linear_mixed_effects_model

Andrew Willems and Tian Hong

3/22/2022

Linear Mixed Effects Model Analysis

In this analysis I seek to determine two things. One, are the variables Condition, Map ID, or Cohort independent or not. Two, if any of these variables are not independent what does a linear mixed effects model look like when we account for those non-independent variables.

Step One: Load Packages

If you do not have any of these packages they can be installed with the following command install.packages() in an R console

```
#Loading needed packages----
library(emmeans)
library(gt)
library(ICC)
library(nlme)
library(modelsummary)
```

```
##
## Attaching package: 'modelsummary'
```

```
## The following object is masked from 'package:gt':
##
##     escape_latex
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::collapse() masks nlme::collapse()
## x dplyr::filter()   masks stats::filter()
## x dplyr::lag()      masks stats::lag()
```

Step Two: Load in the data and filter it into separate data frames containg only left or right hemisphere data

```
#Loading the data----
setwd("~/Documents/PhD Program/Hong Lab/Projects/Neuron_Project/Data/")
pnn_data <- read.csv("processed_data_PNN_3cohorts.csv", sep = "\t")
```

```
pnn_data_left <- filter(pnn_data, Hemisphere=="Left")
pnn_data_right <- filter(pnn_data, Hemisphere=="Right")
```

Step Three: Begin Left Hemisphere analysis. We start be using the Intraclass correlation coefficient (ICC) to determine if the variables Condition, Map ID, or Cohort are truly independent or not. I use the ICCbare command with x representing the grouping variable of interest and y representing the Mean 2 variable

```
#Left hemisphere analysis----
#Intraclass correlation coefficient (ICC)
#For Condition
icc_cond=ICCbare(x=factor(Condition), y= Mean.2,data = pnn_data_left)
#For Map ID
icc_map=ICCbare(x=factor(Map.ID), y= Mean.2,data = pnn_data_left)
#For Cohort
icc_cohort=ICCbare(x=factor(Cohort), y= Mean.2,data = pnn_data_left)

#Making a data frame
icc_df <- data.frame(Condition=icc_cond, Cohort=icc_cohort, Map=icc_map)


#Making that data frame a nicer looking table
gt_icc <- gt(icc_df)
gt_icc %>% tab_header(
  title = "ICC for Left Hemisphere",
  subtitle = "Intraclass Correlation Coefficient (ICC) for left hemisphere PNN data."
 ) %>% cols_align(
  align = "center",
  columns = c("Condition", "Cohort", "Map")
 )
```

ICC for Left Hemisphere

Intraclass Correlation Coefficient (ICC) for left hemisphere PNN data.

Condition	Cohort	Map
0.1608042	0.6676902	0.09202628

Step 4: The Cohort variable has an ICC value (0.668) that indicates it is not independent and therefore we should build a linear-mixed-effects model to account for this. The other two variables have negligable ICC values and do not be considered as random effects in our model

```
#The linear-mixed-effects model
#The fixed effect is the Condition variable. The random variable is Cohort
lme_mod=lme(Mean.2~Condition, data= pnn_data_left, random = ~ 1|Cohort)

modelsummary(
  lme_mod,
  fmt = 1,
  estimate  = "{estimate} {stars} [{conf.low}, {conf.high}]",
```

	Model 1
(Intercept)	69.5 *** [59.2, 79.8]
NW	−0.5 [−2.6, 1.5]
SH	5.3 *** [3.3, 7.4]
SW	−4.4 *** [−6.5, −2.4]
SD (Intercept)	9.0
SD (Observations)	5.5
Num.Obs.	222
R2 Marg.	0.293
R2 Cond.	
AIC	1399.9
BIC	1420.2
RMSE	5.42

```
  statistic = NULL,
  coef_rename = c("ConditionNW" = "NW", "ConditionSH" = "SH", "ConditionSW" = "SW"))
```

```
## Random effect variances not available. Returned R2 does not account for random effects.
```

