A Tutorial for Linear Models and Linear Mixed Effects Models in R

Experimental Phonetics

Ryu, Hyuksu

Naver Clova

Oct. 21, 2017

T.O.C

1. Instruction
2. Example Description
3. Exercise 1: Pitch $\sim$ Sex
4. Exercise 2: Pitch $\sim$ Age
5. Assumption

Outline

1. Instruction
2. Example Description
3. Exercise 1: Pitch $\sim$ Sex
4. Exercise 2: Pitch $\sim$ Age
5. Assumption

Before beginning

Source of this tutorial

- Bodo Winter hompage
  `http://www.bodowinter.com/tutorials.html`
- Linear models and linear mixed effects models in R
  `http://arxiv.org/pdf/1308.5499.pdf`

Citation

- Winter, B. (2013). Linear models and linear mixed effects models in R with linguistic applications. arXiv:1308.5499.

How to download this tutorial

- github page
  `https://github.com/HyuksuRyu/mixedLM_tutorial`

Instruction

What we are dealing with?

1. Linear model ←
2. Linear mixed model

Outline

1. Instruction
2. **Example Description**
3. Exercise 1: Pitch $\sim$ Sex
4. Exercise 2: Pitch $\sim$ Age
5. Assumption

Example description

Question

- Assume you knew nothing about males and females
- you were interested in whether voice pitch of males and females differs, if so, by how much

Experiment

- take a bunch of males and females
- ask them to say a single word
- measure the respective voice pitches

Example description

Subject	Sex	Voice.Pitch(Hz)
1	female	233
2	female	204
3	female	242
4	male	130
5	male	112
6	male	142

Example description

It looks like

- the female values seem to be about 100 Hz above the male ones
- females have higher voice pitch than males

But

- it could be the case that females and males have the same pitch
- you were just unlucky and happened to choose some exceptionally high-pitched females and some exceptionally low-pitched males

Example description

We might want

- a more precise estimate of the difference between males and females
- an estimate about how likely (or unlikely) that difference in voice pitch could have arisen just because of drawing an unlucky sample

→ The linear model comes in

- give some value about voice pitch for males and females
- as well as some probability values as to how likely those values are

Example description

Basic idea

- relationship b/w sex and voice pitch as a simple formula
- `pitch ~ sex`
- This reads
  - pitch predicted by sex
  - pitch as a function of sex
- LEFT: `pitch`
  - response (dependent) variable
  - the thing you measure
- RIGHT: `sex`
  - explanaatory (independent) variable
  - predictor

Example Description

Error term

- Problem: the world is not perfect.
- Pitch is not *completely* determined by sex
- a bunch of different factors such as language, dialect, etc
- We can never measure and control all of these things
- update our formula to capture the existence of these "random" factors.
- `pitch ~ sex` $+ \epsilon$
- ϵ
  - an error term
  - stands for all of the things that affect pitch that are not sex,
  - all of the stuff that is random or uncontrollable

Outline

1. Instruction
2. Example Description
3. **Exercise 1: Pitch $\sim$ Sex**
4. Exercise 2: Pitch $\sim$ Age
5. Assumption

Exercise

Let's create the dataset

```
pitch = c(233,204,242,130,112,142)
sex = c(rep("female",3),rep("male",3))
my.df = data.frame(sex, pitch)
```

```
my.df

##      sex pitch
## 1 female   233
## 2 female   204
## 3 female   242
## 4   male   130
## 5   male   112
## 6   male   142
```

Exercise

Let's create the dataset

```
pitch = c(233,204,242,130,112,142)
sex = c(rep("female",3),rep("male",3))
my.df = data.frame(sex, pitch)
```

```
my.df
```

```
##      sex pitch
## 1 female   233
## 2 female   204
## 3 female   242
## 4   male   130
## 5   male   112
## 6   male   142
```

Exercise

- `lm()`

```
xmdl = lm(pitch ~ sex, my.df)
```

  - Generate the linear model
  - Note that we omit the “ϵ” term
  - we saved the model into an object `xmdl`

