LGN 5822 - Biometrical Genetics

L04 – Linear Models

Michele Jorge Silva Siqueira

2023

Linear Models

Motivation

- Linear models are a class of statistical models that are used to describe the relationship between a dependent variable (or response) and one or more independent variables (or predictors) through a linear relationship

- A model is an **approximation** of reality

Linear Models

Motivation

- Linear models are relatively simple and provide an easily interpreted mathematical formula

Motivation

To reflect

"The most that can be expected from any model is that it can supply a useful approximation to reality: All models are wrong; some models are useful" George Box[1]

George Box (1919-2013)

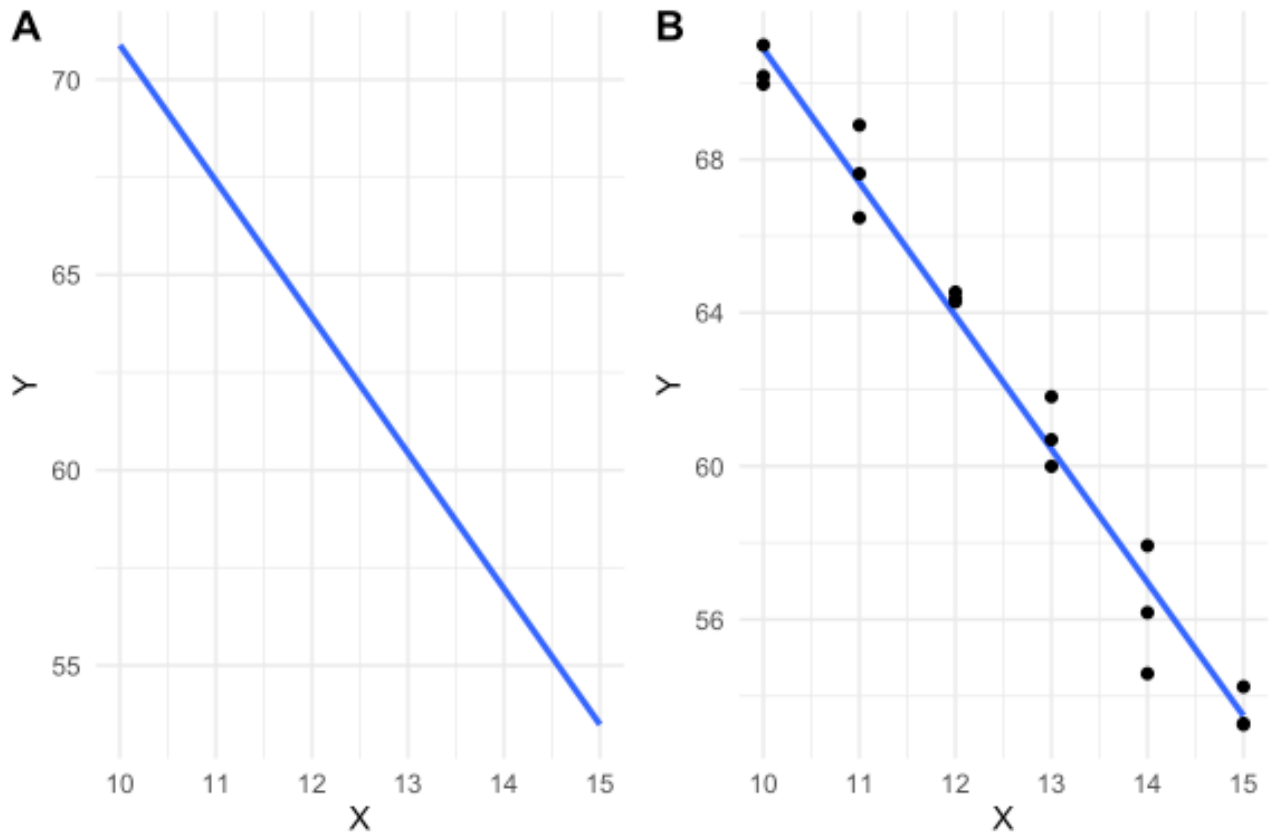
Linear Models

Motivation

Models are never perfect!

