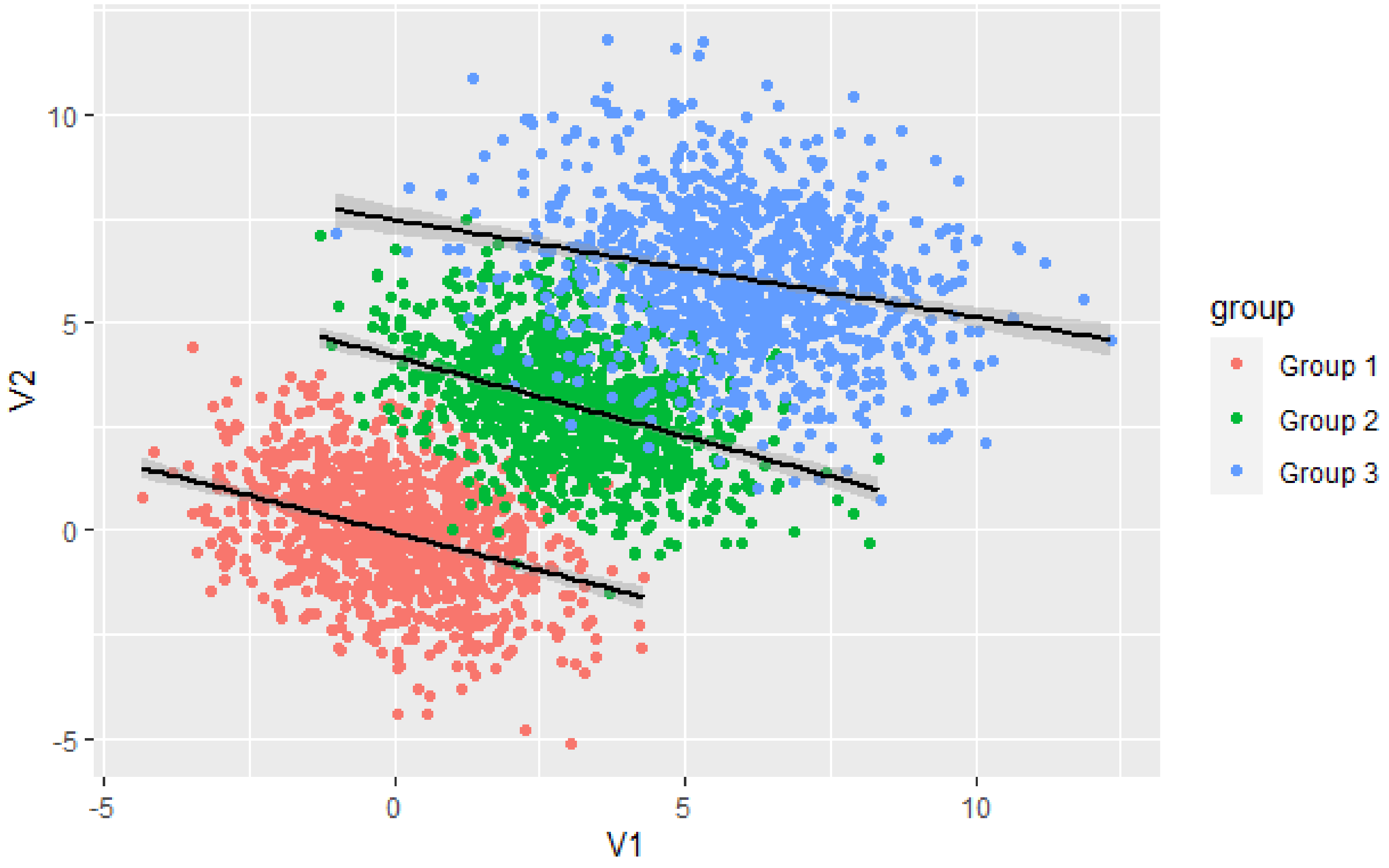
Art of Mixed Effect Modeling

Introduction to Linear Mixed Effects Modeling

Prabakaran Chandran

14th OCT 2021

Have you heard about Simpson paradox?



