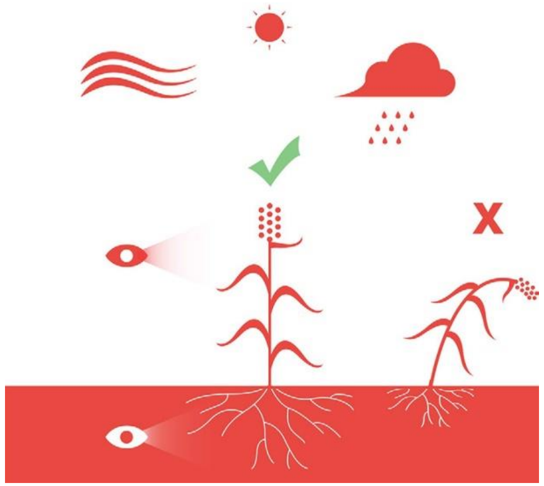
Integrated Data Analysis

Dr. Vuk Đorđević

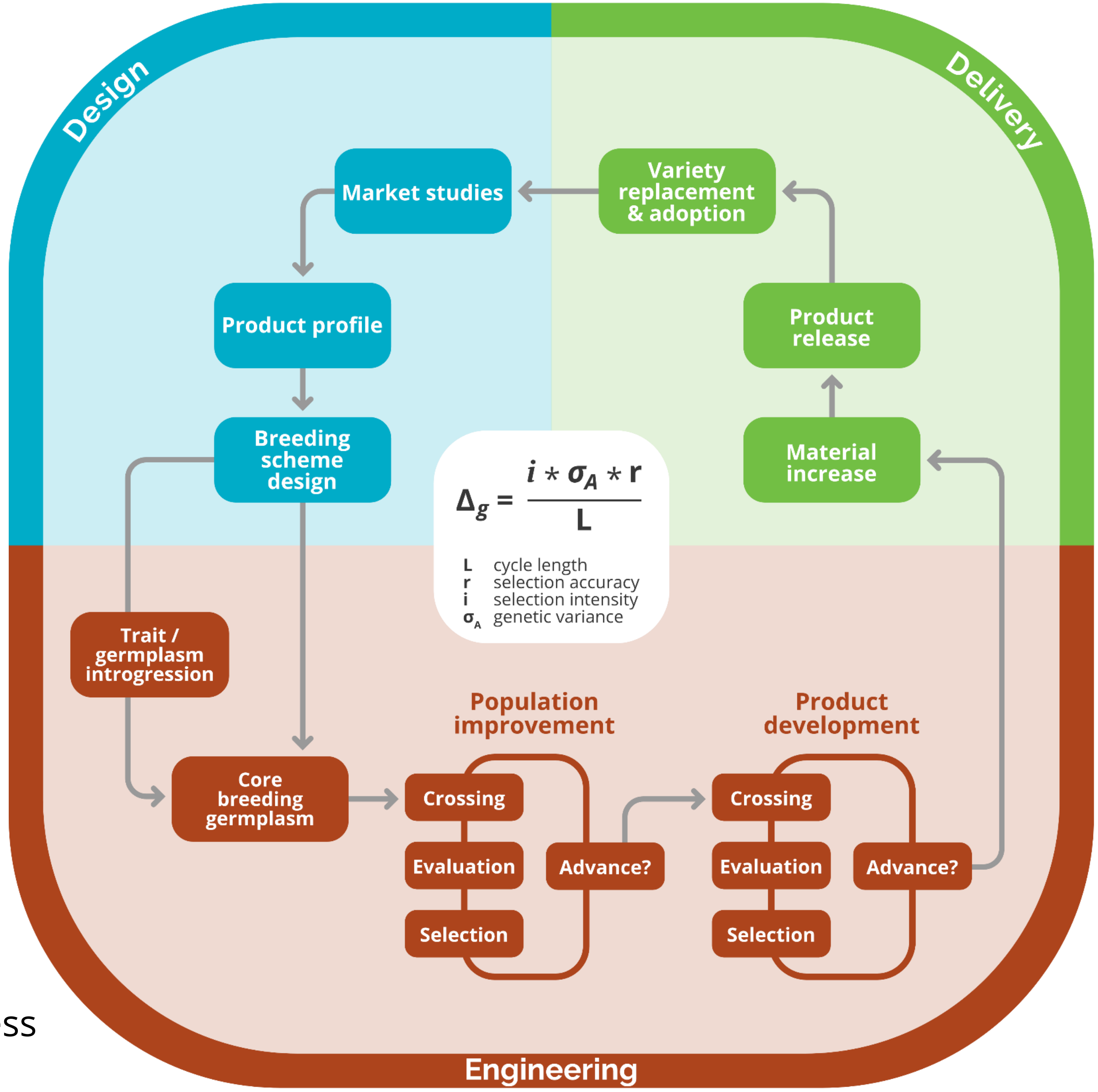
Institute of filed and vegetable crops, Novi Sad
Center of Excellence for Legumes

Key Concepts

- **Pooling Raw Data**
- **Expanded Inquiry**
- **Increased Power**
- **Cumulative Science**



- Integrated Data Analysis is a powerful approach for synthesizing information from multiple sources to advance scientific knowledge and address complex research questions. It offers the potential to move beyond the limitations of individual studies and build a more robust and advanced understanding of phenomena.

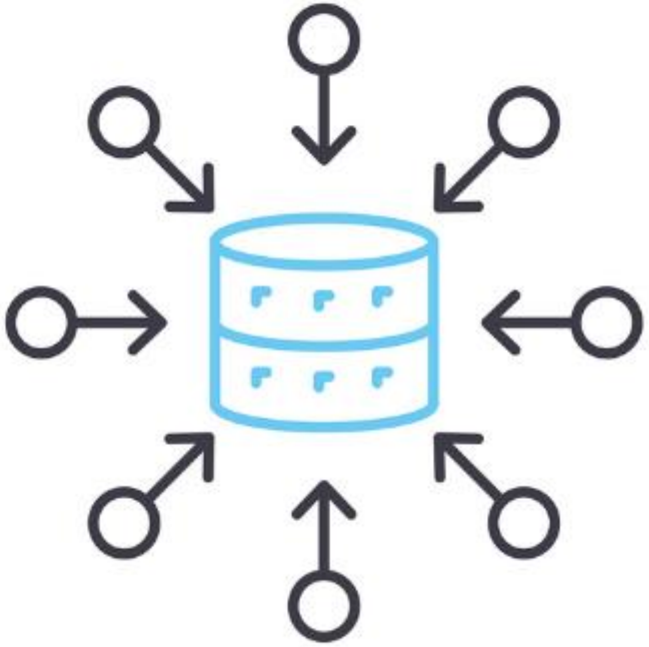


Breeding as a process

Source: EiB

Single data source

- Single experiment/trial
- Single survey
- Data base query
- Simulation and modeling
- ….



Data types

<u>Qualitative Data (Categorical Data):</u>

Nominal Data: categories without any inherent order or ranking.

Examples: colors (red, blue, green), types of fruit (apple, banana, orange)

Ordinal Data: categories with a meaningful order or ranking.

Examples: survey responses like "strongly agree, agree, neutral, disagree, strongly disagree," or disease score (S, MS, M, MR, R).

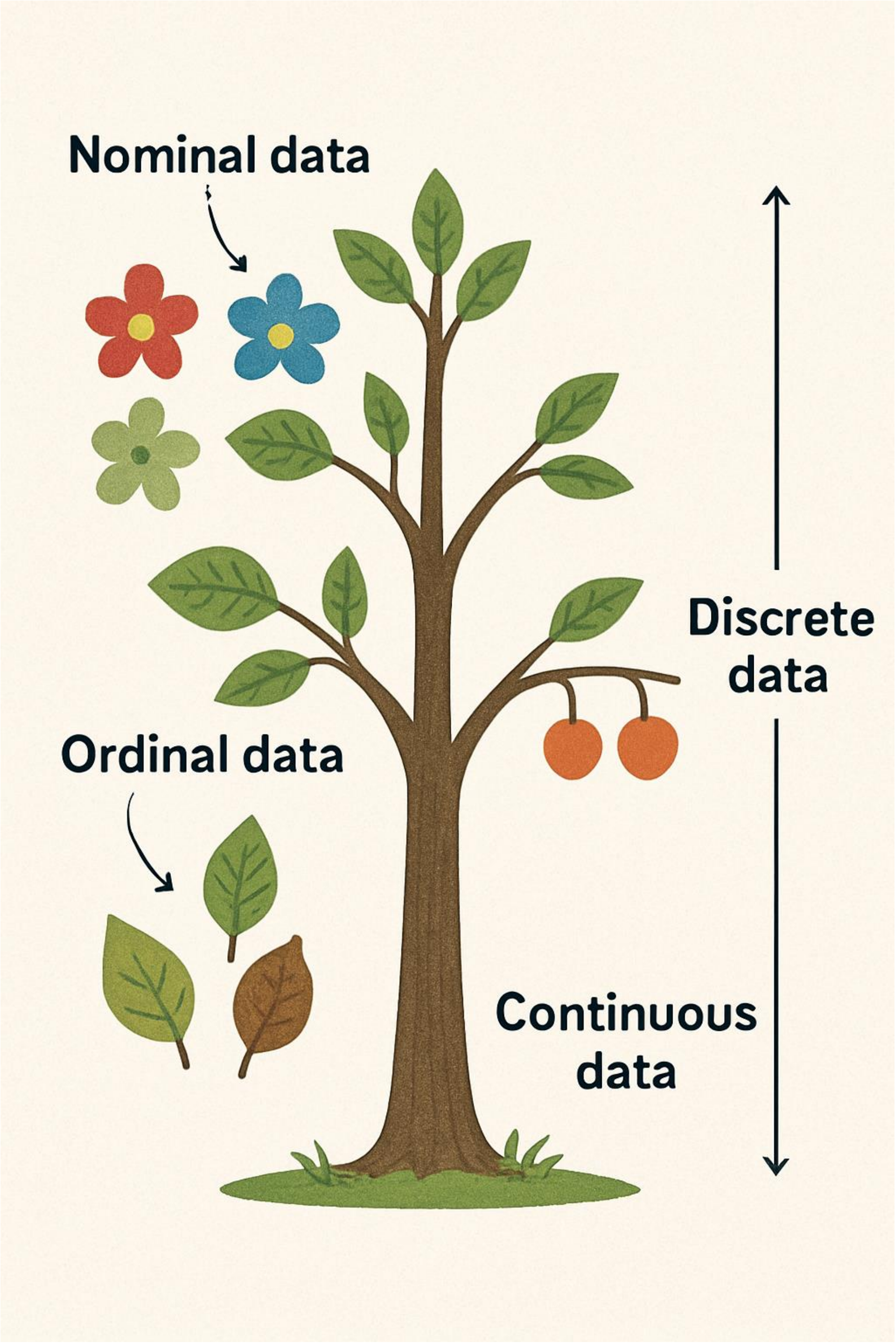
<u>Quantitative Data (Numerical Data):</u>