Exercise

- summary()
  - To see what the linear model did

```
summary(xmdl)
```

Exercise

```
##
## Call:
## lm(formula = pitch ˜ sex, data = my.df)
##
## Residuals:
##       1       2       3       4       5       6
##   6.667 -22.333  15.667   2.000 -16.000  14.000
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   226.33      10.18  22.224 2.43e-05 ***
## sexmale       -98.33      14.40  -6.827  0.00241 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.64 on 4 degrees of freedom
## Multiple R-squared:  0.921,Adjusted R-squared:  0.9012
## F-statistic: 46.61 on 1 and 4 DF,  p-value: 0.002407
```

← the model formula you entered

← the residuals

← the coefficients of the fixed effects

← the output prints some overall results of the model

Output

Multiple R-Squared

- R^2
  - variance explained
  - 0.921 (quite high) means 92.1% of the stuff is "explained" by our model
- In this case, we have only one independent variable (the fixed effect "sex"),
- R^2 reflects how much variance in our data is accounted for by differences b/w males and females

Output

Multiple R-Squared and Adjusted R-squared

- If you have two or more variables (fixed effect), you see Adjusted R-squared (R^2_{adj}), instead of Multiple R-squared.
- R^2 has a property that it is increased when variables are added up, even though the variables are irrelevant.
- "Adjusted" R-squared are adjusted considering how many fixed effects you used.

Output

Meaning of p-value in the F-statistics

- **conditional probability**
  - a probability *under the condition that the* H_0 *is true*
- H_0: Explanary variables (sex) have no effect on the response ($pitch$)
- "statistically significant" when the conditional probability is lower than a threshold, and the alternative hypothesis (H_1) is more likely.
- Report
  - "We constructed a linear model of pitch as a function of sex. This model was significnat (F(1,4)=46.61, p<0.01)."
- Significance of the overall model
- How to distinguish b/w the significance of the overall model (considering all effects together) from the p-value of individual coefficients? → to be continue

Output

Coefficients table

- Significance of the individual coefficients

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 226.33333   10.18441 22.223508 2.426952e-05
## sexmale     -98.33333   14.40293 -6.827314 2.406892e-03
```

- Why `sexmale`, rather than just `sex`?
- Estimate of `intercept`: mean of pitch of female
- Estimate of `sexmale`: difference of pitch b/w female and male
  $\therefore$ Pitch(male) $=$ `intercept` $+$ `sexmale`

	female	male	difference
mean	226.333	128	98.333

Tips

Tip 1 - calculate mean in terms of sex

1. Classic

```
with(my.df, mean(pitch[sex=='female']))

## [1] 226.3333

with(my.df, mean(pitch[sex=='male']))

## [1] 128
```

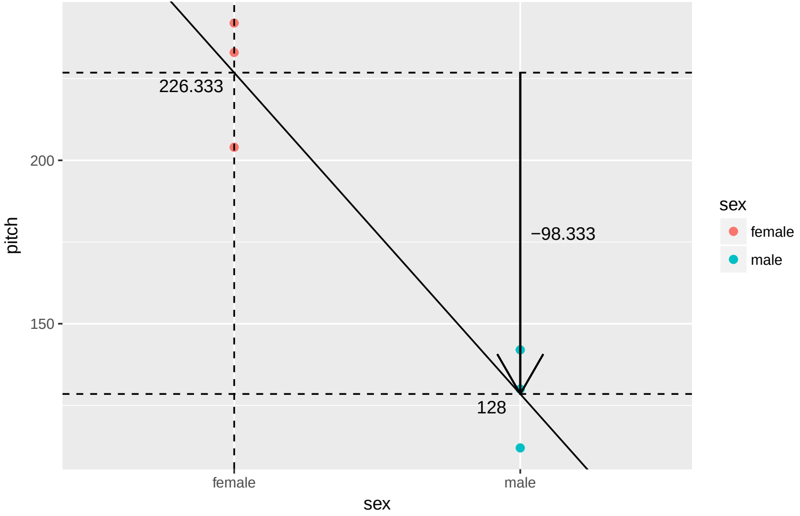
2. Using dplyr

```
library(dplyr)
my.df %>%
  group_by(sex) %>%
  summarise(mean=mean(pitch))

