Additional Regression Techniques

Sean Hellingman ©

Regression for Applied Data Science (ADSC2020)

shellingman@tru.ca

Winter 2025

THOMPSON RIVERS UNIVERSITY

Topics

2. Introduction
3. Mixed Effects Regression
4. Nonparametric Regression
    - Local Regression
5. Exercises and References

Introduction

- Within our regression models we can consider the case that there may be multiple sources of variability.
- Mixed effects regression can be used to model these situations.
- We may also estimate regression models without distribution assumptions.
  - Nonparametric regression.

Mixed Effects Regression

Mixed Models

- (G)LM models assume independent observations.
- Sometimes this is not realistic:
  - Data collected from different regions or states.
  - Payments on the same claim (insurance).
  - Panel data.
  - Nested/hierarchical data.
- Assume a secondary source of variability.

Mixed Models Specifications

- (Generalized) Linear Mixed Models ((G)LMM) contain two types of effects:
    1. Fixed effects: $\mathbb{X}^{\mathbb{T}}\beta$
    2. Random effects: $\mathbb{Z}^{\mathbb{T}}\gamma$

$$g(\mu) = \mathbb{X}^{\mathbb{T}}\beta + \mathbb{Z}^{\mathbb{T}}\gamma$$

- Note: $g(\mu)$ is the identity for LMM.

Assumptions

- **The regular assumptions on the fixed effects of the (G)LM apply and need to be tested.**
- It is assumed that the random effects are normally distributed around the fixed effect (intercept).

Random Intercepts

- By including a random intercept, one assumes that each group in the grouping variable has its own intercept.
- The slopes remain the same and the random intercepts are distributed around the overall intercept.
- Note: *This is what we have done in ADSC2030.*

Random Slopes

- Now we are assuming that the relationships in the explanatory variables differs *within the groups*.
- Again, the random slopes are distributed around the overall slope.

Mixed Models Illustration

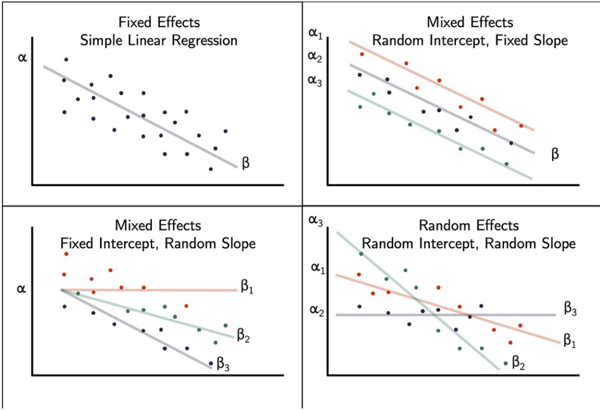

Figure: Source: Wikipedia

Mixed Models in R

- LMM:

```
model <- lmer(Response ~ fixed.variable.1 + ...  +
fixed.variable.p + (1+random.slope|random.intercept),
data = dataset)
```

- GLMM:

```
model <- glmer(Response ~ fixed.variable.1 + ...  +
fixed.variable.p + (1+random.slope|random.intercept),
family = distribution, data = dataset)
```

- Note: *Models are estimated using REML from the lme4 package.*

Example 1

1. Import the *MLSData.csv* dataset.

2. Take some time to get to know the dataset.

3. Estimate two LMM model with `GP` (MLS games played) as the response variable (only random intercepts & both random intercepts and slopes).

4. Using the binary response `Ind30` variable (indicating if a player has played 30 MLS games or not) estimate two GLMM models (only random intercepts & both random intercepts and slopes).

5. Comment on your overall findings.

Example 2

1. Adjust the provided code to check the normality assumptions of the random effects in your models.

Comments on Predictions

- You may use the `predict(GLMM.model, new.data, type = )` function to make predictions.
- **You cannot predict on unseen random effects.**

Nonparametric Regression

Nonparametric Regression

- No parametric relationship between the response and explanatory variables is assumed.
- Instead, the relationship is determined from the data through some algorithm.
- This means that larger sample sizes are needed compared to parametric regression.

Nonparametric Regression Methods

- There are many methods that exist.
- We will cover one very popular methods in this course:
    - Local Regression

Local Regression

- **LOESS: LOcal regrESSion** fits a weighted least squares model for y_i values corresponding to the x_i values that are *near* some given x.
- In the simple regression case:

$$\textit{minimize} \sum_{i=1}^{n} w_i(x)(y_i - \beta_0 - \beta_1 x_i)^2$$

- Where the weights are given by:

$$w_i(x) = \begin{cases} (1 - |x - x_i|^3)^3 & \text{if } \; |x - x_i| < \delta_i \\ 0 & \text{otherwise} \end{cases}$$

Comments on LOESS

- The choice of δ_i (span) is very important.
  - Becomes a cross-validation problem.
- You can add polynomials to add flexibility to your localized regression models.
- Sparse observations can cause problems in your estimation.

LOESS in R

- `model <- loess(response ~ variable.1 + ... variable.k, data = dataset, span =` *span*`)`
- The span goes from 0 - 1 and defines the smoothing span (related to δ).
- You can use the `predict(model, new.data)` function to make predictions.

Example 3

1. Import the *Simulated.csv* dataset.
2. Estimate **three** models with different spans and `X1` as the lone explanatory variable.
3. Ammend the provided code to plot all of your results.
4. What did you notice?

Exercise 1

- Can you identify any places where mixed effects regression could be useful?
- Estimate some LMM and GLMM models and see what you find.
- **Be sure to complete the diagnostic checks.**

Exercise 2

- Split the *Simulated.csv* dataset into training and testing data.
- Use all of the explanatory variables to estimate some LOESS regression models.
- Test the accuracy of your models on the unseen data.
- Are there any problems with overfitting?

References & Resources

- Mixed Effects Example
- lmer()
- glmer()
- Nonparametric Examples
- loess()