36-617: Applied Linear Models
Fall 2022
HW07 – Due Wed Nov 9, 11:59pm

- Please turn the homework in, as a single pdf, online in GradeScope using the link provided on the HW01 assignment page on canvas.cmu.edu, under Assignments. Upload *one* file per person.
- Reading:
  - For this week, you should finish Sheather 10.1 and Gelman & Hill, Ch 12.
  - For next week, please finish Gelman & Hill Ch 13 and read G&H Ch 14.
- There are 3 exercises below; each one has several "parts".

Exercises

1. **Data analysis.** On July 23, 2006, the Houston Chronicle published an article entitled "Reading: First-grade standard too tough for many". The article claimed in part that "more students (across Texas) are having to repeat first grade. Experts attribute the increase partially to an increase in poverty." The article presents data for each of 61 Texas school districts on

$$\left.\begin{array}{rcl} y & = & \text{Percentage of students repeating first grade} \\ x & = & \text{Percentage of low-income students} \end{array}\right\} \tag{1}$$

for both 2004–2005 and 1994–1995. The data in the file `houston.txt` is derived from this study, and contains the following variables:

```
District                - An individual school district
Pct.Repeating.1st.Grade - The percent of students repeating first grade in the district
Pct.Low.income.students - The percent of low-income students in the district
Year                    - The year of the observation: 1994 or 2004
County                  - The county that the school district is in
```

Fit ordinary linear models to this data with `lm()` to determine whether:

(a) An increase in the percentage of low income students is associated with an increase in the percentage of students repeating first grade?

(b) There has been an increase in the percentage of students repeating first grade between 1994–1995 and 2004-2005?

(c) Any association between the percentage of students repeating first grade and the percentage of low-income students differs between 1994–1995 and 2004–2005?

Use appropriate numerical and/or graphical tools to assess fit, compare models, determine whether predictors are significant, etc.

2. In the data for problem 1, it is plausible that there would be some variation among Counties in percentage of students repeating first grade. We will use a multilevel model to explore the variation across counties, *using the 2004 data* **only**.

   (a) Extract from the file `houston.txt` a data frame containing just the data from 2004. Report the following for the extracted data frame:

   - Number of groups, J. (Each County is a group.)
   - For each group j, number of observations n_j.
   - Total number of observations, n. (Each District is an observation.)

   Use only data from this data frame in the rest of this problem.

   (b) Use `lmer()` from `library(lme4)` to fit the multilevel model

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

   where i indexes Districts, j indexes Counties, and y_i is the percentage of students repeating first grade in the i^{th} District. Provide: a `summary()` of your model, and report the following quantities from your summary:

   - The value of $\hat{\sigma}^2$;
   - The value of $\hat{\tau}^2$;
   - The value of $\hat{\beta}_0$.

   (c) Use `library(gglpot2)` (or another R package if you prefer) to make a facet plot of y vs x in equation (1), with a different facet for each County. Add to the facets

   - A black horizontal line indicating the average of y in the whole data set (ignoring groups)
   - A red horizontal line indicating the average of y in each County
   - A green horizontal line indicating the "random effect" estimates $\hat{\alpha}_j = \hat{\eta}_j + \hat{\beta}_0$ for each County, as indicated in equation (2).

   and submit your plot. (Does your plot illustrate the "shrinkage" phenomenon of MLM's?)

   (d) Fit the ordinary linear model `lm(y ~ County)`, where the coefficients on the County dummy variables are constrained to sum to zero. Provide the `summary()` of your model, and decide from the output whether we need to keep the `County` variable in the model (this is also an informal way to decide whether to keep the multilevel model that you fitted in part (a)).

3. **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:

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

where the ϵ's and η's are also independent of each other. In all of the rest of this problem, variances, covariances and correlations are based on theoretical expected values, not on sample quantities. Thus, for any random variable that appears below (e.g., $U = y_i$, $U = \overline{y}_{j.}$, etc.):

- $\mu_U = E[U]$, not $\overline{u} = \frac{1}{m}\sum_{k=1}^{m} u_k$ in some sample
- $\mathrm{Var}\,(U) = E[(U - \mu_U)^2]$, not $s_u^2 = \frac{1}{m-1}\sum_{k=1}^{m}(u_k - \overline{u})^2$ in some sample
- $\mathrm{Cov}\,(U, V) = E[(U - \mu_U)(V - \mu_V)]$, not $c_{uv} = \frac{1}{m-1}\sum_{k=1}^{m}(u_k - \overline{u})(v_k - \overline{v})$ in some sample, and
- $\mathrm{Corr}\,(U, V) = \dfrac{\mathrm{Cov}\,(U,V)}{\sqrt{\mathrm{Var}\,(U)\mathrm{Var}\,(V)}}$, not $\dfrac{c_{uv}}{\sqrt{s_u^2 s_v^2}}$

Prove the following four assertions that were stated in lecture:

(a) If $i \neq i'$ and $j[i] \neq j[i']$, then $\mathrm{Corr}\,(y_i, y_{i'}) = 0$.

(b) If $i \neq i'$ but $j[i] = j[i']$, then $\mathrm{Corr}\,(y_i, y_{i'}) = \frac{\tau^2}{\tau^2+\sigma^2}$.

(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 $\mathrm{Var}\,(\overline{y}_{j.}) = \tau^2 + \sigma^2/n_j$

(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 and groups 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 sample with new measurement errors within each group]. Form the group averages $\overline{y}_{j.}^*$, analogous to $\overline{y}_{j.}$. Then

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

(This is another interpretation of the reliability coefficient $\frac{\tau^2}{\tau^2+\sigma^2/n_j}$.)

In all four parts, be sure to state any assumptions that you need.