- Simpson's paradox, which also goes by several other names, is a phenomenon in probability and statistics in which a trend appears in several groups of data but disappears or reverses when the groups are combined.

- It is also referred to as Simpson's reversal, Yule–Simpson effect, amalgamation paradox, or reversal paradox

This phenomenon needs to be handled and investigated with proper statistical modeling.

Problem of Non-Independence in Data

- Traditional Mathematical Statistics is based to a large extent on assumptions of the Maximum Likelihood principal and Normal distribution.

- In case of multiple linear regression these assumptions might be violated if there is non-independence in the data.

- There can be two types of non-independence in the data:
  - ❑ non-independent variables / features (multicollinearity)
  - ❑ non-independent statistical observations (grouping of samples)

Store	Location	Month	Price	Offer	Comp.Price	Sales
A	Bangalore	Oct	100	10%	101	2000
B	Chennai	Oct	103	7%	99	1300
C	Coimbatore	Oct	120	5%	131	900

Scenario 1 : Price and Offer might be dependent on each other

Store	Location	Month	Price	Offer	Comp.Price	Sales
A	Bangalore	Aug	100	10%	101	2000
A	Bangalore	Sep	103	7%	99	1300
A	Bangalore	Oct	120	5%	131	900

Scenario 2 : Price of September might be dependent on each other

What if samples don't follow non-Independence?

Store	Location	Month	Price	Offer	Comp . Price	Sales
A	Bangalore	Aug	100	10%	101	2000
A	Bangalore	Sep	103	7%	99	1300
A	Bangalore	Oct	120	5%	131	900
B	Coimbatore	Aug	120	10%	101	2000
B	Coimbatore	Sep	120	7%	99	1400
B	Coimbatore	Oct	115	5%	131	3000
C	Chennai	Aug	120	10%	101	2000
C	Chennai	Sep	110	7%	99	2000
C	Chennai	Oct	105	5%	131	1000

Dependence among price exists
(either positive or negative)

Longitudinal data, sometimes called panel data, is data that is collected through a series of repeated observations of the same subjects over some extended time frame—and is useful for measuring change

- To overcome the problem of non-independent variables, one can for example select most informative variables with LASSO, Ridge or Elastic Net regression,
- But the non-independence among statistical observations cannot be solved using the regularized regression techniques

Mixed Effects Modeling

- Mixed Effects ➔ Mixture of Two effects ; Fixed Effects + Random Effects
- It is an Extension of Linear Model
- Applied in Economics , Biology , Business and Life sciences
- powerful tool for linear regression models when data contains global and group-level trends.
- Repeated measurements are made on the same statistical units (longitudinal study), or where measurements are made on clusters of related statistical units.
- In the field of ecological and biological data are often complex and messy and sometimes bi-modal. We may have different **grouping factors** like populations, species, sites, gender ,etc.
- Allows measurements to be made repeatedly over time.
- Can work on other types of dependent variable:- categorical, continuous, ordinal, discrete count, etc.
- Works for correlated data regression models, including repeated measures, longitudinal, time series, clustered & other related methods.
- Types of mixed models: Within-Subject Designs ,Repeated Measures , Longitudinal Studies , Hierarchical or Multilevel Models

- **Linear Model ➔ Y = Fixed Effect + Error**

- **Linear Mixed Model**
- **Y = Fixed Effect + Random Effects + Error**

$$y = X\beta + Zu + \epsilon$$

- **y** is a known vector of observations;
- β is an unknown vector of fixed effects;
- **u** is an unknown vector of random effects;
- ϵ is an unknown vector of random errors;
- **X** and **Z** are known design matrices relating the observations **y** to β and **u** , respectively.

What are Fixed Effects?

