hs.grad` / `pop`) $\times$ 100% and add it to the data frame that you created for `cdi.dat` in R.

   (a) Fit the multilevel model

$$\begin{aligned}
(\texttt{per.cap.income})_i &= \alpha_{0j[i]} + \alpha_{1j[i]} \cdot (\texttt{pct.hs.grad})_i + \epsilon_i, \;\; \epsilon_i \overset{iid}{\sim} N(0, \sigma^2) \\
\alpha_{0j} &= \beta_0 + \eta_{0j}, \;\; \eta_{0j} \overset{iid}{\sim} N(0, \tau_0^2) \\
\alpha_{1j} &= \beta_1 + \eta_{1j}, \;\; \eta_{1j} \overset{iid}{\sim} N(0, \tau_1^2), \\
i &= 1, \ldots, 440 \;\; \text{(counties)} \\
j &= 1, \ldots, 48 \;\; \text{(states)}
\end{aligned}$$

   i.e. `per.cap.income ~ 1 + pct.hs.grad + (1 + pct.hs.grad | state)`, produce a summary of the model, and provide estimates of the following, from the summary:

   - $\widehat{\sigma^2}$
   - $\widehat{\tau_0^2}$
   - $\widehat{\tau_0^2}$
   - $\widehat{\text{Corr}}(\eta_0, \eta_1)$
   - $\hat{\beta}_0$
   - $\hat{\beta}_1$
   - $\text{SE}(\hat{\beta}_0)$
   - $\text{SE}(\hat{\beta}_1)$

   Note: R will probably complain that the model[1] *failed to converge, complain that the model is nearly unidentifiable, and suggest rescaling the variables. You could improve all of this by standardizing (subtract the mean and divide by the SD) the two continuous variables, but that is not necessary for this excercise*[2] *and would just complicate the rest of the problem, so let's just ignore the complaints.*

   (b) Create a facet plot for the CDI data like that on slide 7 of lecture 24 (mlm residuals), with x = `pct.hs.grad` and y = `per.cap.income` showing:
   - The raw data in each state;
   - The fitted regression line from the completely pooled lm() model that fits a single regression line to all of the data, ignoring differences among the states.
   - The fitted regression lines from the completely unpooled lm() model that fits a different regression line to each state, ignoring the data in the other states.
   - The fitted regression line using estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ from part (a).
   - The fitted regression lines using estimates $\hat{\alpha}_{0j}$ and $\hat{\alpha}_{1j}$ from part (a).

   (c) Comment on similarities and differences among the four regression lines plotted in each panel in part (b). A sentence or two will do.

   (d) Use the `Sigma.y()` function in `residual-functions.r` to create an `image()` plot of the variance-covarance matrix for `per.cap.income` under the model you fitted in part (a). Comment on any patterns you see. A sentence or two will do. *Note: you may need to sort the rows and columns of the variance-covariance matrix by state to see anything.*

   (e) Examine the marginal, conditional, and random effects residuals (again, using functions in `residual-functions.r`) from your model in part (a), separately by state in facet plots, and pooled across states in normal qq plots. Comment briefly on the residuals, and make a suggestion for improving the model. A sentence or two will do.

[1] ...by which it means the **model-fitting algorithm**...

[2] If you are really curious you could try this exercise both with the raw variables and with the standardized variables and compare the results; they are basically the same and using the raw variables is a bit more interpretable.