Discrete Data: countable values that can only take specific, separate values with gaps in between.

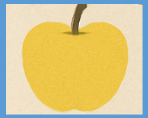
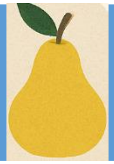
Examples: number of students, number of fruits.

Continuous Data: values that can fall anywhere within a given range.

Examples: height, weight, temperature, time.



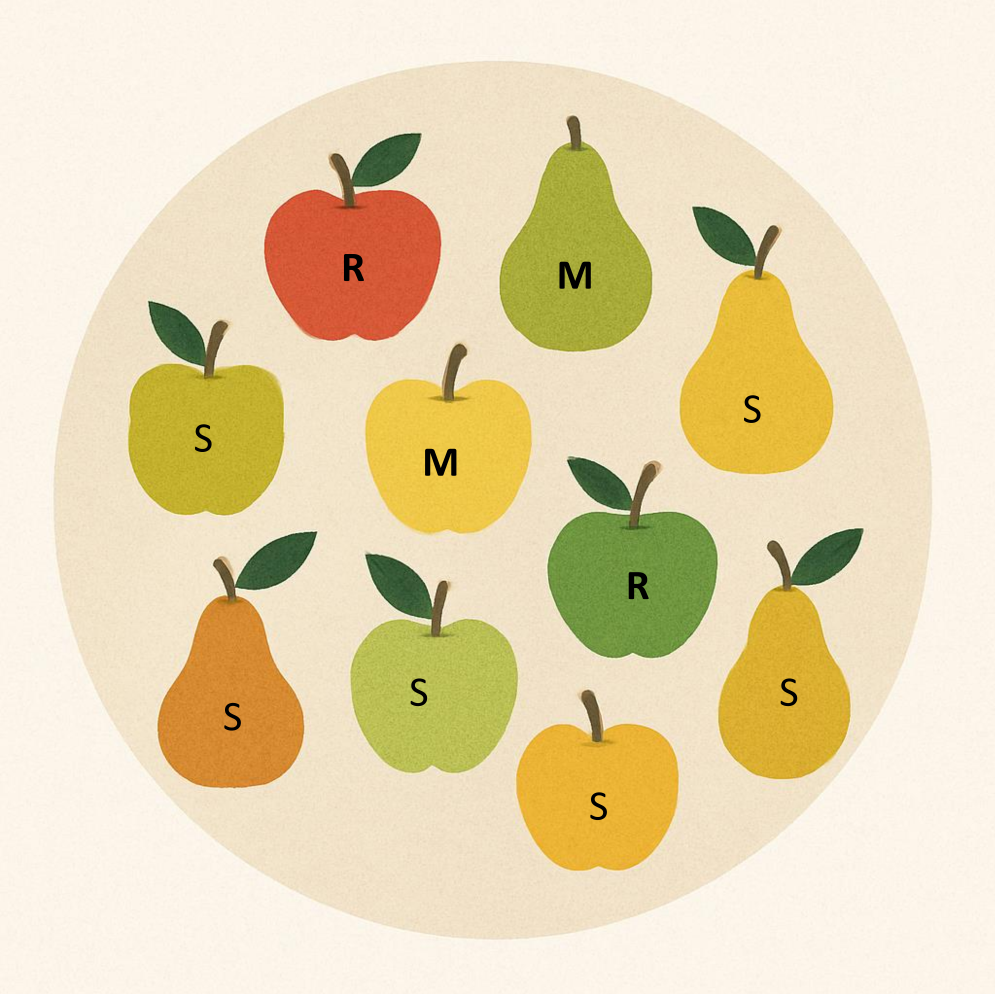
Nominal Data

			
Frequencies	The number of times each category appears in the data	6	4
Proportions Percentages	The frequency of a category divided by the total number of observations	60 %	40 %
Mode	The category with the highest frequency	x	

Do not use mean and standard deviation !

Ordinal Data

		R (1)	M (2)	S (3)
Frequency	raw counts or percentages	2	2	6
Mode	most frequently occurring value			x
Median	middle value when the data is ordered	1, 1, 2, 2, **3, 3**, 3, 3, 3, 3		
Range	difference between the highest and lowest values	2		
Interquartile Range (IQR)	difference between the 75th and 25th percentiles	Q75 = 3, Q25 = 2 IQR = 1		



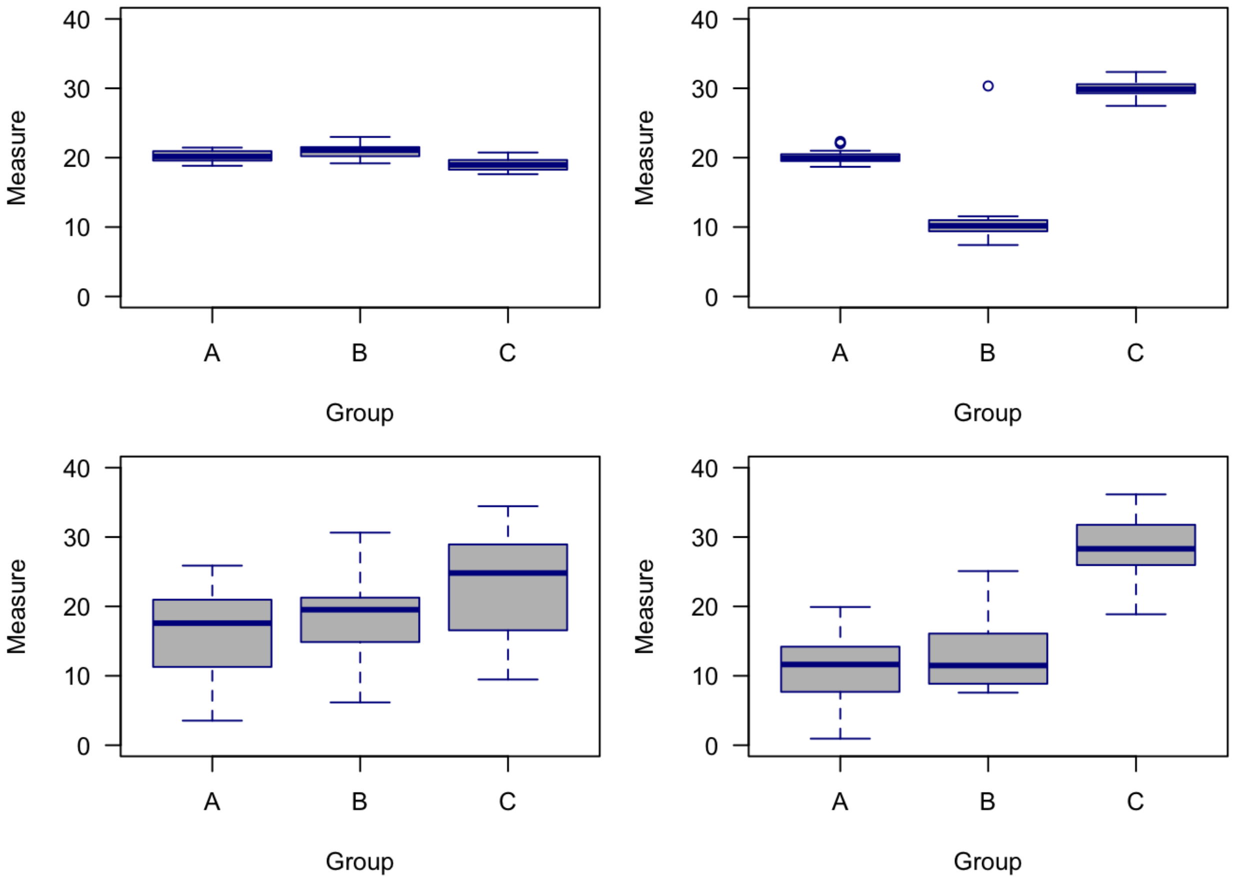
Do not use mean and standard deviation !