Linear Models

Motivation: Whats is the difference?

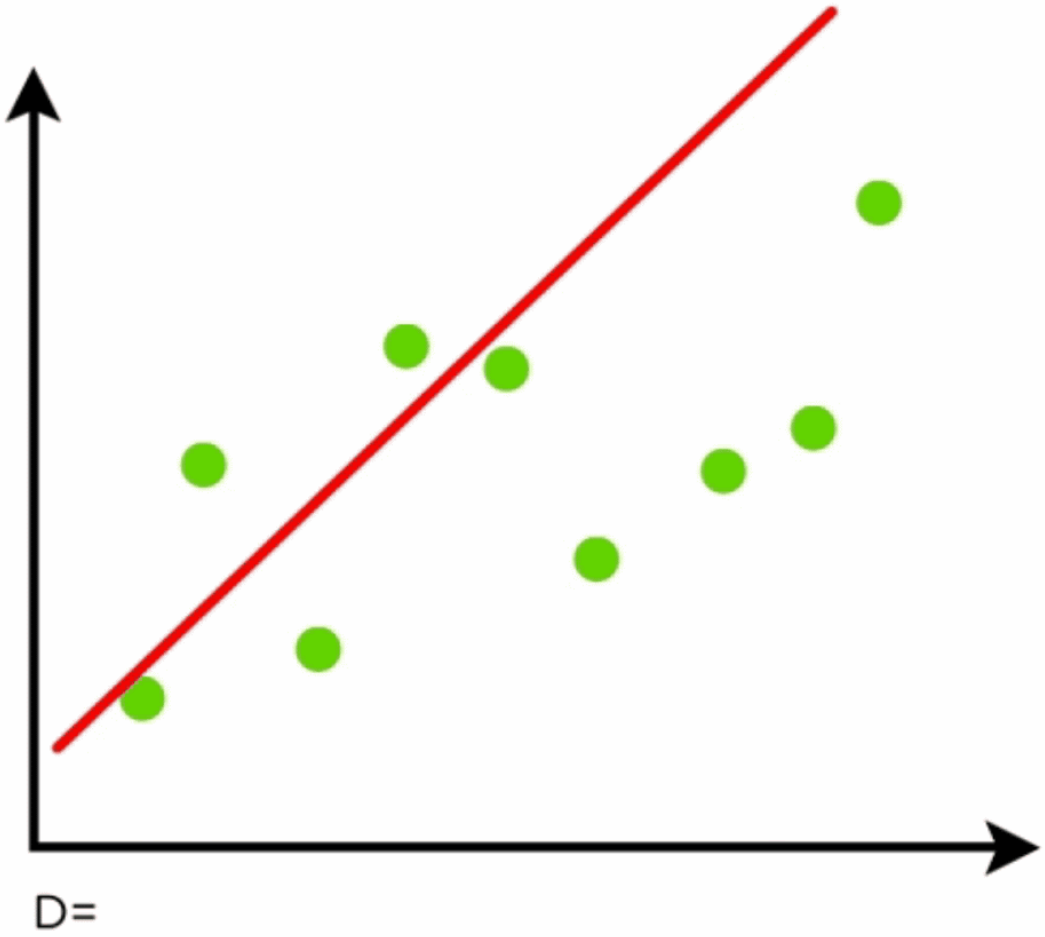


The line in **Figure A** is just a line, but the line in **Figure B** is a linear model fit to the data