Step 5: Perform an overall model Wald test. Here we see that the when comparing our intercept-only model to our predictor-included model we have strong statistical support for significance (p-value < 0.0001)

```
#ANOVA Wald test for overall model p-value
wald_test <- anova(lme_mod)
wald_test$terms <- rownames(wald_test)
gt(wald_test)
```

numDF	denDF	F-value	p-value	terms
1	216	178.43055	0.000000e+00	(Intercept)
3	216	30.32082	2.220446e-16	Condition

Step 6: Perform an overall model likelihood ratio (LRT) test. Here again we see that we have strong support for concluding that our predictor-included model better fits the data with a p-value <0.0001. Our Wald and LRT stats are similar and this further supports our claim that our model with Condition as a predictor of Mean 2 is well-supported. Finally, LRT is considered to be prefereable to Wald which is another point in our favor.

```
#ANOVA likelihood ratio (LRT) test for overall model p-value
#Here we are comparing the intercept only model with the model that includes
#the predictor of Condition
base_lme_mod=lme(Mean.2~1, data= pnn_data_left,
                 random = ~ 1|Cohort, method="ML")
pred_lme_mod=lme(Mean.2~Condition, data= pnn_data_left,
                 random = ~ 1|Cohort, method="ML")

lrt_test <- anova(base_lme_mod, pred_lme_mod)

lrt_test_df <- data.frame(lrt_test$Model, lrt_test$df, lrt_test$AIC,
                          lrt_test$BIC, lrt_test$logLik, lrt_test$Test,
                          lrt_test$L.Ratio, lrt_test$`p-value`[2])

colnames(lrt_test_df) <- c("Model", "df", "AIC", "BIC", "logLik", "Test",
                           "L.Ratio", "P-value")
```

```
gt_lrt %>% tab_header(
  title = "LRT for Left Hemisphere",
  subtitle = "LRT for left hemisphere PNN data."
 ) %>% cols_align(
  align = "center",
  columns = c("Model","df","AIC","BIC","logLik", "Test","L.Ratio", "P-value")
 )
```

LRT for Left Hemisphere

LRT for left hemisphere PNN data.

Model	df	AIC	BIC	logLik	Test	L.Ratio	P-value
1	3	1480.919	1491.127	-737.4596		NA	1.355982e-16
2	6	1409.928	1430.344	-698.9639	1 vs 2	76.99127	1.355982e-16

Step 8: Doing a pairwise comparison of our various conditions. Note that these p-values are adjusted with the tukey method. From the results we see that the NH-NW comparison is not signficant while all other comparisons are less than 0.05 and many are less than 0.0001.

```
#Doing pairwise adjusted-p-value comparison
left_hemi_pairwise <- contrast(emmeans(lme_mod, specs="Condition"), "pairwise")
left_hemi_pairwise <- as.data.frame(left_hemi_pairwise)
colnames(left_hemi_pairwise) <- c("Contrast", "Estimate", "SE", "df",
                                  "T-ratio", "P-value")
gt_left_hemi_pw <- gt(left_hemi_pairwise)
gt_left_hemi_pw %>% tab_header(
  title = "Left Hemisphere Pairwise Analysis"
) %>% cols_align(
  align = "center",
  columns = c("Contrast", "Estimate", "SE", "df", "T-ratio", "P-value")
)
```

Left Hemisphere Pairwise Analysis

Contrast	Estimate	SE	df	T-ratio	P-value
NH - NW	0.5489036	1.062466	216	0.5166319	9.550444e-01
NH - SH	-5.3248395	1.046404	216	-5.0887019	4.653225e-06
NH - SW	4.4302806	1.029673	216	4.3026100	1.494768e-04
NW - SH	-5.8737432	1.062220	216	-5.5296883	5.502432e-07
NW - SW	3.8813770	1.045400	216	3.7128163	1.479054e-03
SH - SW	9.7551202	1.027465	216	9.4943620	0.000000e+00