Quantitative Data (Numerical Data)

Discrete Data and Continuous Data

Measures of Center	
Mean	The average of all values in the dataset
Median	The middle value when the data is sorted
Mode	The value that appears most frequently
Measures of Dispersion	
Range	The difference between the maximum and minimum values
Variance	A measure of how spread out the data is from the mean
Standard Deviation	The square root of the variance, providing a more interpretable measure of spread

ANalysis Of VAriance



Steve Midway: **Data Analysis in R**
https://bookdown.org/steve_midway/DAR/

two sources of variance

between-group variance

comparing the mean of each group with the overall mean

within-group variance

variation of each observation from its group mean

sums of squares (SS) = is the numerical metric distances of each point to the mean

ANOVA terminology

- **Factor**: different groups (eg. Genotypes, locations, years….)
- **Level**: categories within the factor (# of genotypes, # locations…
- One way ANOA, Two way ANOVA

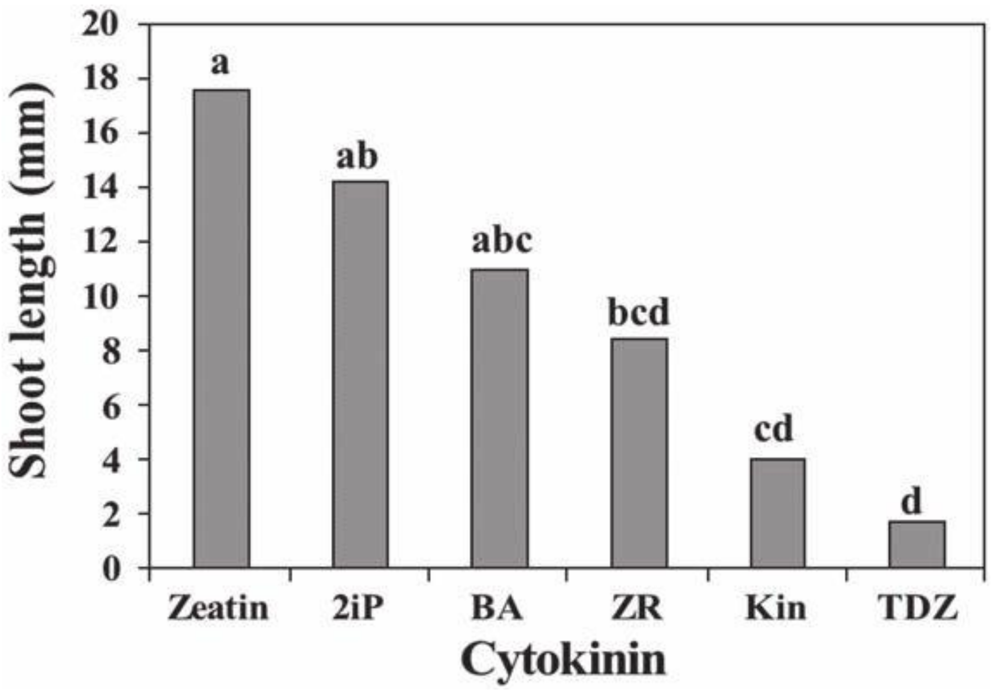
```r
aov.model <- aov(size ~ pop)
summary(aov.model)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## pop          2  34.67  17.333    10.4 0.00457 **
## Residuals    9  15.00   1.667
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

p-value <0.05

significant difference between the means, it only indicates that at least two means are different, not which specific pairs

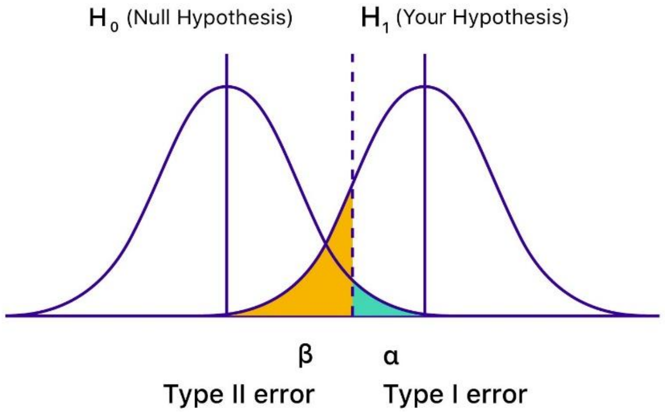
http://dx.doi.org/10.1007/978-1-61779-818-4_4



Tukey's Honest Significant Differences (HSD)

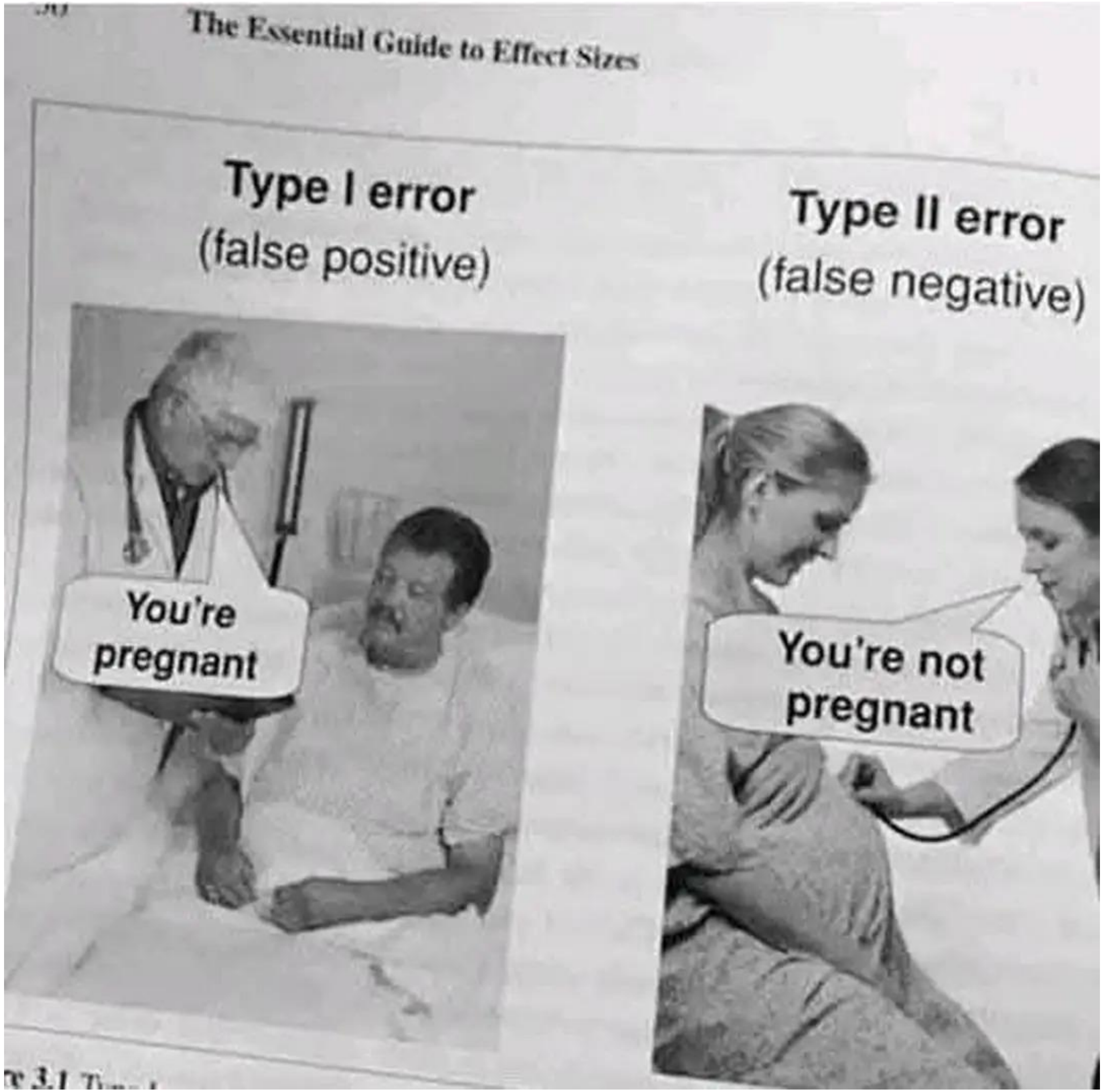
post-hoc comparisons or means comparisons

Error types



Reality

		TRUE	FALES
Test	TRUE		Type II
	FALES	Type I	

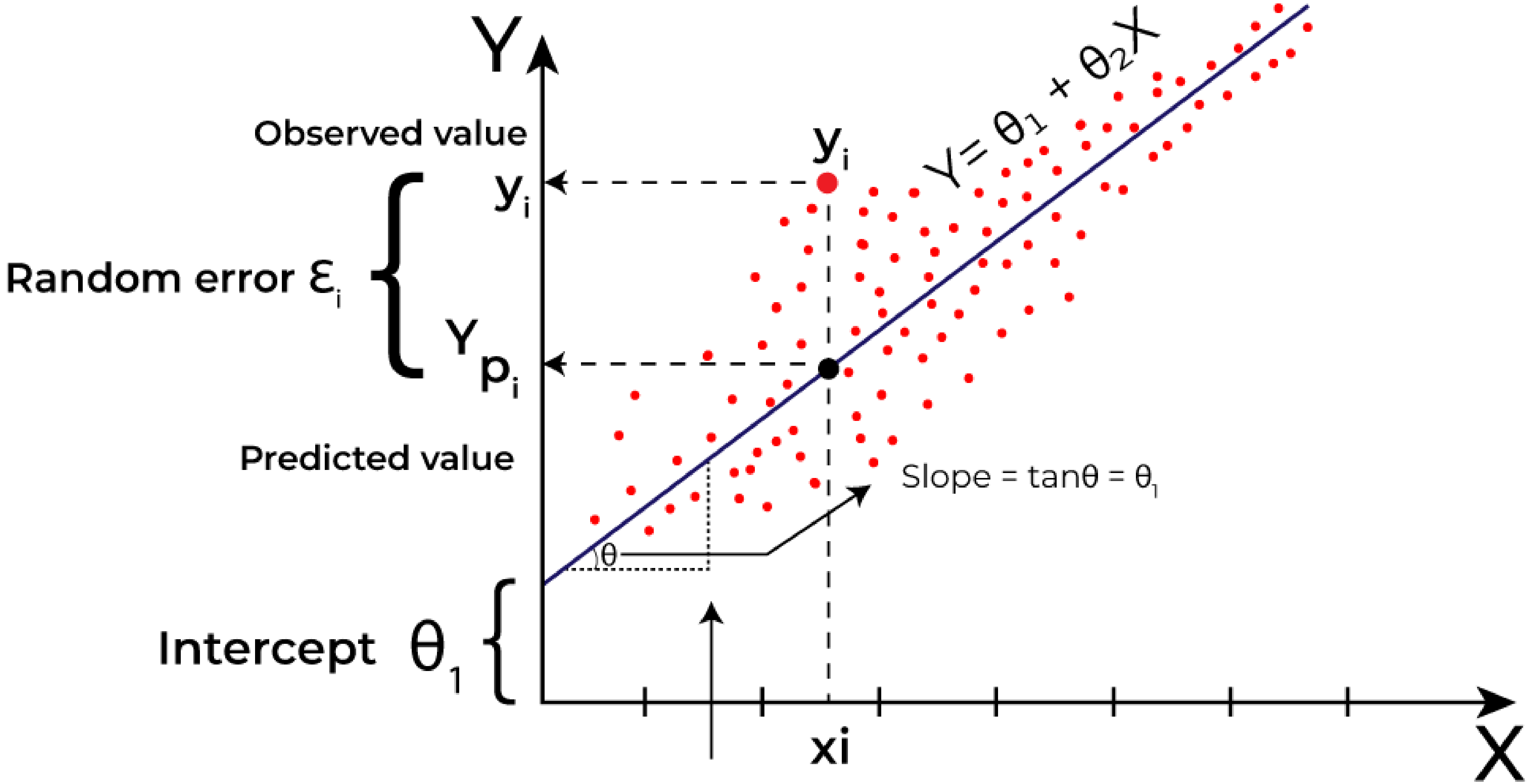


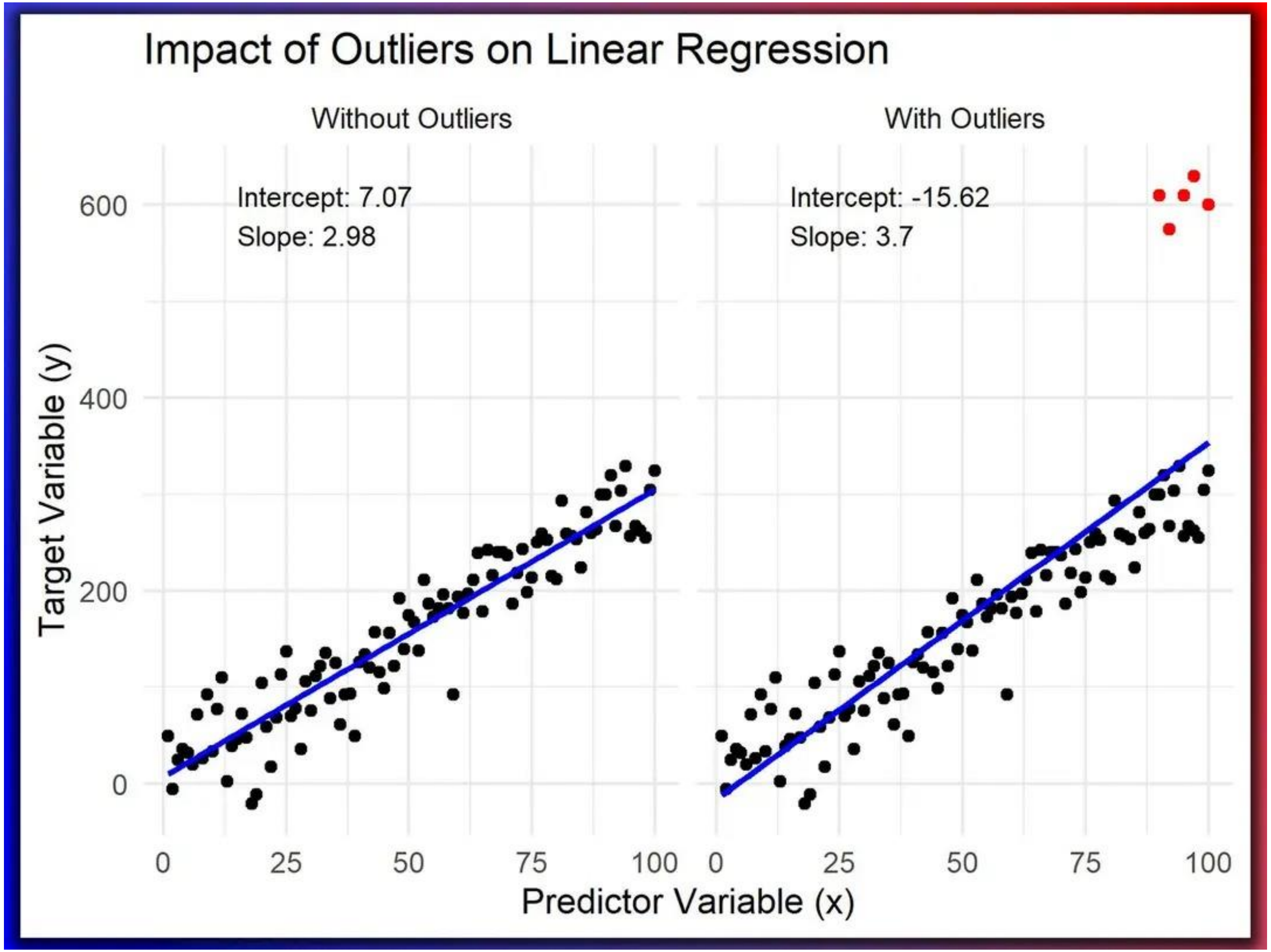
