Homework 07 Solutions

2022-11-04

```
library(lme4)
library(ggplot2)
```

Problem 1.

Problem 1a.

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 data in the file `houston.txt` is derived from this study. Here is a quick look:

```
houston <- read.table("houston.txt",header=T)

attach(houston)
houston$y <- Pct.Repeating.1st.Grade
houston$x <- Pct.Low.income.students
houston$Year <- factor(Year)
detach()

summary(houston)
```

```
##    District         Pct.Repeating.1st.Grade Pct.Low.income.students   Year
##  Length:122         Min.   : 0.000          Min.   : 3.20           1994:61
##  Class :character   1st Qu.: 2.825          1st Qu.:27.15           2004:61
##  Mode  :character   Median : 5.050          Median :41.35
##                     Mean   : 5.993          Mean   :41.88
##                     3rd Qu.: 8.475          3rd Qu.:53.02
##                     Max.   :17.800          Max.   :98.10
##     County                y                x
##  Length:122         Min.   : 0.000   Min.   : 3.20
##  Class :character   1st Qu.: 2.825   1st Qu.:27.15
##  Mode  :character   Median : 5.050   Median :41.35
##                     Mean   : 5.993   Mean   :41.88
##                     3rd Qu.: 8.475   3rd Qu.:53.02
##                     Max.   :17.800   Max.   :98.10
```

Here, I use ordinary linear models to this data with `lm()` *to answer each question. There is code below to make plots, but they are not very informative, so I commented them out to save space. Basically the residuals for each of the models are a bit right-skewed, but not enough to motivate me to make a transformation and complicate interpretation of the models.*

Problem 1a.

Is an increase in the percentage of low income students associated with an increase in the percentage of students repeating first grade?

We just have to regress `y` *on* `x` *and see if the coefficient on* `x` *is significantly different from zero.*

```
summary(lm.1 <- lm(y ~ x, data=houston))
```

```
##
## Call:
## lm(formula = y ~ x, data = houston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6458 -2.5077 -0.5413  1.7291 11.1484
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.99929    0.85959   3.489 0.000678 ***
## x            0.07147    0.01870   3.823 0.000211 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.918 on 120 degrees of freedom
## Multiple R-squared:  0.1086, Adjusted R-squared:  0.1011
## F-statistic: 14.61 on 1 and 120 DF,  p-value: 0.0002106
```

```
## par(mfrow=c(2,2))
## plot(lm.1)
```

From the summary we see that the coefficient on x is significantly different from zero, which suggests that there is an association between the percentage of low income students and the percentage of students repeating first grade.

Problem 1b.

Has there been an increase in the percentage of students repeating first grade between 1994–1995 and 2004—2005?

Similar to part (a), we regress on the categorical variable `Year` *and check for a significant coefficient.*

```
summary(lm.2 <- lm(y ~ Year, data=houston))
```

```
##
## Call:
## lm(formula = y ~ Year, data = houston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.608 -3.108 -0.377  2.390 11.823
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.3770     0.5253  10.236   <2e-16 ***
## Year2004      1.2311     0.7429   1.657      0.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.103 on 120 degrees of freedom
## Multiple R-squared:  0.02238,    Adjusted R-squared:  0.01423
## F-statistic: 2.747 on 1 and 120 DF,  p-value: 0.1001
```

```
## par(mfrow=c(2,2))
## plot(lm.2)
```

The coefficient on the `Year2004` *dummy variable is not significantly different from zero, so it does not look like there is a real increase in the percentage of students repeating first grade between 1994–1995 and 2004—2005.*

Problem 1c.

Is the association (if any) between the percentage of students repeating first grade and the percentage of low-income students different between 1994–1995 and 2004–2005?

The association is quantified by the slope of the regression line between `y` *and* `x`*. To see if the association changes, we should see if there is a significant difference in the slopes of the regression lines for the two years. We can do this by putting an interaction between* `Year` *and* `x` *in the model, and checking to see if the coefficient on the interaction is significantly different from zero (since that coefficient will be the difference in slopes between the two years.)*

```
summary(lm.3 <- lm(y ~ x*Year, data=houston))
```

```
##
## Call:
## lm(formula = y ~ x * Year, data = houston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.7671 -2.4825 -0.4808  1.7034 11.0178
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.27778    1.25429   2.613   0.0101 *
## x            0.05797    0.03171   1.828   0.0700 .
## Year2004    -0.19035    1.80545  -0.105   0.9162
## x:Year2004   0.01607    0.04048   0.397   0.6921
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.942 on 118 degrees of freedom
## Multiple R-squared:  0.1126, Adjusted R-squared:  0.09006
## F-statistic: 4.992 on 3 and 118 DF,  p-value: 0.002698
```

```
## par(mfrow=c(2,2))
## plot(lm.3)
```

We can see from the summary that the coefficient for the interaction is not significantly different from zero, so the slope doesn't change significantly from 1994 to 2004. Actually the main effect for `Year2004` *is not significant either, so it doesn't even look like the intercept changes between years.*

Problem 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***.