Step 9: Comparingn NW to all other treatment conditions (i.e. treatment vs. control). Note that these p-values are adjusted with the dunnettx method (similar to Tukey). We see that in the SH-NW comparison we have a p-value <0.0001. We also have a statistically significanct difference between SW-NW. We might want to investigate why our two wildtypes are different from one another because this also holds true in the pairwise analysis. From this step and step 8 we can confidently conclude that there are differences between the NW and the other groups being examined in the left hemisphere for Mean 2 expression.

```
#Doing treatment vs. control p-value comparison (Specifying NW as control)
treat_v_control <- contrast(emmeans(lme_mod, specs="Condition"), "trt.vs.ctrl", ref=2)
treat_v_control <- as.data.frame(treat_v_control)
colnames(treat_v_control) <- c("Contrast", "Estimate", "SE", "df",
                               "T-ratio", "P-value")


gt_left_hemi_tc <- gt(treat_v_control)
gt_left_hemi_tc %>% tab_header(
  title = "Left Hemisphere Treatment vs. Control Analysis"
) %>% cols_align(
  align = "center",
  columns = c("Contrast", "Estimate", "SE", "df", "T-ratio", "P-value")
)
```

Left Hemisphere Treatment vs. Control Analysis

Contrast	Estimate	SE	df	T-ratio	P-value
NH - NW	0.5489036	1.062466	216	0.5166319	8.884827e-01
SH - NW	5.8737432	1.062220	216	5.5296883	2.759778e-07
SW - NW	-3.8813770	1.045400	216	-3.7128163	7.675654e-04

Step 10: Constructing 95% Confidence intervals. We construct 95% confidence intervals to further support our analysis as this is now becoming a standard value to report in literature to help with rigor and reproducibility.

```
#Constructing 95% Confidence intervals
left_ci <- contrast(emmeans(lme_mod, specs="Condition"),
         "pairwise") %>% summary(infer = TRUE)

left_ci <- as.data.frame(left_ci)
colnames(left_ci) <- c("Contrast", "Estimate", "SE", "df", "Lower CL",
                       "Upper CL","T-ratio", "P-value")

left_ci <- gt(left_ci)
left_ci %>% tab_header(
  title = "Left Hemisphere 95% Confidence Interval"
) %>% cols_align(
  align = "center",
  columns = c("Contrast", "Estimate", "SE", "df", "Lower CL",
                       "Upper CL","T-ratio", "P-value")
)
```

Left Hemisphere 95% Confidence Interval

Contrast	Estimate	SE	df	Lower CL	Upper CL	T-ratio	P-value
NH - NW	0.5489036	1.062466	216	-2.201971	3.299778	0.5166319	9.550444e-01
NH - SH	-5.3248395	1.046404	216	-8.034129	-2.615550	-5.0887019	4.653225e-06
NH - SW	4.4302806	1.029673	216	1.764311	7.096250	4.3026100	1.494768e-04
NW - SH	-5.8737432	1.062220	216	-8.623981	-3.123505	-5.5296883	5.502432e-07
NW - SW	3.8813770	1.045400	216	1.174688	6.588066	3.7128163	1.479054e-03
SH - SW	9.7551202	1.027465	216	7.094868	12.415372	9.4943620	0.000000e+00

Step 11: Doing all of the same anlyses we did above but on right hemisphere data

```
#Right hemisphere analysis----
#Intraclass correlation coefficient (ICC)
icc_cond=ICCbare(x=factor(Condition), y= Mean.2,data = pnn_data_right)
icc_map=ICCbare(x=factor(Map.ID), y= Mean.2,data = pnn_data_right)
icc_cohort=ICCbare(x=factor(Cohort), y= Mean.2,data = pnn_data_right)


#Making a data frame
icc_df <- data.frame(Condition=icc_cond, Cohort=icc_cohort, Map=icc_map)


#Making that data frame a nicer looking table
gt_icc <- gt(icc_df)
gt_icc %>% tab_header(
  title = "ICC for Right Hemisphere",
  subtitle = "Intraclass Correlation Coefficient (ICC) for right hemisphere PNN data."
 ) %>% cols_align(
  align = "center",
  columns = c("Condition", "Cohort", "Map")
 )
```

ICC for Right Hemisphere

Intraclass Correlation Coefficient (ICC) for right hemisphere PNN data.

