Homework 6

Tim Vigers

11 April 2019

```
# Categorical variables
cat_vars <- c("ID","current_smoker","race")
# Read in data
inv1 <- read.csv("/Users/timvigers/Documents/School/Biostatistical Methods 2/Ho
meworks/Homework 6/file2_FEV1.csv")
inv1[,c(cat_vars,"trt")] <- lapply(inv1[,c(cat_vars,"trt")], as.factor)
inv2 <- read.csv("/Users/timvigers/Documents/School/Biostatistical Methods 2/Ho
meworks/Homework 6/file1_FEV1.csv")
inv2[,cat_vars] <- lapply(inv2[,cat_vars], as.factor)
inv3 <- inv1
```

1. Fit the models

```
# Models
inv1_mod <- lm(FEV1~pack_years+current_smoker+emphysema+race+height+bmi+trt,dat
a=inv1)
inv2_mod <- lm(delta_FEV1~pre_FEV1+pack_years+current_smoker+emphysema+race+hei
ght+bmi,data=inv2)

# Random intercept
inv3_ri_nlme <- lme(FEV1~pack_years+current_smoker+emphysema+race+height+bmi+tr
t,random = ~1|ID,data=inv3)
inv3_ri_lmer <- lmer(FEV1~pack_years+current_smoker+emphysema+race+height+bmi+t
rt+(1|ID),data=inv3)

# Random intercept and random slope
inv3_ris_nlme <- lme(FEV1~pack_years+current_smoker+emphysema+race+height+bmi+t
rt,random = ~1+trt|ID,data=inv3)
#inv3_ris_lmer <- lmer(FEV1~trt+pack_years+current_smoker+emphysema+race+height
+bmi+(1+trt|ID),data=inv3)
```

Check that nlme, lme4, and SAS match for random intercept model:

```
summary(inv3_ri_nlme)$tTable
```

```
##                       Value    Std.Error  DF    t-value      p-value
## (Intercept)     -3.974321111 0.3768523646 999 -10.546096 9.976179e-25
## pack_years      -0.005940267 0.0008909274 993  -6.667509 4.310627e-11
## current_smoker1  0.126245174 0.0477035097 993   2.646455 8.262725e-03
## emphysema       -0.045342350 0.0021789234 993 -20.809520 4.175029e-80
## race1           -0.228276464 0.0543776596 993  -4.197983 2.934597e-05
## height           0.041328790 0.0021392308 993  19.319463 7.583287e-71
## bmi             -0.013998050 0.0036609811 993  -3.823579 1.397226e-04
## trt1             0.094254000 0.0048051714 999  19.615117 1.007163e-72
```

```
summary(inv3_ri_lmer)$coefficients
```

```
##                     Estimate   Std. Error    t value
## (Intercept)     -3.974321110 0.3768524339 -10.546094
## pack_years      -0.005940267 0.0008909276  -6.667508
## current_smoker1  0.126245174 0.0477035184   2.646454
## emphysema       -0.045342350 0.0021789238 -20.809516
## race1           -0.228276464 0.0543776696  -4.197982
## height           0.041328790 0.0021392312  19.319460
## bmi             -0.013998050 0.0036609818  -3.823578
## trt1             0.094254000 0.0048051706  19.615121
```

Solution for Fixed Effects					
Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	-3.9743	0.3769	994	-10.55	<.0001
pack_years	-0.00594	0.000891	998	-6.67	<.0001
current_smoker	0.1262	0.04770	998	2.65	0.0083
emphysema	-0.04534	0.002179	998	-20.81	<.0001
race	-0.2283	0.05438	998	-4.20	<.0001
height	0.04133	0.002139	998	19.32	<.0001
bmi	-0.01400	0.003661	998	-3.82	0.0001
trt	0.09425	0.004805	998	19.62	<.0001

Check random intercept and random slope model:

lme() fits a model, but lmer() throws an error:

Error: number of observations (=2000) <= number of random effects (=2000) for term (1 + trt | ID); the random-effects parameters and the residual variance (or scale parameter) are probably unidentifiable

The estimates from SAS match those from lme(), but have 0 DF (and therefore no p values) for several variables:

Solution for Fixed Effects					
Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	-3.9477	0.3768	994	-10.48	<.0001
pack_years	-0.00594	0.000891	0	-6.67	.
current_smoker	0.1253	0.04770	0	2.63	.
emphysema	-0.04533	0.002179	0	-20.80	.
race	-0.2253	0.05437	0	-4.14	.
height	0.04118	0.002139	0	19.25	.
bmi	-0.01405	0.003661	0	-3.84	.
trt	0.09425	0.004805	999	19.62	<.0001

```
summary(inv3_ris_nlme)$tTable
```

```
##                       Value    Std.Error  DF    t-value      p-value
## (Intercept)     -3.947704109 0.3768167380 999 -10.476456 1.939769e-24
## pack_years      -0.005941661 0.0008908472 993  -6.669675 4.250185e-11
## current_smoker1  0.125288957 0.0476992128 993   2.626646 8.755887e-03
## emphysema       -0.045327087 0.0021787271 993 -20.804389 4.498000e-80
## race1           -0.225331447 0.0543727616 993  -4.144197 3.700798e-05
## height           0.041177730 0.0021390381 993  19.250583 1.998398e-70
## bmi             -0.014046011 0.0036606514 993  -3.837025 1.323982e-04
## trt1             0.094254000 0.0048051831 999  19.615069 1.007843e-72
```

All of this suggests that the random slope model should not be used.

All model results

Investigator 1 used a linear model to examine the effect of bronchodilator use, pack-years of smoking history, current smoking status, emphysema, race, height and BMI on FEV1. According to this model, emphysema is significantly associated with pre- and post-bronchodilator FEV1 ($p < 0.001$).

```
##
## Call:
## lm(formula = FEV1 ~ pack_years + current_smoker + emphysema +
##     race + height + bmi + trt, data = inv1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.91517 -0.45972 -0.00778  0.46019  2.15681
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -3.9743211  0.2682629 -14.815  < 2e-16 ***
## pack_years      -0.0059403  0.0006333  -9.380  < 2e-16 ***
## current_smoker1  0.1262452  0.0339072   3.723 0.000202 ***
## emphysema       -0.0453423  0.0015488 -29.277  < 2e-16 ***
## race1           -0.2282765  0.0386511  -5.906 4.11e-09 ***
## height           0.0413288  0.0015205  27.180  < 2e-16 ***
## bmi             -0.0139981  0.0026022  -5.379 8.36e-08 ***
## trt1             0.0942540  0.0294878   3.196 0.001413 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6594 on 1992 degrees of freedom
## Multiple R-squared:  0.4967, Adjusted R-squared:  0.4949
## F-statistic: 280.8 on 7 and 1992 DF,  p-value: < 2.2e-16
```

Investigator 2 also used a linear regression, but modeled the change from pre to post bronchodilator FEV1, adjusting for baseline FEV1, pack-years of smoking history, current smoking status, emphysema, race, height and BMI. According to this model, emphysema is significantly associated with change in FEV1 ($p = 0.03$).

```
##
## Call:
## lm(formula = delta_FEV1 ~ pre_FEV1 + pack_years + current_smoker +
##     emphysema + race + height + bmi, data = inv2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.88549 -0.08113 -0.00338  0.08001  0.50470
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -0.4464446  0.0893226  -4.998 6.84e-07 ***
## pre_FEV1        -0.0225718  0.0071413  -3.161  0.00162 **
## pack_years      -0.0001105  0.0002059  -0.536  0.59178
## current_smoker1  0.0190515  0.0108209   1.761  0.07861 .
## emphysema       -0.0012821  0.0005891  -2.176  0.02977 *
## race1           -0.0550522  0.0123825  -4.446 9.74e-06 ***
## height           0.0034924  0.0005619   6.215 7.55e-10 ***
## bmi              0.0004967  0.0008343   0.595  0.55177
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1483 on 992 degrees of freedom
## Multiple R-squared:  0.05366,    Adjusted R-squared:  0.04698
## F-statistic: 8.035 on 7 and 992 DF,  p-value: 1.599e-09
```

Investigator 3 first used a random intercept and random slope model on FEV1 adjusting for bronchodilator use, pack-years of smoking history, current smoking status, emphysema, race, height and BMI. According to this model, emphysema is significantly associated with change in FEV1 ($p < 0.001$), but I wouldn't trust this model.

```
##                       Value    Std.Error  DF    t-value      p-value
## (Intercept)     -3.947704109 0.3768167380 999 -10.476456 1.939769e-24
## pack_years      -0.005941661 0.0008908472 993  -6.669675 4.250185e-11
## current_smoker1  0.125288957 0.0476992128 993   2.626646 8.755887e-03
## emphysema       -0.045327087 0.0021787271 993 -20.804389 4.498000e-80
## race1           -0.225331447 0.0543727616 993  -4.144197 3.700798e-05
## height           0.041177730 0.0021390381 993  19.250583 1.998398e-70
## bmi             -0.014046011 0.0036606514 993  -3.837025 1.323982e-04
## trt1             0.094254000 0.0048051831 999  19.615069 1.007843e-72
```