Problem 2a.

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

```
houston04 <- houston[houston$Year==2004,]

attach(houston04)
```

- Number of groups, J. (Each County is a group.)

```
cat("J =",J <- length(unique(County)))
```

```
## J = 8
```

- For each group j, number of observations n_j.

```
nj <- c(table(County))
cat("nj =\n")
```

```
## nj =
```

```
nj
```

```
##   Brazoria   Chambers   FortBend  Galveston     Harris    Liberty Montgomery
##          8          3          5          9         20          7          6
##     Waller
##          3
```

- Total number of observations, n. (Each District is an observation.)

```
cat("n =",n <- sum(nj))
```

```
## n = 61
```

```
detach()
```

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

Problem 2b.

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

$$\left.\begin{array}{rcl} y_i & = & \alpha_{j[i]} + \epsilon_i,\ \epsilon_i \overset{iid}{\sim} N(0,\sigma^2) \\ \alpha_j & = & \beta_0 + \eta_j,\ \eta_j \overset{iid}{\sim} N(0,\tau^2) \end{array}\right\} \tag{1}$$

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:

```
summary(lmer.1 <- lmer(y ~ 1 + (1|County), data=houston04))
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: y ~ 1 + (1 | County)
##    Data: houston04
##
## REML criterion at convergence: 335.8
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.68743 -0.69475 -0.07737  0.48252  2.72625
##
```

```
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  County   (Intercept)  1.028   1.014
##  Residual             14.068   3.751
## Number of obs: 61, groups:  County, 8
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   6.7283     0.6252   10.76
```

We can read these values off the summary:

- $\hat{\sigma}^2 = 14.068$;
- $\hat{\tau}^2 = 1.028$;
- $\hat{\beta}_0 = 6.7283$.

Problem 2c.

Use `library(ggplot2)` (or another R package if you prefer) to make a facet plot of y=Pct.Repeating.1st.Grade vs x=Pct.Low.income.students, 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 (1).

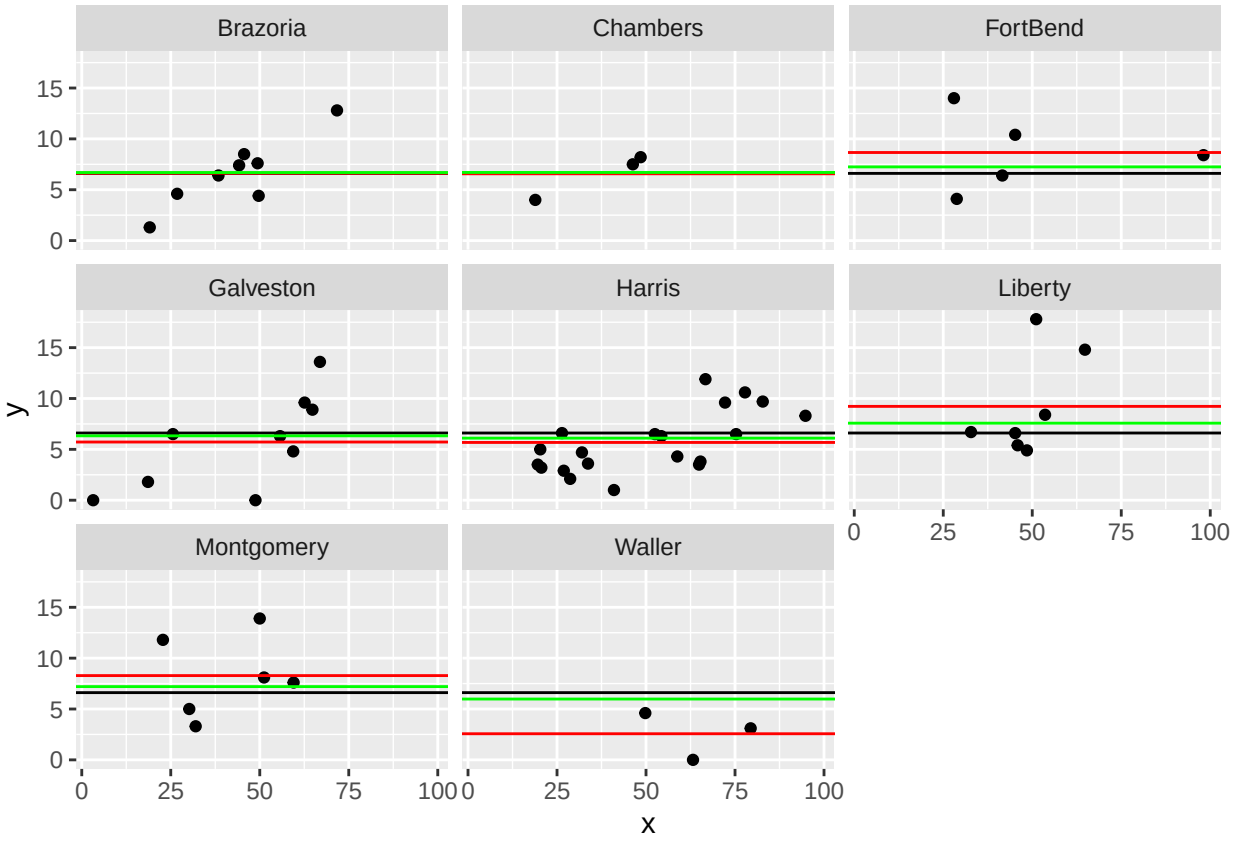
and submit your plot. (Does your plot illustrate the "shrinkage" phenomenon of MLM's?)