Linear Models

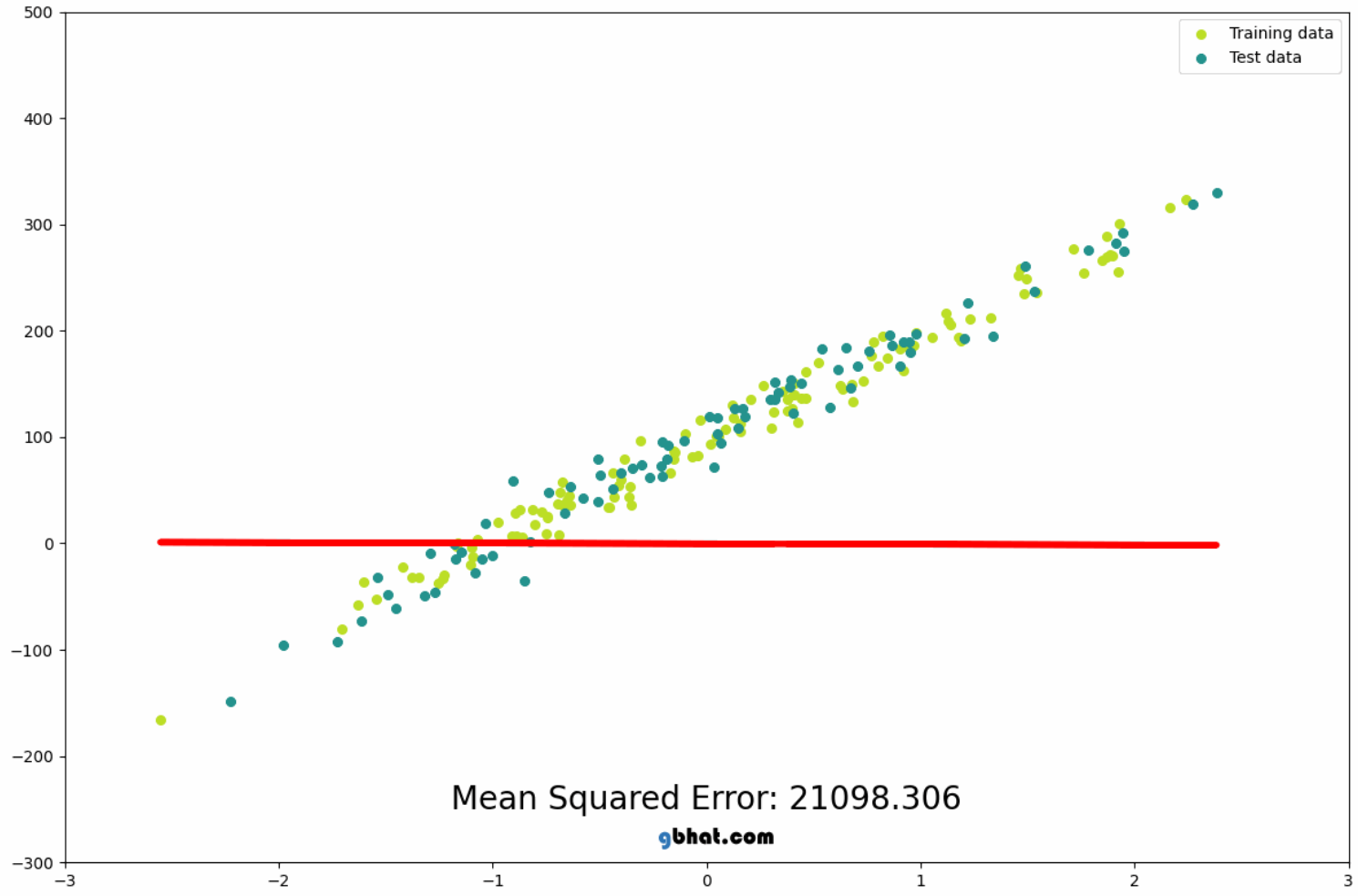
Motivation

Residual = Real Value - Predicted Value



Linear Models

Motivation



Which equation will best fit the straight line to the set of data points?

Linear Models

Motivation

- We will focus on **analysis of variance** (ANOVA) models
  - These are frequently used for experimental data analysis
- Most of the discussion of our course also applies to regression models

Linear Models

Linear models and linear regression are synonymous?