## # A tibble: 2 x 2
##        sex     mean
##     <fctr>    <dbl>
## 1 female 226.3333
## 2   male 128.0000
```

Output

Graphical interpretation



Output

Secret of Intercept: Why did the model choose females to be the intercept?

∵ `lm()` function simply takes whatever comes first in alphabet
"f" comes before "m"

Output

Secret of Intercept: Why did the model choose females to be the intercept?

$\because$ `lm()` function simply takes whatever comes first in alphabet
“f” comes before “m”

Tips

Tip 2 - How to change order of coefficients' level

```
my.df$sex = factor(my.df$sex, levels=c("male", "female"))
```

- Before (order: female-male)

```
##             Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 226.33333   10.18441 22.223508 2.426952e-05
## sexmale     -98.33333   14.40293 -6.827314 2.406892e-03
```

- After (order: male-female)

```
##             Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 128.00000   10.18441 12.568228 0.0002306453
## sexfemale    98.33333   14.40293  6.827314 0.0024068918
```

Outline

1. Instruction
2. Example Description
3. Exercise 1: Pitch ∼ Sex
4. **Exercise 2: Pitch ∼ Age**
5. Assumption

New model

Whether age predicts voice pitch

- continuous as explanatory
- pitch $\sim$ age $+ \epsilon$

Subject	Age	Pitch(Hz)
1	14	252
2	23	244
3	35	240
4	48	233
5	52	212
6	67	204

Output

```
age = c(14,23,35,48,52,67)
pitch = c(252,244,240,233,212,204)
my.df = data.frame(age, pitch)
xmdl = lm(pitch~age, my.df)
summary(xmdl)
```

```
## 
## Call:
## lm(formula = pitch ~ age, data = my.df)
## 
## Residuals:
##      1      2      3      4      5      6 
## -2.338 -2.149  4.769  9.597 -7.763 -2.115 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 267.0765     6.8522   38.98 2.59e-06 ***
## age          -0.9099     0.1569   -5.80  0.00439 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 6.886 on 4 degrees of freedom
## Multiple R-squared:  0.8937,Adjusted R-squared:  0.8672 
## F-statistic: 33.64 on 1 and 4 DF,  p-value: 0.004395
```

Output

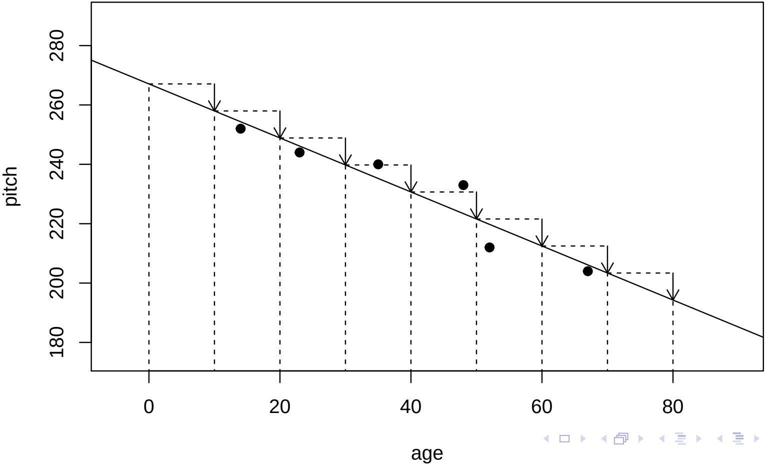
Coefficient table

```
##                Estimate Std. Error  t value     Pr(>|t|)
## (Intercept) 267.0764640  6.8521942 38.976780 2.588356e-06
## age          -0.9098694  0.1568771 -5.799888 4.394969e-03
```

- the significance of the intercept is NOT important
    - intercept means predicted pitch for people with age 0
- the significance of the age IS real interest.
    - every increase of age by 1 → decrease voice pitch by 0.9099

Output

Graphical interpretation



Output

Meaningful and meaningless intercepts

```
my.df$age.c = my.df$age - mean(my.df$age)
xmdl = lm(pitch~age.c, my.df)
```

- new column “age.c” → “centered” data

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 230.8333333  2.8112916 82.109353 1.318716e-07
## age.c        -0.9098694  0.1568771 -5.799888 4.394969e-03
```