Here's the code to make the plot...

```
params <- with(houston04,
                 data.frame(
                    sort(unique(County)),
                    rep(mean(y),J),
                    0,
                    sapply(split(y,County),mean),
                    0,
                    coef(lmer.1)$County[,1], ## gives \beta_0 + \eta_j
                    0
                 )
          )
names(params) = c("County","int0","slo0","int1","slo1","int2","slo2")

g <- ggplot(houston04,aes(x=x,y=y)) +
  facet_wrap( ~ County) +
  geom_point()

g + geom_abline(data=params,aes(intercept=int0,slope=slo0),color="black") +
  geom_abline(data=params,aes(intercept=int1,slope=slo1),color="red") +
  geom_abline(data=params,aes(intercept=int2,slope=slo2),color="green")
```



The facet plot does indeed show the "shrinkage" phenomenon: The fitted county means from the MLM (the green lines) are between the grand mean (the black lines) that would be estimated from `lm(y ~ 1)` *and the raw county means (the red lines) that would be estimated from* `lm(y ~ County)`.

Problem 2d.

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)).

```
cty <- as.factor(houston04$County)
contrasts(cty) <- contr.sum(J)

summary(lm.1.04 <- lm(y ~ cty,data=houston04))
```

```
##
## Call:
## lm(formula = y ~ cty, data = houston04)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.722 -2.529 -0.260  1.875  8.571
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.66656    0.56261  11.849   <2e-16 ***
```

```
## cty1        -0.04156    1.27094  -0.033   0.9740
## cty2        -0.09989    1.94419  -0.051   0.9592
## cty3         1.99344    1.54743   1.288   0.2033
## cty4        -0.94434    1.21284  -0.779   0.4397
## cty5        -0.98656    0.91435  -1.079   0.2855
## cty6         2.56201    1.34195   1.909   0.0617 .
## cty7         1.61678    1.43116   1.130   0.2637
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.722 on 53 degrees of freedom
## Multiple R-squared:  0.1782, Adjusted R-squared:  0.06971
## F-statistic: 1.642 on 7 and 53 DF,  p-value: 0.1439
```

Only one of the `County` *coefficients is even close to significantly different from zero.*

The F statistic, 1.642 on 7 and 53 df, tests the fitted model against the intercept-only model `lm(y ~ 1)` *The p-value is 0.1439, which suggests there is not enough evidence in the data to reject the intercept-only model and keep* `County` *in the model.*

We would get exactly the same result comparing the models directly with the `anova()` *function:*

```
lm.0.04 <- lm(y ~ 1, data=houston04)
anova(lm.0.04,lm.1.04)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ 1
## Model 2: y ~ cty
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     60 893.49
## 2     53 734.23  7    159.25 1.6423 0.1439
```

Problem 3.

In the 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,*

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

we also assume that the ϵ*'s and* η*'s are independent of each other.*

Problem 3a.