Condition	Cohort	Map
0.1191752	0.5796049	0.04871027

Right hemisphere ICC conclusion: We see that we have 0.58 ICC for the cohort variable. Given this, we have moderate correlation among the measurements from this variable and we should consider it as a random effect in a mixed effects model. Finally, we examine the Map variable it has an ICC of only 0.049 which indicates there is very little correlation among measurements for this variable and does not need to be considered as a random effect in our model.

```
#The linear-mixed-effects model
lme_mod=lme(Mean.2~Condition, data= pnn_data_right, random = ~ 1|Cohort)

modelsummary(
```

	Model 1
(Intercept)	69.3 *** [60.4, 78.1]
NW	−0.8 [−3.0, 1.5]
SH	4.3 *** [2.1, 6.5]
SW	−3.2 ** [−5.5, −1.0]
SD (Intercept)	7.6
SD (Observations)	5.9
Num.Obs.	227
R2 Marg.	0.183
R2 Cond.	
AIC	1463.6
BIC	1484.0
RMSE	5.84

```
  lme_mod,
  fmt = 1,
  estimate  = "{estimate} {stars} [{conf.low}, {conf.high}]",
  statistic = NULL,
  coef_rename = c("ConditionNW" = "NW", "ConditionSH" = "SH", "ConditionSW" = "SW"))
```

```
## Random effect variances not available. Returned R2 does not account for random effects.
```

The Wald test shows us that our Condition-included model better fits our data than our intercept-only model with a p-value <0.0001

```
#ANOVA Wald test for overall model p-value
wald_test <- anova(lme_mod)
wald_test$terms <- rownames(wald_test)
gt(wald_test)
```

numDF	denDF	F-value	p-value	terms
1	221	246.20329	0.000000e+00	(Intercept)
3	221	16.84328	6.874759e-10	Condition

LRT shows the same thing with p-value <0.0001. Given that this is prefered to Wald this further supports our claim that using Condition better models the data we observe and better captures the relationship to Mean 2 than the intercept-only model

```
#ANOVA likelihood ratio (LRT) test for overall model p-value
base_lme_mod=lme(Mean.2~1, data= pnn_data_right,
                 random = ~ 1|Cohort, method="ML")
pred_lme_mod=lme(Mean.2~Condition, data= pnn_data_right,
                 random = ~ 1|Cohort, method="ML")

lrt_test <- anova(base_lme_mod, pred_lme_mod)

lrt_test_df <- data.frame(lrt_test$Model, lrt_test$df, lrt_test$AIC,
                          lrt_test$BIC, lrt_test$logLik, lrt_test$Test,
                          lrt_test$L.Ratio, lrt_test$`p-value`[2])

colnames(lrt_test_df) <- c("Model", "df", "AIC", "BIC", "logLik", "Test",
                           "L.Ratio", "P-value")

gt_lrt <- gt(lrt_test_df)
gt_lrt
```

```
  subtitle = "LRT for right hemisphere PNN data."
 ) %>% cols_align(
  align = "center",
  columns = c("Model","df","AIC","BIC","logLik", "Test","L.Ratio", "P-value")
 )
```

LRT for Right Hemisphere

LRT for right hemisphere PNN data.

Model	df	AIC	BIC	logLik	Test	L.Ratio	P-value
1	3	1513.770	1524.045	-753.8850		NA	5.27968e-10
2	6	1473.624	1494.173	-730.8119	1 vs 2	46.14622	5.27968e-10

These p-values are adjusted with the Tukey method. The NH-NW comparison is not signficant, and the NW-SW comparison is not signficanct. All other comparisons are less than 0.05 and some are less than 0.0001. This indicates that certain conditions are highly different from others and their values of Mean 2

```
#Doing pairwise adjusted p-value comparison
right_hemi_pairwise <- contrast(emmeans(lme_mod, specs="Condition"), "pairwise")
right_hemi_pairwise <- as.data.frame(right_hemi_pairwise)
colnames(right_hemi_pairwise) <- c("Contrast", "Estimate", "SE", "df",
                                   "T-ratio", "P-value")
gt_right_hemi_pw <- gt(right_hemi_pairwise)
gt_right_hemi_pw %>% tab_header(
  title = "Right Hemisphere Pairwise Analysis"
) %>% cols_align(
  align = "center",
  columns = c("Contrast", "Estimate", "SE", "df", "T-ratio", "P-value")
)
```