Store	Location	Month	Price	Offer	Comp . Price	Sales
A	Bangalore	Aug	100	10%	101	2000
A	Bangalore	Sep	103	7%	99	1300
A	Bangalore	Oct	120	5%	131	900
B	Coimbatore	Aug	120	10%	101	2000
B	Coimbatore	Sep	120	7%	99	1400
B	Coimbatore	Oct	115	5%	131	3000
C	Chennai	Aug	120	10%	101	2000
C	Chennai	Sep	110	7%	99	2000
C	Chennai	Oct	105	5%	131	1000

- A **fixed effects model** is a statistical model in which the model parameters are fixed or non-random quantities. It is assumed that the observations are independent.
- Fixed effects are, essentially, your predictor variables. This is the effect you are interested in after accounting for random variability
- Sales = **Effect of Price * Price + Effect of Offer * Offer** + Intercept

- Here Effect of Price and Offer is a Fixed variables

What are Random Effects?

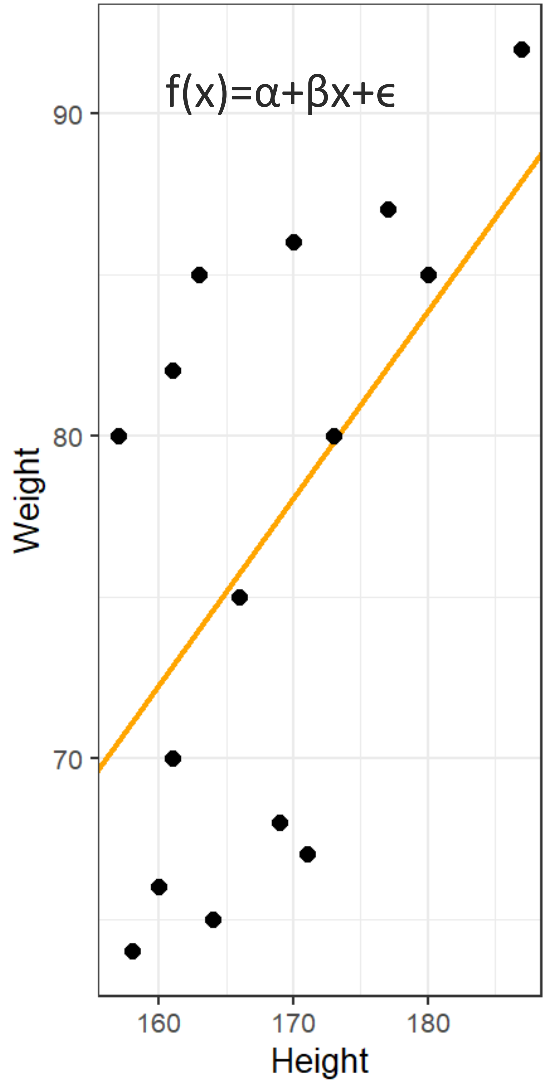
Store	Location	Month	Price	Offer	Comp . Price	Sales
A	**Bangalore**	Aug	100	10%	101	2000
A	**Bangalore**	Sep	103	7%	99	1300
A	**Bangalore**	Oct	120	5%	131	900
B	**Coimbatore**	Aug	120	10%	101	2000
B	**Coimbatore**	Sep	120	7%	99	1400
B	**Coimbatore**	Oct	115	5%	131	3000
C	**Chennai**	Aug	120	10%	101	2000
C	**Chennai**	Sep	110	7%	99	2000
C	**Chennai**	Oct	105	5%	131	1000

- A **random effects model** is a statistical model where the model parameters are random variables. It is assumed that some type of relationship exists between some observations.
- Random effects are best defined as noise in your data. These are effects that arise from uncontrollable variability within the sample. Subject level variability is often a random effect.
- **Eg :-** In the above table effect of Location on sales is random, It will vary within the group
- Sales = **Effect of Location * Location** + Intercept