ANOVA interpretation

SCENARIO

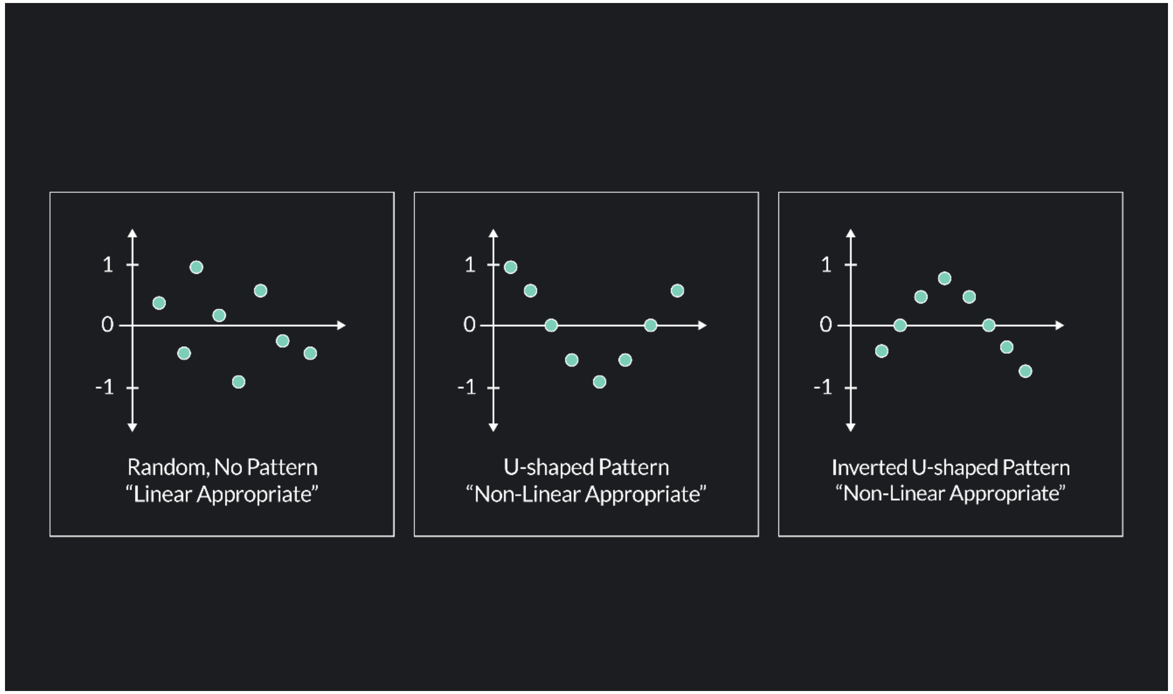
	1	2	3	4
Genotype	**0.45**	**0.30**	**0.30**	**0.25**
Locations	**0.30**	**0.20**	0.10	0.15
Year	0.10	0.05	0.20	0.15
Genotype x Location	0.05	**0.35**	0.05	0.05
Genotype x Year	0.05	0.05	**0.30**	0.05
Genotype x Location x Year	0.05	0.05	0.05	**0.35**

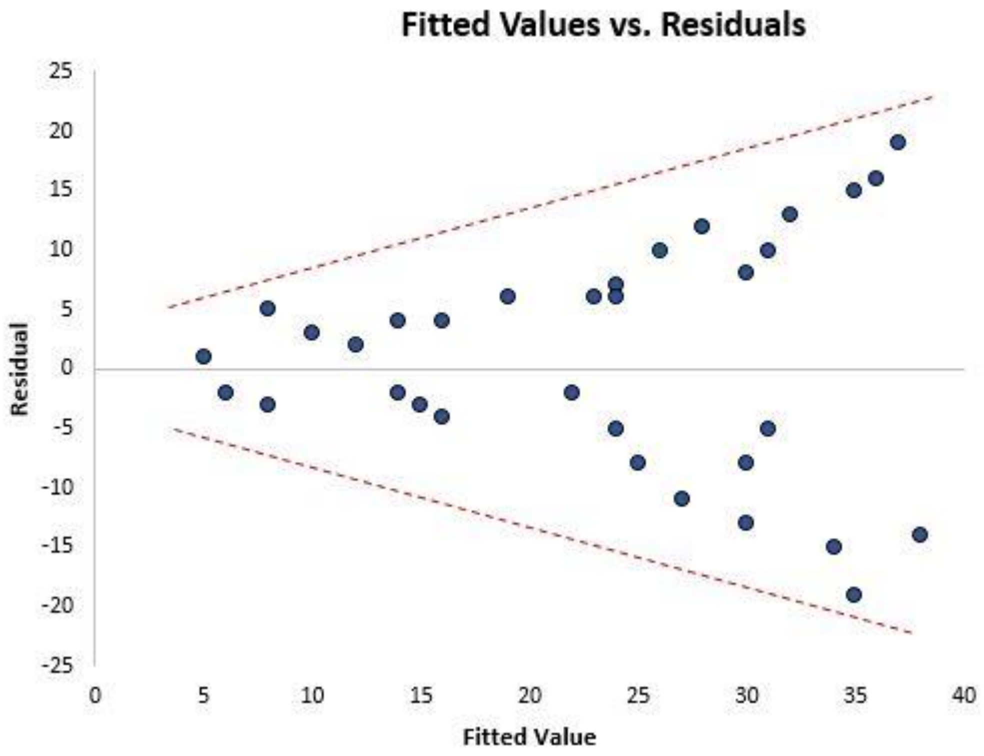
Linear regression




Impact of Outliers on Linear Regression
Without Outliers
With Outliers
Intercept: 7.07
Slope: 2.98
Intercept: -15.62
Slope: 3.7
Target Variable (y)
Predictor Variable (x)
0
200
400
600
0
25
50
75
100

Residuals (actual vs predicted)





Fix vs random linear models

Model is a relationship between variables

Fixed Effects Models: In fixed effects models, **the effects of the independent variables are assumed to be constant across all groups or levels** in the data.

standard regression coefficients representing factors

Random Effects Models: Random effects models, on the other hand, assume that the **variation across entities can be captured in a random component of the model**.

variation introduced by grouping or clustering

Why Use Mixed Models?

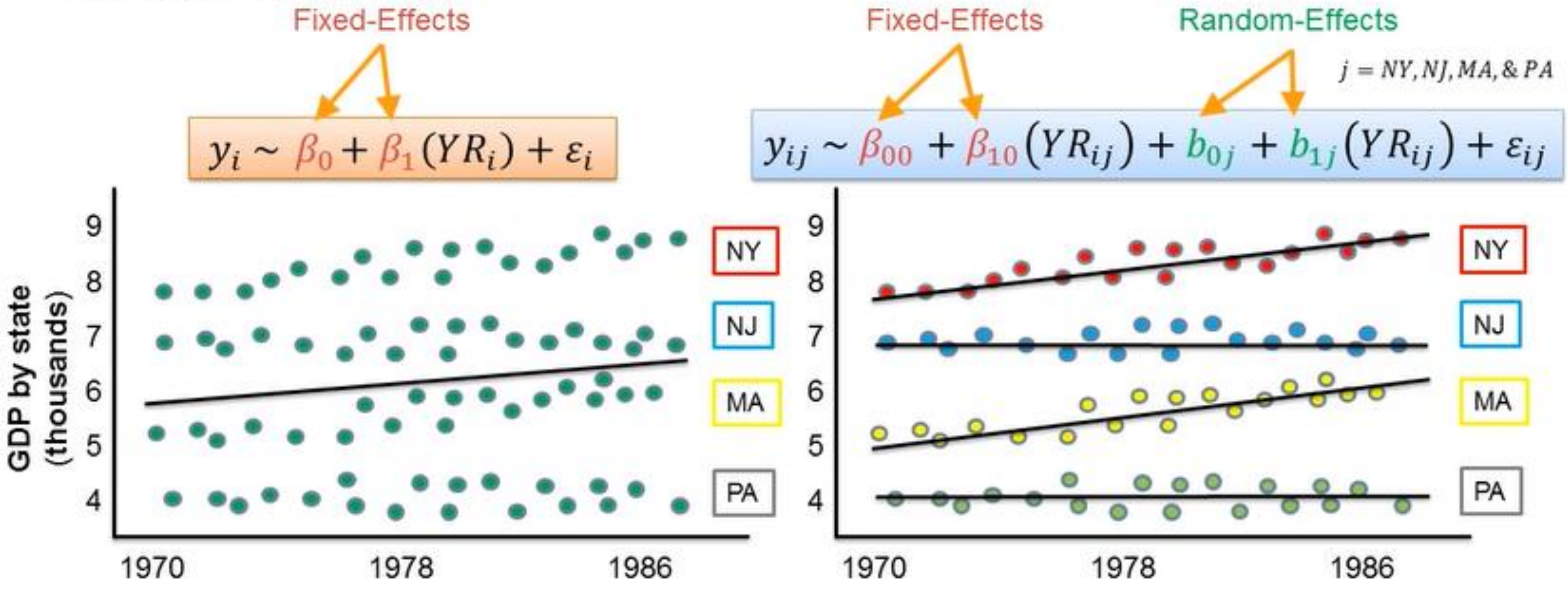
Account for Non-Independence: Mixed models explicitly handle the non-independence that arises from hierarchical data, where observations within the same group are correlated.

Reduce Pseudoreplication: They avoid the problem of pseudoreplication, which occurs when analyses treat non-independent data points as if they were independent.

Model Variability: By modeling variance at different levels, mixed models provide a more accurate understanding of the relationships between variables.

Linear mixed models (LMMs) are applied across many scientific fields to analyze clustered, longitudinal, or repeated-measures data where observations within groups are not independent