We calculate

$$\begin{aligned} \mathrm{Cov}(y_i, y_{i'}) &= \mathrm{Cov}(\beta_0 + \eta_{j[i]} + \epsilon_i, \beta_0 + \eta_{j[i']} + \epsilon_{i'}) && (2) \\ &= \mathrm{Cov}(\eta_{j[i]} + \epsilon_i, \eta_{j[i']} + \epsilon_{i'}) && (3) \\ &= \mathrm{Cov}(\eta_{j[i]}, \eta_{j[i']}) + \mathrm{Cov}(\eta_{j[i]}, \epsilon_{i'}) + \mathrm{Cov}(\epsilon_i, \eta_{j[i']}) + \mathrm{Cov}(\epsilon_i, \epsilon_{i'}) && (4) \\ &= \mathrm{Cov}(\eta_{j[i]}, \eta_{j[i']}) + 0 + 0 + 0 && (5) \end{aligned}$$

where (3) *follows because adding the constant* β_0 *does not change the covariance, and* (5) *follows because the* ϵ*'s are independent of each other and the* η*'s are independent of the* ϵ*'s.*

If $j[i] \neq j[i']$ *then* $\mathrm{Cov}(\eta_{j[i]}, \eta_{j[i']}) = 0$ *too, because different* η*'s are also independent of each other, and so* $\mathrm{Corr}(\eta_{j[i]}, \eta_{j[i']}) = 0$ *also.*

Problem 3b.

If $j[i] = j[i']$ in (4), then the η's in (5) are the same and so

$$\begin{aligned} \mathrm{Cov}(\eta_{j[i]}, \eta_{j[i']}) + 0 + 0 + 0 &= \mathrm{Var}(\eta_{j[i]}) + 0 + 0 + 0 && (6) \\ &= \tau^2 && (7) \end{aligned}$$

Finally, when $j[i] = j[i']$,

$$\begin{aligned} \mathrm{Corr}(y_i, y_{i'}) &= \frac{\mathrm{Cov}(y_i, y_{i'})}{\sqrt{\sigma^2_{y_i}\sigma^2_{y_{i'}}}} && (8) \\ &= \frac{\tau^2}{\sqrt{(\sigma^2+\tau^2)(\sigma^2+\tau^2)}} && (9) \\ &= \frac{\tau^2}{\sigma^2+\tau^2} && (10) \end{aligned}$$

where the numerator in (9) comes from (7) and the terms in the denominator follows from

$$\begin{aligned} \mathrm{Var}(y_i) &= \mathrm{Var}(\beta_0 + \eta_j[i] + \epsilon_i) && (11) \\ &= 0 + \tau^2 + \sigma^2 && (12) \end{aligned}$$

Problem 3c.

$$\begin{aligned} \frac{1}{n_j}\sum_{\{i:j[i]=j\}} Y_i &= \frac{1}{n_j}\left[\sum_{\{i:j[i]=j\}}\beta_0 + \sum_{\{i:j[i]=j\}}\eta_j + \sum_{\{i:j[i]=j\}}\epsilon_i\right] && (13) \\ &= \beta_0 + \eta_j + \frac{1}{n_j}\sum_{\{i:j[i]=j\}}\epsilon_i && (14) \end{aligned}$$

which means

$$\begin{aligned} \mathrm{Var}(\bar{y}_j) &= \mathrm{Var}(\eta_j) + \frac{1}{n_j^2}\sum_{\{i:j[i]=j\}}\mathrm{Var}(\epsilon_i) && (15) \\ &= \tau^2 + \frac{\sigma^2}{n_j} && (16) \end{aligned}$$

Problem 3d.

Using the same kinds of ideas as above, we can calculate

$$\begin{aligned} \mathrm{Var}(\bar{y}_j) &= \tau^2 + \frac{\sigma^2}{n_j} \quad \textit{(from (16))} && (17) \\ \mathrm{Var}(\bar{y}^*_j) &= \tau^2 + \frac{\sigma^2}{n_j} \quad \textit{(from (16) again)} && (18) \\ \mathrm{Cov}(\bar{y}_j, \bar{y}^*_j) &= \mathrm{Cov}\left(\beta_0 + \eta_j + \frac{1}{n_j}\sum_{\{i:j[i]=j\}}\epsilon_i, \beta_0 + \eta_j + \frac{1}{n_j}\sum_{\{i:j[i]=j\}}\epsilon^*_i\right) && (19) \\ &= \tau^2 \quad \textit{(similarly to (7))} && (20) \end{aligned}$$

and therefore

$$\begin{aligned}
\mathrm{Corr}(\bar{y}_j, \bar{y}_j^*) &= \frac{\mathrm{Cov}(\bar{y}_j, \bar{y}_j^*)}{\sqrt{\mathrm{Var}(\bar{y}_j)\,\mathrm{Var}(\bar{y}_j^*)}} && (21)\\
&= \frac{\tau^2}{\sqrt{(\tau^2 + \sigma^2/n_j)(\tau^2 + \sigma^2/n_j)}} && (22)\\
&= \frac{\tau^2}{\tau^2 + \sigma^2/n_j} && (23)
\end{aligned}$$