- Same slope, but different intercept
- intercept here means pitch at mean age → mean pitch
- intercept becomes more meaningful than previous

Tips

Tip 3 - Manipulating data frame and add to a new column

1. Classic

```
my.df$age.c = my.df$age - mean(my.df$age)
```

2. Using dplyr

```
my.df %>%
  mutate(age.c = age-mean(age))

##   age pitch      age.c
## 1  14   252 -25.833333
## 2  23   244 -16.833333
## 3  35   240  -4.833333
## 4  48   233   8.166667
## 5  52   212  12.166667
## 6  67   204  27.166667
```

Multiple regression

Scaling up

- What if we measured two factors, such as age and sex?
- Multiple regression
    - a function of multiple predictor variables

$$pitch \sim sex + age + \epsilon$$

- But, same interpretation
    - Significance of overall model
    - Significance of each coefficient

Multiple regression

Example

```
##
## Call:
## lm(formula = pitch ~ sex + age, data = my.df)
##
## Residuals:
##        1        2        3        4        5        6
##   -2.218   -7.103    9.321   13.846  -11.769   -2.077
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   202.2179    13.1658  15.359 0.000599 ***
## sexmale      -100.6603    10.3476  -9.728 0.002307 **
## age             0.6346     0.2887   2.198 0.115376
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.61 on 3 degrees of freedom
## Multiple R-squared:  0.9697,Adjusted R-squared:  0.9495
## F-statistic: 48.05 on 2 and 3 DF,  p-value: 0.005268
```

The end of Linear model

The end of linear model!
But one more thing!

Outline

1. Instruction
2. Example Description
3. Exercise 1: Pitch $\sim$ Sex
4. Exercise 2: Pitch $\sim$ Age
5. **Assumption**

Assumptions for LM

Assumptions for applying a linear model

1. Linearity
2. Absense of collinearity
3. Homoskedasticity
4. Normality of residuals
5. Absense of influential data points
6. Independence

Linearity

Linearity?

- linear what?
- linearity of residuals
- So, what is residuals?

Linearity

Linearity?

- linear what?
- linearity of residuals
- So, what is residuals?

Linearity

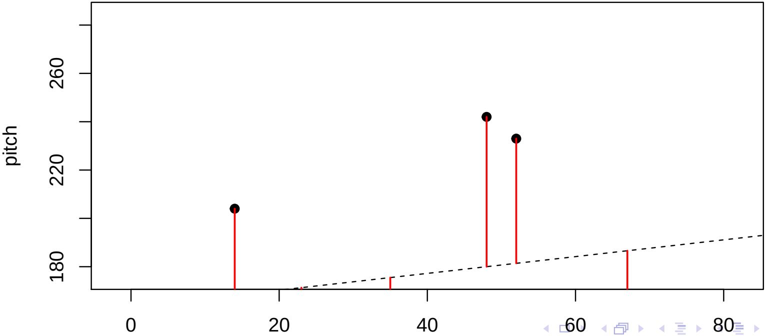
Linearity?

- linear what?
- linearity of residuals
- So, what is residuals?

Linearity

Residuals

- Deviations of the observed data points from the predicted values (fitted values)
- In this case, the residuals very small → well predicted

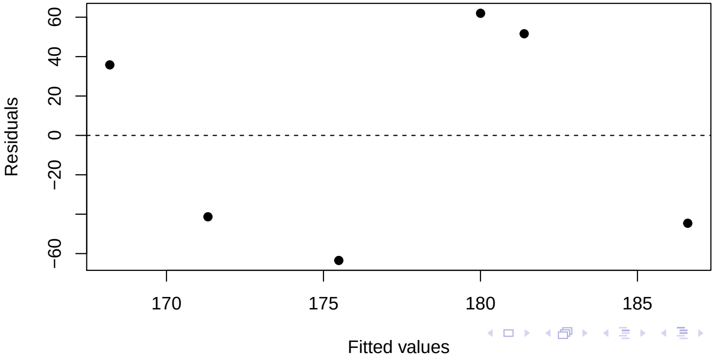


Linearity

Residuals

- Rotate the plot → Residual plot

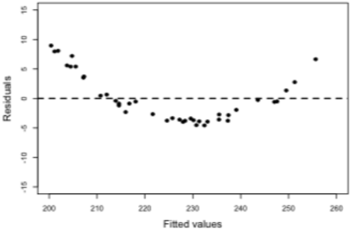
```r
plot(fitted(xmdl), residuals(xmdl),pch=19,
          xlab = 'Fitted values', ylab='Residuals' )
```