Extensions of Linear regression models for data that are collected and summarized in groups.



Trial design

Handles Non-Independence

Multi-environment testing

Accounts for Within- and Between-Subject Variability

Genotype by environment interaction

Flexible for Unbalanced Data

Mathematical Representation of Linear Mixed-Effects Models

Mathematical Representation of Linear Mixed-Effects Models

The general form of a linear mixed-effects model can be represented as:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_j + \epsilon_{ij}$$

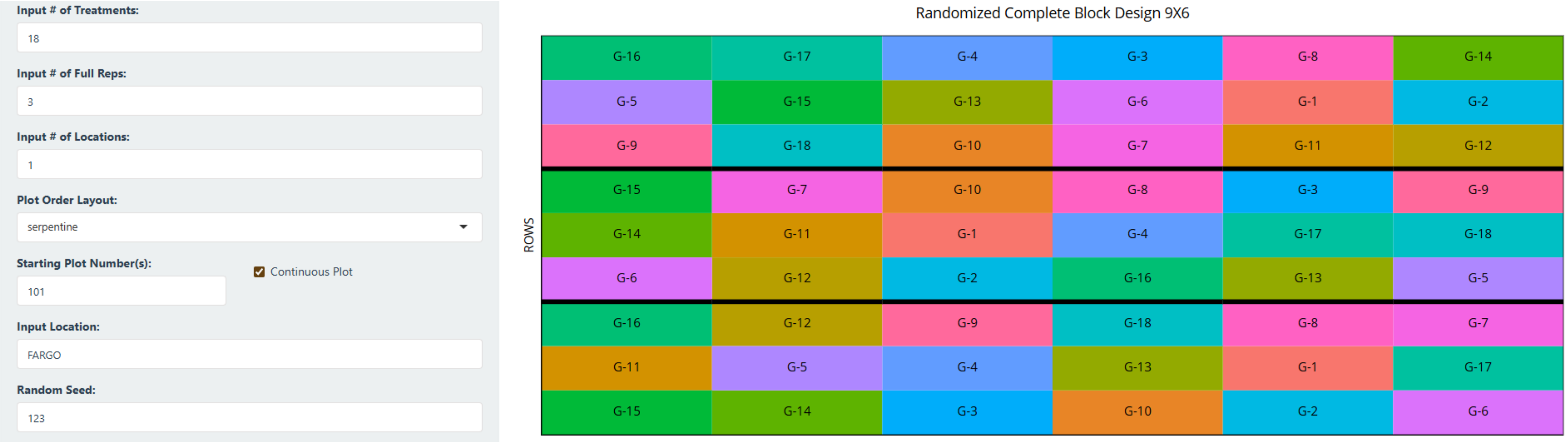
Where:

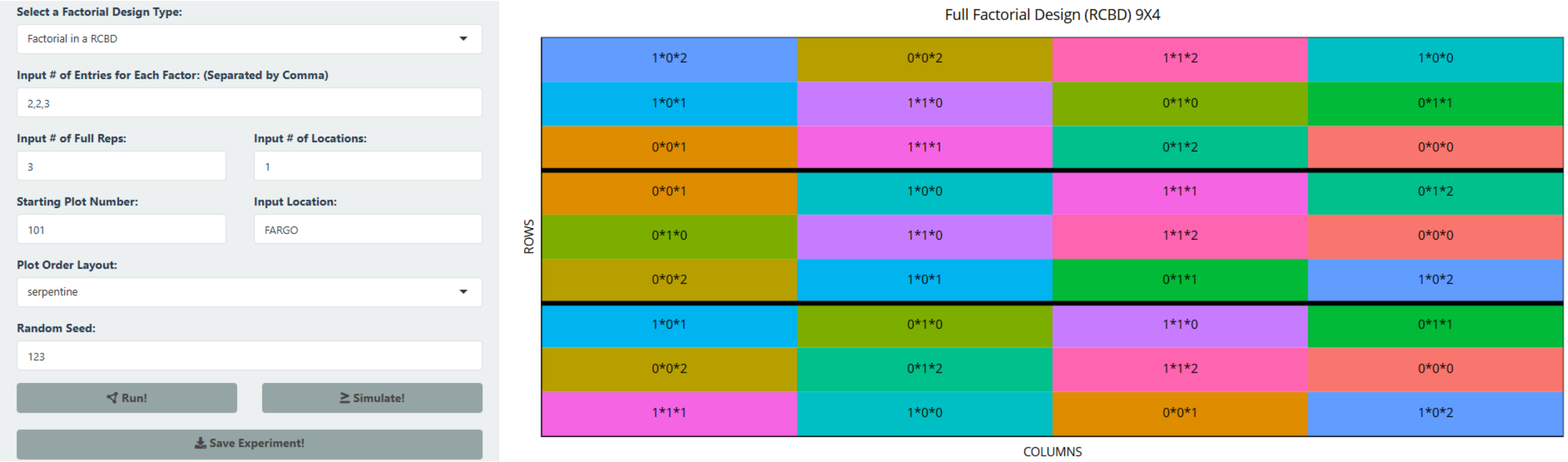
- y_{ij} is the outcome variable for the i-th observation in the j-th group.
- $\beta_0$and β_1 are the fixed-effect coefficients (intercept and slope).
- x_{ij} is the predictor variable for the i-th observation in the j-th group.
- u_j is the random effect for the j-th group.
- ϵ_{ij} is the residual error for the i-th observation in the j-th group.

The random effect u_j is assumed to follow a normal distribution with mean zero and variance σ_u^2, and the residuals ϵ_{ij} are also assumed to be normally distributed with varianceσ^2.

Fixed effects give the estimate of the average effect of the characteristic in the population

Random effects model group-specific deviations of this common average


Input # of Treatments:
18
Input # of Full Reps:
3
Input # of Locations:
1
Plot Order Layout:
serpentine
Starting Plot Number(s):
101
Continuous Plot
Input Location:
FARGO
Random Seed:
123
Randomized Complete Block Design 9X6
ROWS
G-16 G-17 G-4 G-3 G-8 G-14
G-5 G-15 G-13 G-6 G-1 G-2
G-9 G-18 G-10 G-7 G-11 G-12
G-15 G-7 G-10 G-8 G-3 G-9
G-14 G-11 G-1 G-4 G-17 G-18
G-6 G-12 G-2 G-16 G-13 G-5
G-16 G-12 G-9 G-18 G-8 G-7