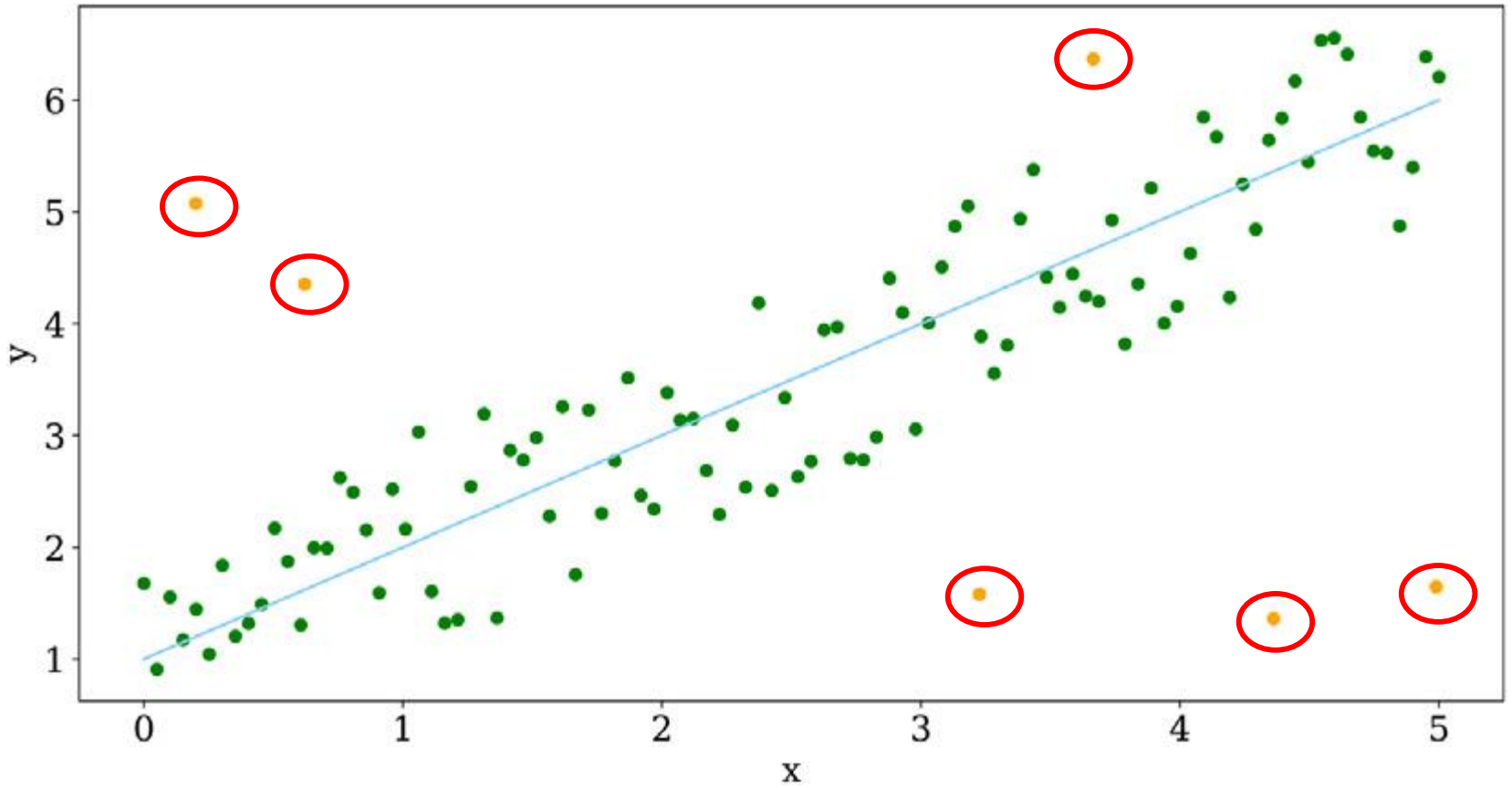
- **"Linear model"** is a large category of statistical models that describe the relationship between a dependent variable and one or more independent variables linearly

≠

- **"Linear regression"** is a specific technique within the category of linear models that focuses on modeling a linear relationship between a dependent variable and one or more independent variables.

Linear Models

Regression Models



The linear regression model (blue line) explains the relationship between the explanatory variable and the response variable

Linear Models

How to deal with outliers in linear models?

- **Outlier Identification:** Identify outliers in your data set using, for example, scatter plots
- **Data Transformations:** In some cases, it is possible to apply mathematical transformations to the data
- **Truncation or Cut:** Consider removing the most extreme outliers from the data set if they are found to be invalid values or measurement errors
- **Robustness:** More robust models are less sensitive to outliers
- **Cross Validation:** Use cross validation to evaluate how the model handles outliers (training and validation)

Linear Models

Why do linear models have significant importance in statistics?