Linearity

Residual plot

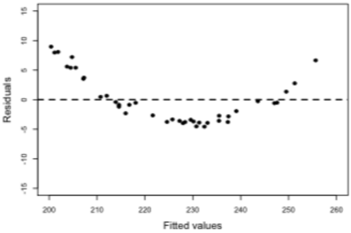
- The fitted values - on the horizontal line
- residuals - the vertical deviations from the line
- No obvious pattern in the residuals → **Linear**
- What if there were a nonlinear or curvy pattern?
- this would indicate a violation of the linearity assumption



Linearity

Residual plot

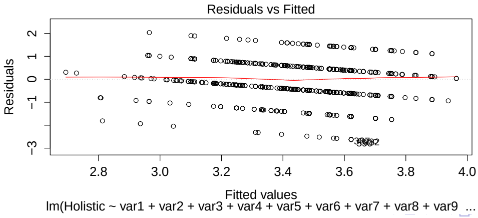
- The fitted values - on the horizontal line
- residuals - the vertical deviations from the line
- No obvious pattern in the residuals → **Linear**
- What if there were a nonlinear or curvy pattern?
- this would indicate a violation of the linearity assumption



Linearity

What to do in case of non-linearity?

1. You might miss an important fixed effects. Add them
2. Perform a nonlinear transformation of your response, e.g., log-transform (*commonly chosen*)
3. Perfrom a nonlinear transformation of your fixed effects
    - if *age* showed in a U-shaped
    - add age and age^2 as predictors
4. if stripes in residual plot, then you're most likely dealing with categorical data → different model such as logistic models



Collinearity

What is collinearity?

- When two fixed effects (predictors) are correlated with each other,
- they are said to be **collinear**

Example

- you were interested in how average talking speed affects intellignece ratings

$$intelligence\ ratings \sim talking\ speed$$

- you measured several different indicators of talking speed
  - syllables/sec, words/sec, sentences/sec
- they are likely to be highly correlated with each other
- if you use all of them as predictors within the same model, there will be **collinearity problem**

Collinearity

If there is collinearity

- the interpretation of the model becomes unstable
- the significance of these correlated or collinear fixed effects is not easily interpretable
  - ∵ they might steal each other's explanatory power
- if multiple predictors are very similar to each other
  - it becomes very difficult to decide what, in fact, is playing a big role

Collinearity

How to get rid of collinearity?

1. pre-empt the problem in the design stage
   - focus on a few fixed effects that are not correlated with each other
2. Or, think about which one is the most meaningful and drop the others
   - DO NOT base this dropping decision on the **significance** (circular logic problem)
3. Or, consider dimension-reduction techniques such as Principal Component Analysis
   - transform several correlated variables into a smaller set of variables which you can then use as new fixed effects.

Homoskedasticity

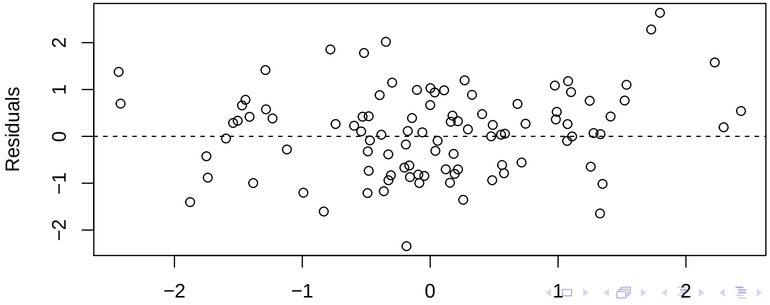
What is Homoskedasticity?