G-11 G-5 G-4 G-13 G-1 G-17
G-15 G-14 G-3 G-10 G-2 G-6


Select a Factorial Design Type:
Factorial in a RCBD
Input # of Entries for Each Factor: (Separated by Comma)
2,2,3
Input # of Full Reps:
3
Input # of Locations:
1
Starting Plot Number:
101
Input Location:
FARGO
Plot Order Layout:
serpentine
Random Seed:
123
Run!
Simulate!
Save Experiment!
Full Factorial Design (RCBD) 9X4
ROWS
1*0*2 0*0*2 1*1*2 1*0*0
1*0*1 1*1*0 0*1*0 0*1*1
0*0*1 1*1*1 0*1*2 0*0*0
0*0*1 1*0*0 1*1*1 0*1*2
0*1*0 1*1*0 1*1*2 0*0*0
0*0*2 1*0*1 0*1*1 1*0*2
1*0*1 0*1*0 1*1*0 0*1*1
0*0*2 0*1*2 1*1*2 0*0*0
1*1*1 1*0*0 0*0*1 1*0*2
COLUMNS

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
10	28	71	20	9	7	21	76	22	30	37
9	77	40	5	10	78	13	6	14	18	2
8	17	12	18	74	64	62	16	25	12	11
7	8	24	31	49	68	73	60	33	5	38
6	57	59	58	53	75	27	3	34	63	70
5	72	19	15	42	44	29	50	51	20	48
4	3	36	43	65	52	9	39	41	80	[illegible]
3	4	69	11	47	56	55	54	61	19	
2	46	66	67	14	35	7	2	13	10	
1	17	45	1	16	26	6	4	23	8	

p-rep design

	V1	V2	V3	V4	V5	V6	V7	V8	V9
Row5	3	8	29	2	6	11	4	1	5
Row4	15	22	18	4	1	2	26	3	9
Row3	10	14	1	4	23	3	27	20	2
Row2	2	3	19	21	4	13	24	16	1
Row1	1	4	7	2	28	12	3	25	17

Augmented RCBD

	LOC1	LOC2	LOC3	LOC4	LOC5	LOC6	Copies	Avg
Gen-3	1	1	1	1	2	2	8	1.3
Gen-4	1	1	1	2	1	2	8	1.3
Gen-6	1	1	2	1	2	1	8	1.3
Gen-7	1	2	2	1	1	1	8	1.3
Gen-9	1	1	1	2	2	1	8	1.3
Gen-11	1	1	1	2	1	2	8	1.3
Gen-12	1	2	1	1	2	1	8	1.3
Gen-13	1	1	2	2	1	1	8	1.3
Gen-16	1	2	1	1	1	2	8	1.3

Multi-Location P-rep Design

Example of LMM

Dataset: Sleep Study
A survey-based study of the sleeping habits of individuals

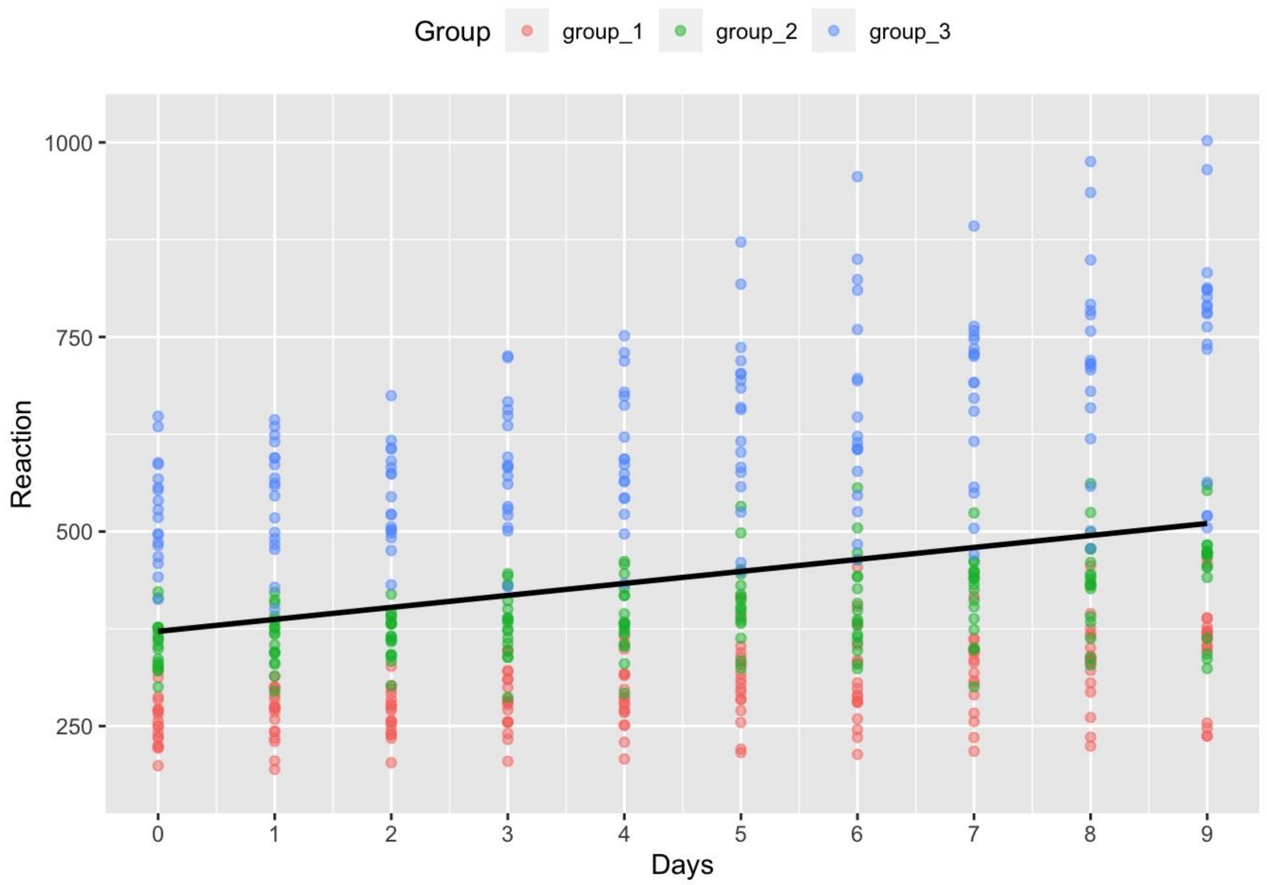
The sleepstudy data set is from an experiment that examined reaction times of sleep deprived individuals over the course of a few days (individuals only got 3h of sleep each day), with three different groups of people