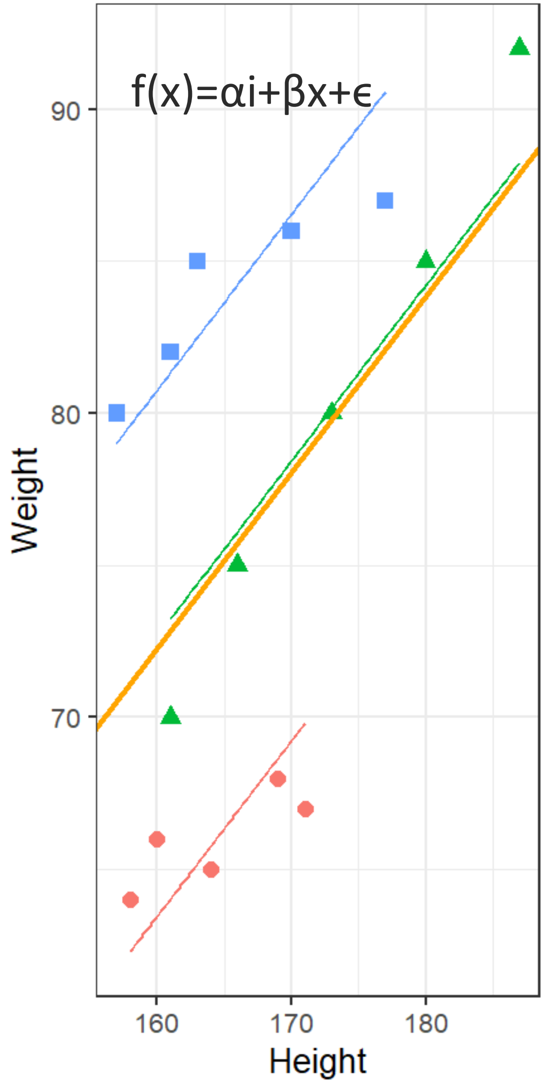
- Here Effect of Location is a Random one.

Example of Height vs Weights of 4 individuals

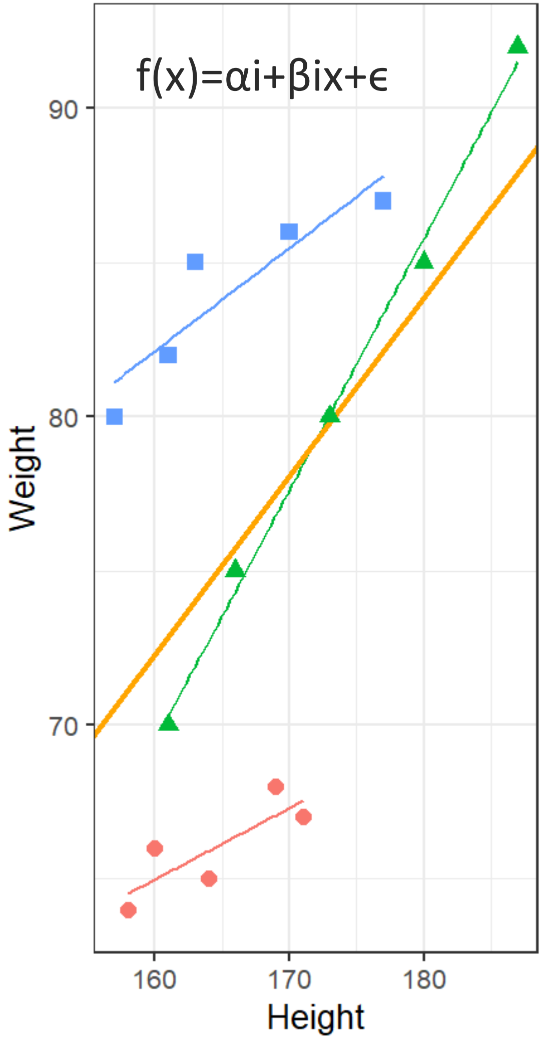
Fixed-effects model
(with fixed intercept)

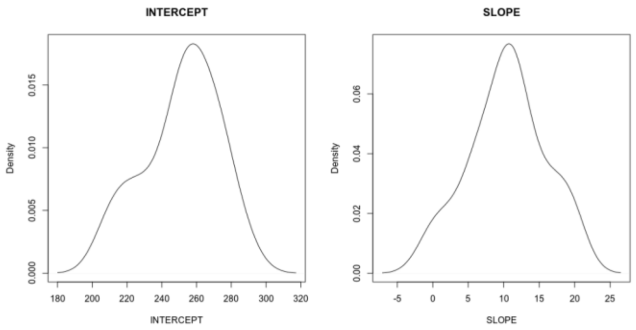


Mixed-effects model
(with random intercepts)