Investigator 3 then used a random intercept only model on FEV1 adjusting for bronchodilator use, pack-years of smoking history, current smoking status, emphysema, race, height and BMI. According to this model, emphysema is significantly associated with change in FEV1 ($p < 0.001$).

```
##                       Value    Std.Error  DF    t-value      p-value
## (Intercept)     -3.974321111 0.3768523646 999 -10.546096 9.976179e-25
## pack_years      -0.005940267 0.0008909274 993  -6.667509 4.310627e-11
## current_smoker1  0.126245174 0.0477035097 993   2.646455 8.262725e-03
## emphysema       -0.045342350 0.0021789234 993 -20.809520 4.175029e-80
## race1           -0.228276464 0.0543776596 993  -4.197983 2.934597e-05
## height           0.041328790 0.0021392308 993  19.319463 7.583287e-71
## bmi             -0.013998050 0.0036609811 993  -3.823579 1.397226e-04
## trt1             0.094254000 0.0048051714 999  19.615117 1.007163e-72
```

2. Investigator 2 vs. investigators 1 and 3

a.

The model fit by investigator 2 is really asking whether the change in FEV1 is associated with emphysema after adjusting for the base set of confounders, and also after adjusting for baseline FEV1. Investigators 1 and 3 are asking whether both pre- and post-bronchodilator FEV1 are associated with emphysema after adjusting for the base set of confounders, including treatment with a bronchodilator.

b.

The goal of the analysis is to determine if pre- and post-bronchodilator FEV1 are jointly associated with emphysema after adjusting for the base set of confounders. So, investigators 1 and 3 are answering the question correctly.

3. Investigator 3

a.

Investigator 3 should choose the random intercept only model, since the random slope model throws an error when using the lme4 package. The error says that the random effects are unidentifiable, and I think this essentially means that there aren't enough data points to support a random slope model. Also, the SAS output for the model including a random slope doesn't look right, and doesn't provide any p values with which to make inference.

b.

First of all, you can't (or shouldn't) compare models when one of them doesn't converge.

Also, χ^2_2 wouldn't be the correct distribution under the null hypothesis. According to Stram & Lee (1994):

> "The use of likelihood ratio methods for constructing tests for nonzero variance components is a nonstandard problem in the use of maximum likelihood because the null hypothesis, that such a component is zero, places the true value of the variance parameters on the boundary of the parameter space defined by the alternative hypothesis. This has an effect on the large-sample behavior of likelihood ratio tests so that the limiting distribution of -2 times the logarithm of the likelihood ratio (denoted -2 ln λ_N) cannot be treated as that of a χ^2 random variable."

Under H_0, the variance of the random slope is 0 which is the lowest possible variance (since variance can't be negative). And because this is on the boundary of the sample space, the asymptotic behavior of the LRT statistic is a little weird, and can't be treated as χ^2.

4. Best model

I think the random intercept model run by investigator 3 is the best, because it answers the scientific question while accounting for within subject correlation (and the model actually converged). Investigator 1's model answers the correct question, but does not consider the fact that measurements within each subject will be correlated, so the standard errors and p values are probably too small. Finally, investigator 2's model doesn't even answer the correct question.

5. Matrices

a.

$$Z_i = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$$

b.

$$G = \begin{pmatrix} \sigma_0^2 & \sigma_{0s}^2 \\ \sigma_{0s}^2 & \sigma_s^2 \end{pmatrix}$$

c.

$$\text{Var}(Y_i) = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \sigma_0^2 & \sigma_{0s}^2 \\ \sigma_{0s}^2 & \sigma_s^2 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} \sigma_\epsilon^2 & 0 \\ 0 & \sigma_\epsilon^2 \end{pmatrix} =$$

$$\begin{pmatrix} \sigma_0^2 & \sigma_{0s}^2 \\ \sigma_0^2 + \sigma_{0s}^2 & \sigma_{0s}^2 + \sigma_s^2 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} \sigma_\epsilon^2 & 0 \\ 0 & \sigma_\epsilon^2 \end{pmatrix} =$$

$$\begin{pmatrix} \sigma_0^2 & \sigma_0^2 + \sigma_{0s}^2 \\ \sigma_0^2 + \sigma_{0s}^2 & \sigma_0^2 + 2\sigma_{0s}^2 + \sigma_s^2 \end{pmatrix} + \begin{pmatrix} \sigma_\epsilon^2 & 0 \\ 0 & \sigma_\epsilon^2 \end{pmatrix} =$$

$$\begin{pmatrix} \sigma_0^2 + \sigma_\epsilon^2 & \sigma_0^2 + \sigma_{0s}^2 \\ \sigma_0^2 + \sigma_{0s}^2 & \sigma_0^2 + 2\sigma_{0s}^2 + \sigma_s^2 + \sigma_\epsilon^2 \end{pmatrix}$$

This matrix has 4 different parameters that need to be estimated (σ_0^2,σ_s^2,σ_{0s}^2, and σ_ϵ^2), but because there are only two timepoints, there isn't enough data to support a random intercept and slope model. There are only two repeated measurements, so from this data we can only estimate 3 covariance parameters.