```
head(sleepstudy,20)
```

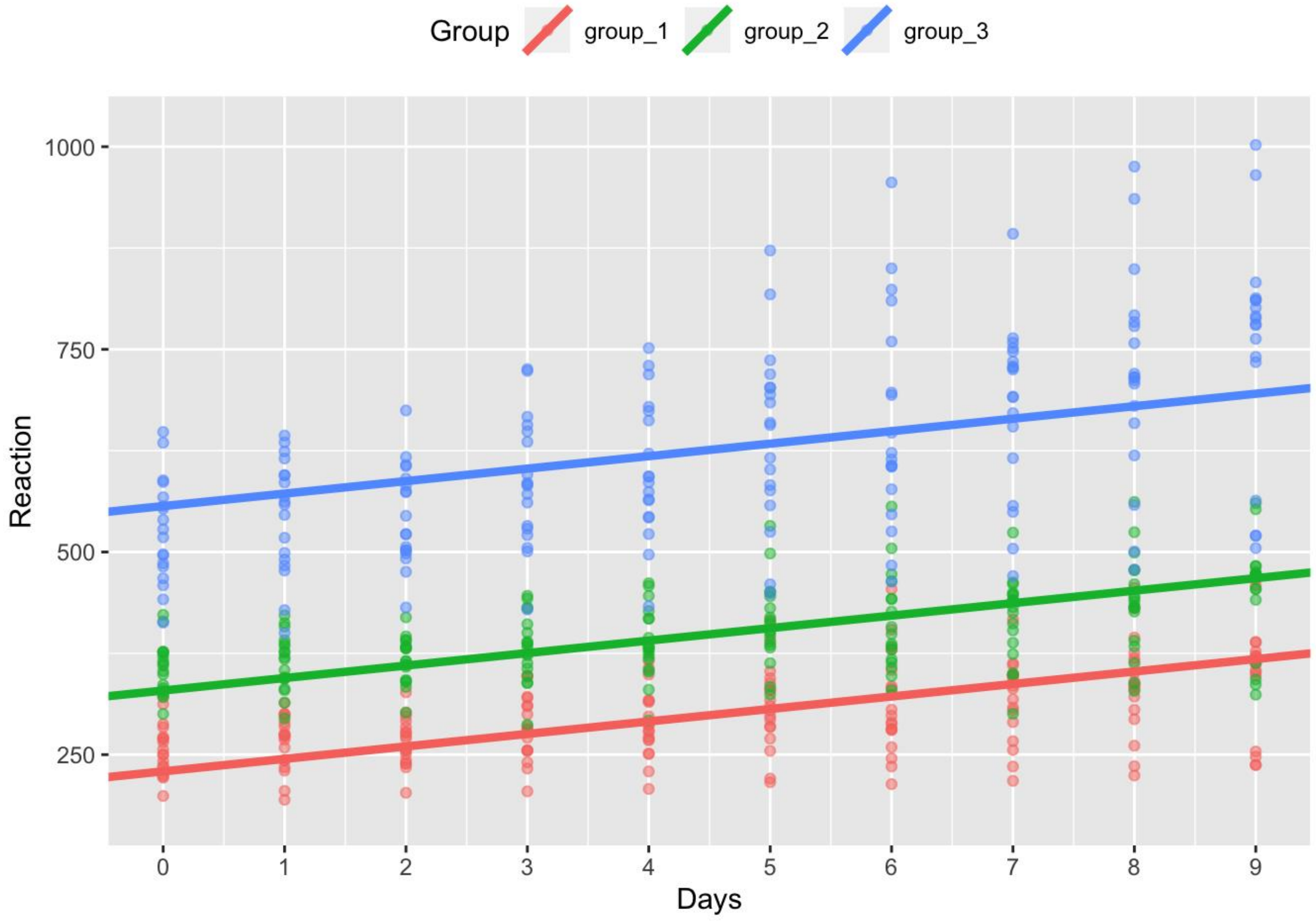
```
##    Reaction Days Subject
## 1  249.5600    0     308
## 2  258.7047    1     308
## 3  250.8006    2     308
## 4  321.4398    3     308
## 5  356.8519    4     308
## 6  414.6901    5     308
## 7  382.2038    6     308
## 8  290.1486    7     308
## 9  430.5853    8     308
## 10 466.3535    9     308
## 11 222.7339    0     309
## 12 205.2658    1     309
## 13 202.9778    2     309
## 14 204.7070    3     309
## 15 207.7161    4     309
## 16 215.9618    5     309
## 17 213.6303    6     309
## 18 217.7272    7     309
## 19 224.2957    8     309
## 20 237.3142    9     309
```

https://glennwilliams.me/r4psych/

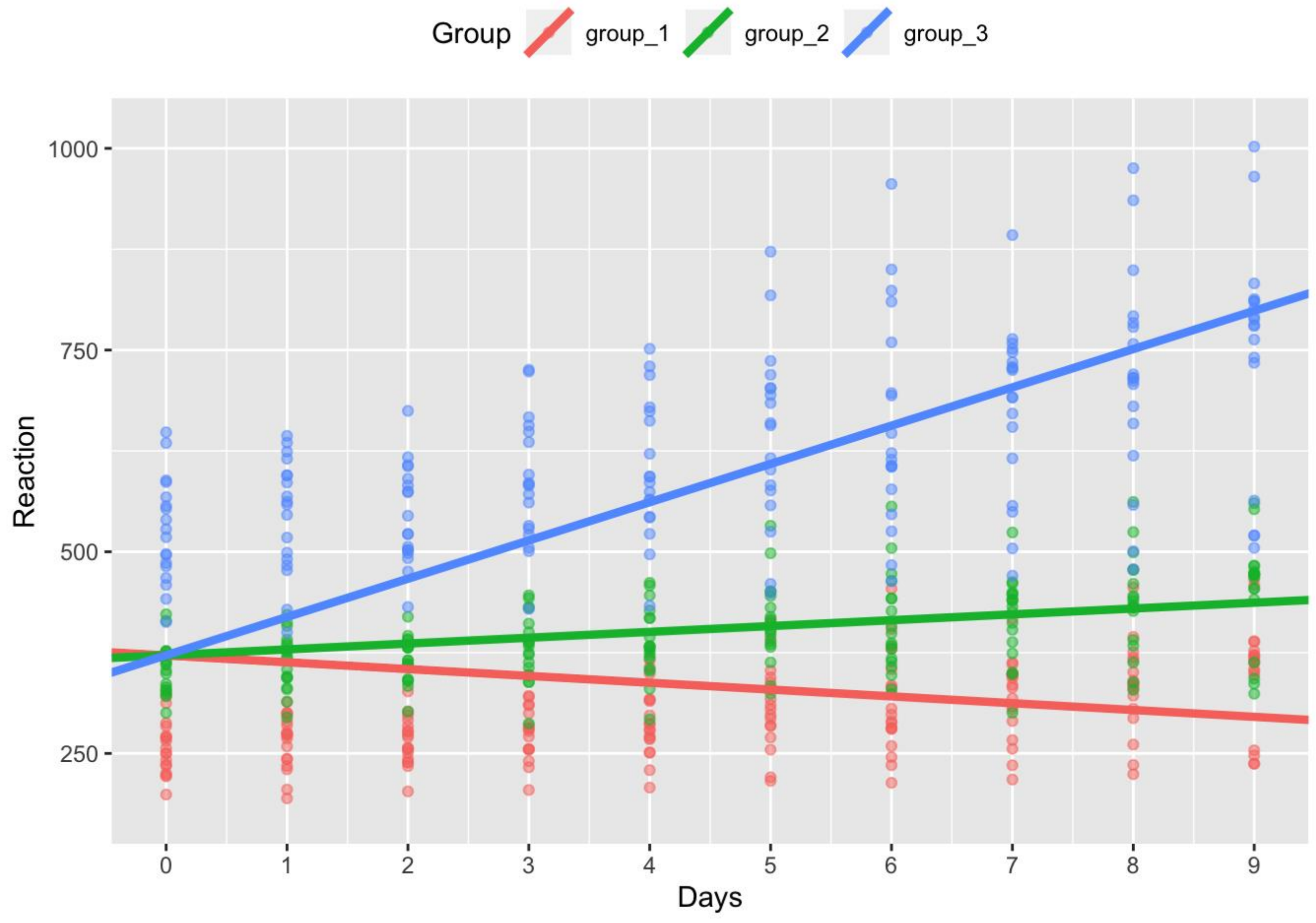
Fix slope and intercept



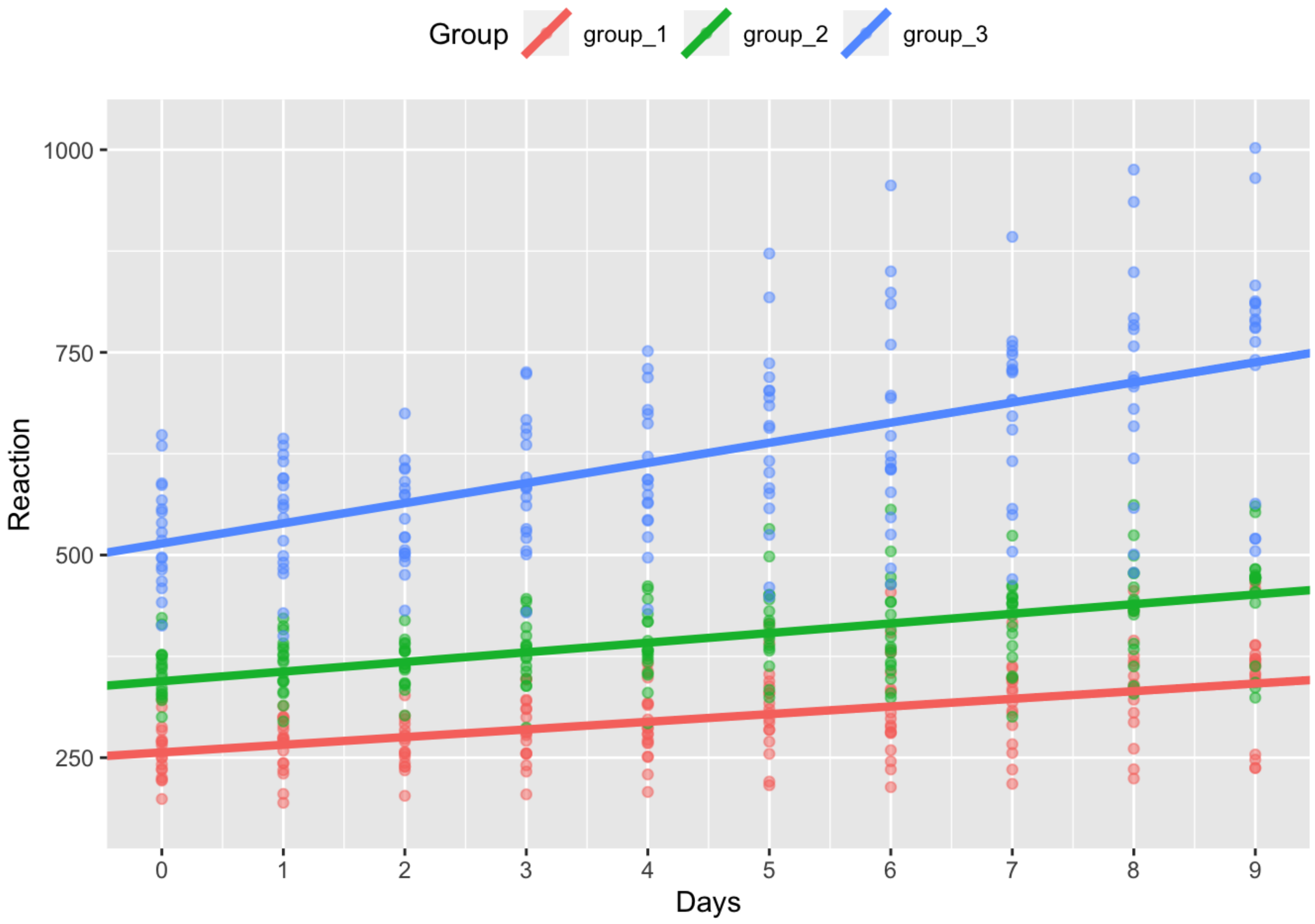
Random intercept

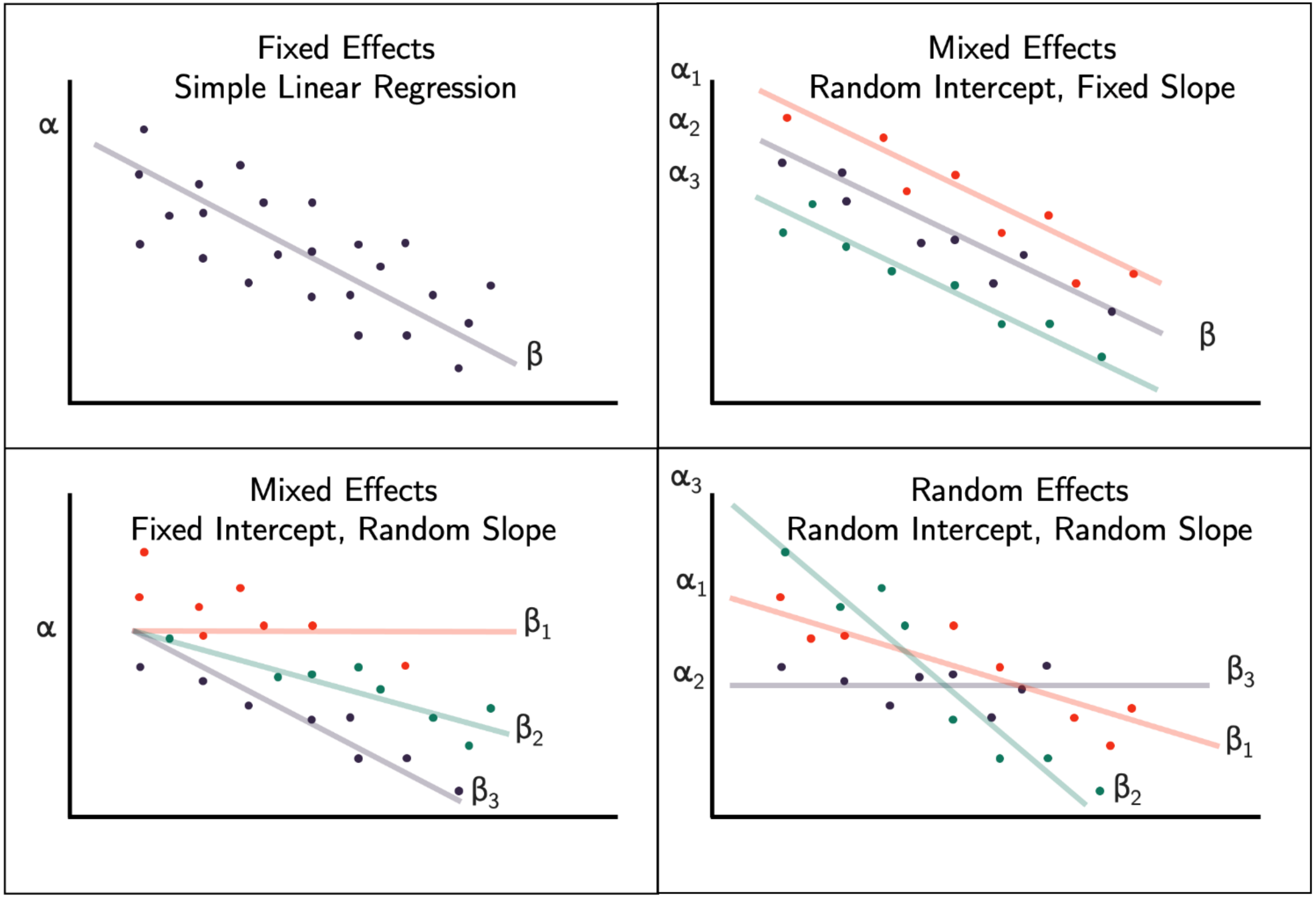


Random slope



Random intercept and slope




Fixed Effects
Simple Linear Regression