Right Hemisphere Pairwise Analysis

Contrast	Estimate	SE	df	T-ratio	P-value
NH - NW	0.7762686	1.141615	221	0.6799739	9.046244e-01
NH - SH	-4.2935451	1.119521	221	-3.8351614	9.358582e-04
NH - SW	3.2407663	1.141462	221	2.8391364	2.530779e-02
NW - SH	-5.0698137	1.085844	221	-4.6690056	3.100586e-05
NW - SW	2.4644977	1.108517	221	2.2232389	1.201444e-01
SH - SW	7.5343114	1.086036	221	6.9374394	2.614634e-10

These p-values are adjusted with the dunnettx method (similar to Tukey). We see that only SH is statistically signficant when comparing to the control of NW.

```
#Doing treatment vs. control p-value comparison (Specifying NW as control)
treat_v_control <- contrast(emmeans(lme_mod, specs="Condition"), "trt.vs.ctrl", ref=2)
treat_v_control <- as.data.frame(treat_v_control)
colnames(treat_v_control) <- c("Contrast", "Estimate", "SE", "df",
                               "T-ratio", "P-value")
```

```
gt_right_hemi_tc <- gt(treat_v_control)
gt_right_hemi_tc %>% tab_header(
  title = "Right Hemisphere Treatment vs. Control Analysis"
) %>% cols_align(
  align = "center",
  columns = c("Contrast", "Estimate", "SE", "df", "T-ratio", "P-value")
)
```

Right Hemisphere Treatment vs. Control Analysis

Contrast	Estimate	SE	df	T-ratio	P-value
NH - NW	0.7762686	1.141615	221	0.6799739	8.079062e-01
SH - NW	5.0698137	1.085844	221	4.6690056	1.566667e-05
SW - NW	-2.4644977	1.108517	221	-2.2232389	7.281333e-02

Constructing 95% confidence intervals to further improve rigor and reproducibility

```
#Constructing 95% Confidence intervals
right_ci <- contrast(emmeans(lme_mod, specs="Condition"),
         "pairwise") %>% summary(infer = TRUE)

right_ci <- as.data.frame(right_ci)
colnames(right_ci) <- c("Contrast", "Estimate", "SE", "df", "Lower CL",
                        "Upper CL","T-ratio", "P-value")

right_ci <- gt(right_ci)
right_ci %>% tab_header(
  title = "Right Hemisphere 95% Confidence Interval"
) %>% cols_align(
  align = "center",
  columns = c("Contrast", "Estimate", "SE", "df", "Lower CL",
                        "Upper CL","T-ratio", "P-value")
)
```

Right Hemisphere 95% Confidence Interval

Contrast	Estimate	SE	df	Lower CL	Upper CL	T-ratio	P-value
NH - NW	0.7762686	1.141615	221	-2.1790135	3.731551	0.6799739	9.046244e-01
NH - SH	-4.2935451	1.119521	221	-7.1916330	-1.395457	-3.8351614	9.358582e-04
NH - SW	3.2407663	1.141462	221	0.2858809	6.195652	2.8391364	2.530779e-02
NW - SH	-5.0698137	1.085844	221	-7.8807227	-2.258905	-4.6690056	3.100586e-05
NW - SW	2.4644977	1.108517	221	-0.4051029	5.334098	2.2232389	1.201444e-01
SH - SW	7.5343114	1.086036	221	4.7229056	10.345717	6.9374394	2.614634e-10

Conclusion: The Cohort variable is not indepedent and therefore should be considered as a random effect in a mixed effects model. We see that when using this type of model we still see statistically signficant differences for many of the different conditions as outlined in the written sections above. It appears that we have strong support so far that SH mice are indeed different from NW whether we look at either left or right hemisphere regardless of if we are doing 'treatment vs. control' or 'pairwise' comparisons for mean expression of PNN. We should test to see if this relationship holds up in other cohorts for future directions.