Mixed-effects model
(with random intercepts +
random slops)





Mixed Effects Models – Coefficient Equations

- The response $N \times 1$ vector $\boldsymbol{Y}$ is modelled as

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{u} + \boldsymbol{\epsilon}, \quad \text{where} \quad E\begin{bmatrix}\boldsymbol{u}\\ \boldsymbol{\epsilon}\end{bmatrix} = \begin{bmatrix}0\\ 0\end{bmatrix} \quad \text{and} \quad \text{var}\begin{bmatrix}\boldsymbol{u}\\ \boldsymbol{\epsilon}\end{bmatrix} = \begin{bmatrix}\mathbf{G} & 0\\ 0 & \mathbf{R}\end{bmatrix}$$

  where
  - $\mathbf{X}$ is the $N \times p$ design matrix for the fixed effects $\boldsymbol{\beta}$,
  - $\mathbf{Z}$ is the $N \times q$ design matrix for the random effects $\boldsymbol{u}$,
  - $\boldsymbol{\epsilon}$ is the $N \times 1$ vector of error.
- The above model is referred to as a linear mixed model. Some also refer it to as mixed linear model, mixed-effects model, linear mixed-effects model, hierarchical model, multi-level model, nested models (latter three usual specific to a structure in the data) ...
- We usually assume $\begin{bmatrix}\boldsymbol{u}\\ \boldsymbol{\epsilon}\end{bmatrix} \sim N\left(\begin{bmatrix}0\\ 0\end{bmatrix}, \begin{bmatrix}\mathbf{G} & 0\\ 0 & \mathbf{R}\end{bmatrix}\right)$ hence $\boldsymbol{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{ZGZ}^\top + \mathbf{R})$.
- Note $\text{var}(\boldsymbol{Y}) = \mathbf{V} = \mathbf{ZGZ}^\top + \mathbf{R}$, where we often assume $\mathbf{R} = \sigma^2 \boldsymbol{I}_N$.

Mixed Effects Models – Coefficient Equations

- The log-density function of the joint distribution of $\boldsymbol{y}$ and $\boldsymbol{u}$ is given by

$$\ell = -\frac{1}{2}\log|\mathbf{R}| - \frac{1}{2}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{u})^\top \mathbf{R}^{-1}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{u}) - \frac{1}{2}\log|\mathbf{G}| - \boldsymbol{u}^\top \mathbf{G}^{-1}\boldsymbol{u} + \text{constant}.$$

- The $(\hat{\boldsymbol{\tau}}, \tilde{\boldsymbol{u}})$ that jointly maximises ℓ (assuming $\mathbf{R}$ and $\mathbf{G}$ are known) leads to the mixed model equations (MME), sometimes referred to as Henderson's equations:

$$\begin{bmatrix} \mathbf{X}^\top \mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^\top \mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^\top \mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^\top \mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \tilde{\boldsymbol{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^\top \mathbf{R}^{-1}\boldsymbol{y} \\ \mathbf{Z}^\top \mathbf{R}^{-1}\boldsymbol{y} \end{bmatrix}$$

- which gives the solution (assuming that $\mathbf{X}$ is full-rank)

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{X}^\top \mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top \mathbf{V}^{-1}\boldsymbol{y} \\ \tilde{\boldsymbol{u}} &= \mathbf{G}\mathbf{Z}^\top \mathbf{V}^{-1}(\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}}). \end{aligned}$$

- The solutions of MME are referred to as best linear unbiased estimate (BLUE) for $\hat{\boldsymbol{\beta}}$ and best linear unbiased predictor (BLUP) for $\tilde{\boldsymbol{u}}$.
- When the variance parameters are estimated and "plugged in" the above solution, we refer to them as empirical BLUE (E-BLUE) and empirical BLUP (E-BLUP).