α
β
Mixed Effects
Random Intercept, Fixed Slope
α1
α2
α3
β
Mixed Effects
Fixed Intercept, Random Slope
α
β1
β2
β3
Random Effects
Random Intercept, Random Slope
α3
α1
α2
β3
β1
β2

“simplest, most complete”

Model selection

AIC (Akaike Information Criterion)

$$AIC = -2\ln(\text{Likelihood}) + 2k$$

•**Likelihood** measures how well the model explains the data.
•**k** is the number of parameters in the model.

BIC (Bayesian Information Criterion)

$$BIC = -2\ln(\text{Likelihood}) + k\ln(n)$$

•**n** is the number of observations in the dataset.
•The **penalty term** increases with the size of the dataset.

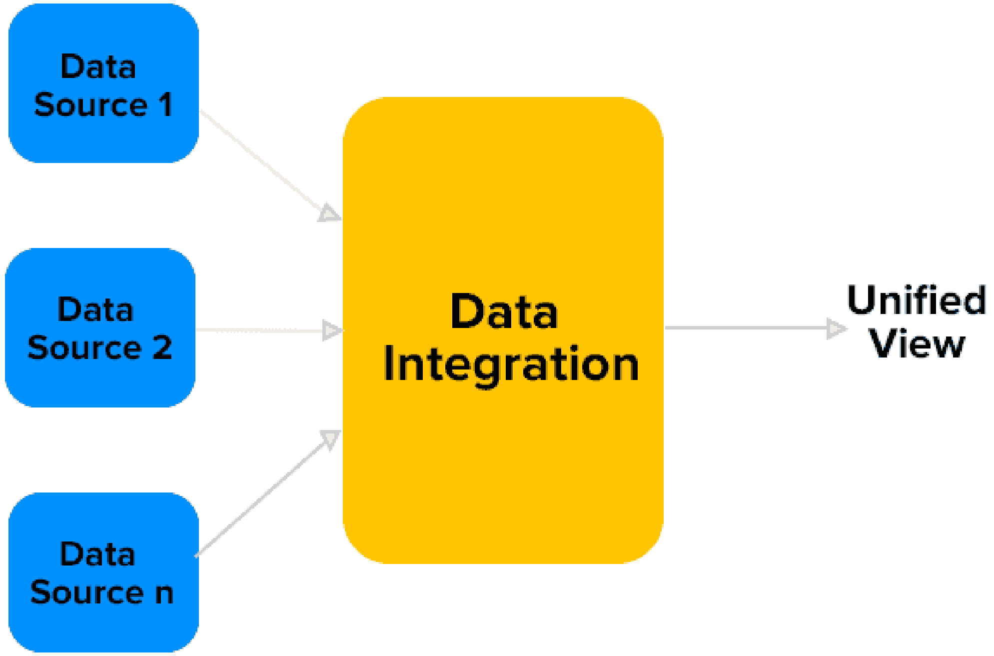
lower values indicating a better model

Data Integration and Analysis

combining or merging of data from multiple sources in an effort to extract better and/or more information

merging data by common data elements

linking data sets through a common factor



example: meta analysis