- The variance of your data should be approximately equal across the range of your predicted values
- it is **extremely important assumption**
- If homoscedasticiy is violated → a problem with unequal variances

Homoskedasticity

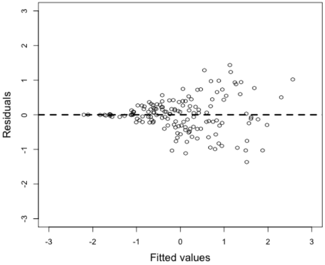
How to check whether homoscedasticity assumption were met?

- the **residuals** of your model need to roughly have a similar amount of deviation from the predicted values
- See Residual Plot
- Good residual plot essentially looks blob-like



Homoskedasticity

Example of **heteroskedasticity**



- higher fitted values have larger residuals

What to do?

- As mentioned earlier, consider a log-transform

Normality of residuals

Normality of residuals

- It is the one that is least important
- LM is relatively robust agains violation of the normality assumption
- Gellman and Hill (2007), a famous book on LM and mixed models, DO NOT EVEN RECOMMEND diagnostics of the normality assumption

If you want to test the assumption

- Histogram

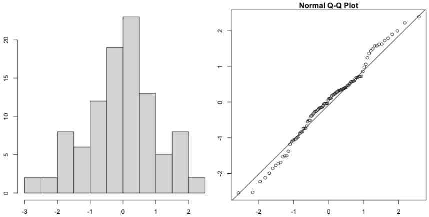
```
hist(residuals(xmdl))
```

- Q-Q plot

```
qqnorm(residuals(xmdl))
```

Normality of residuals

Example of normality



- Thses look good
- The histogram is relatively bell-shaped
- The Q-Q plot indicates that the data falls on a straight line
  - which means that it's similar to a normal distribution
- can conclude that there are no obvious violations of the normalitry assumption

Absence of influential data points

What is the influential data point?

- If a particular data point is excluded, when values with which the coefficient is adjusted is large, it is an influential data point.
- Influential data points can drastically change the interpretation of the results, it can lead to instable results

Absence of influential data points

How to check?

- Using `dfbeta()` function

```
dfbeta(xmdl)

##   (Intercept)        age
## 1  -5.7886394  0.4307550
## 2  51.5196947 -0.9857490
## 3  -0.1678502  0.3289864
## 4 -30.9878812  0.5260128
## 5 -20.6098951  0.1939283
## 6  39.1182640 -1.3964689
```

- DFbeta values are the values of coefficient as a result of *leave-one-out diagnostics*
- For example, if data point 1 is excluded, the coefficient for age has to be adjsted by 0.0644 from -0.9099, so -0.8455

Absence of influential data points

What is the criteria for decision of influential data point

- There is no clear, sharp criteria
- One thing for sure
  - any value that changes the sign of the slope is **definitely** an influential point
  - be cautious to DFbeta value which is at least half of the absolute value of the slope (To author)

Absence of influential data points

How to proceed if there are influential data points?

- DO NOT SIMPLY EXCLUDE those points and report only the results on the reduced set
  - The only case to exclude influential points is when
  - there is an obvious error (negative age)
  - or there is a value that obviously is the result due to a technical error (voice pitch value of 0)
- Run the analysis **with** the influential points and **without** the points, reports both analyses, state whether the interpretation of the results does or does not change

Independence

What is independence?

- easy example - coin flip or roll of a dice
- each try is not influenced by another try
- each coin flip and each roll of a dice is absolutely independent from the outcome of the preceding coin flips or dice rolls
- The same should hold for your data points for LM analysis
- the data points should come from DIFFERENT SUBJECT
- Each subject should only contribute one data point
- Independence assumption is by far **the most important one**

Independence

When you violate the indepence assumption?

- may greatly inflate chance of finding a *spurious result*
- and it results in a p-value that is *completely meaningless*.

How can guarantee independence?

- Independence is a **question of the experimental design**
- by only collecting one data point per subject

Independence

If you want to collect more data per subject?

- such as in repreated measures design
- need to resolve these non-independence at the analysis stage
- This is where **MIXED MODELS** comes in

Mixed models will be proceeded in Tutorial 2

Independence

If you want to collect more data per subject?

- such as in repreated measures design
- need to resolve these non-independence at the analysis stage
- This is where **MIXED MODELS** comes in

Mixed models will be proceeded in Tutorial 2