- **Interpretability:** Linear models are relatively simple to understand

Linear Models

Why do linear models have significant importance in statistics?

- **Versatility:** Linear models can be applied to a wide variety of problems, from simple regression to classification problems such as regression

Linear Models

Why do linear models have significant importance in statistics?

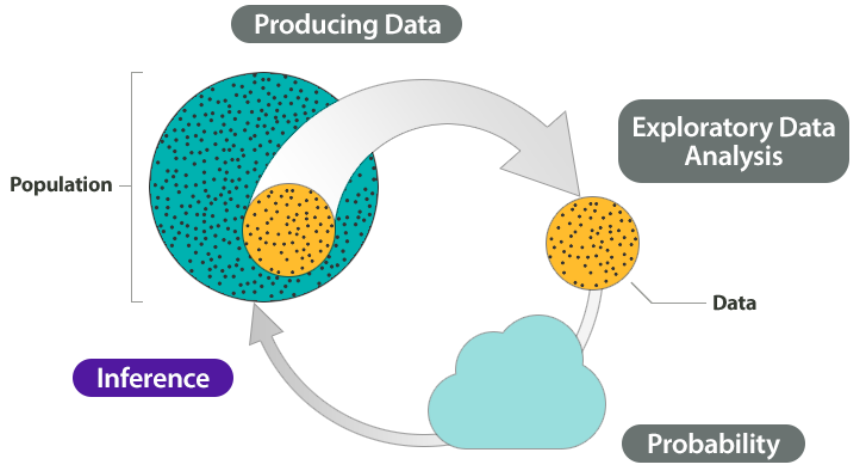
- **Computational Efficiency:** Linear models are computationally efficient and can be trained on large datasets with less computational and resource effort

Linear Models

Why do linear models have significant importance in statistics?

- **Statistical Inference:** Linear models allows the application of hypothesis tests and the obtaining of confidence intervals for the coefficients

- This is important when you want to make statistically significant statements about relationships between variables



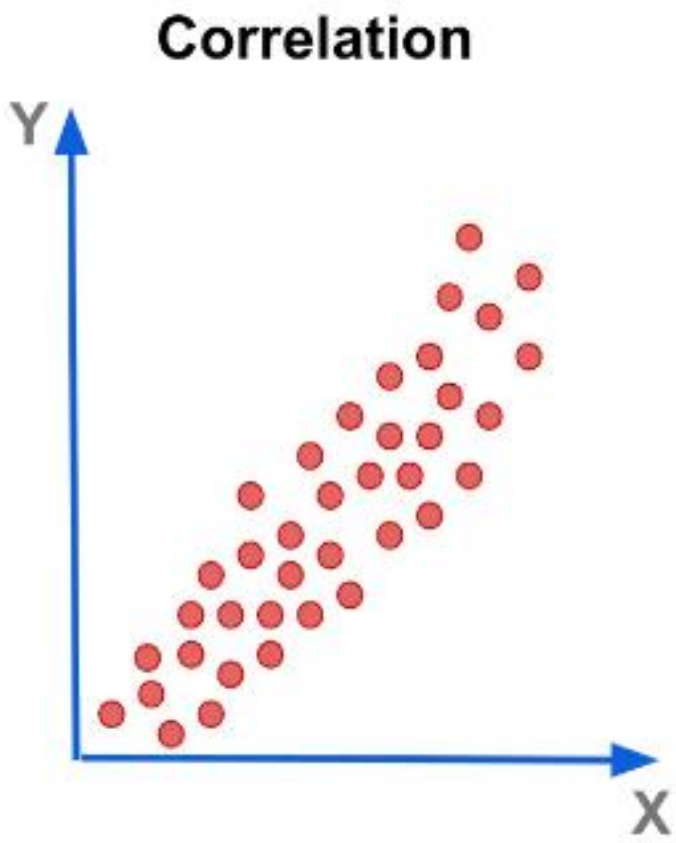
Linear Models

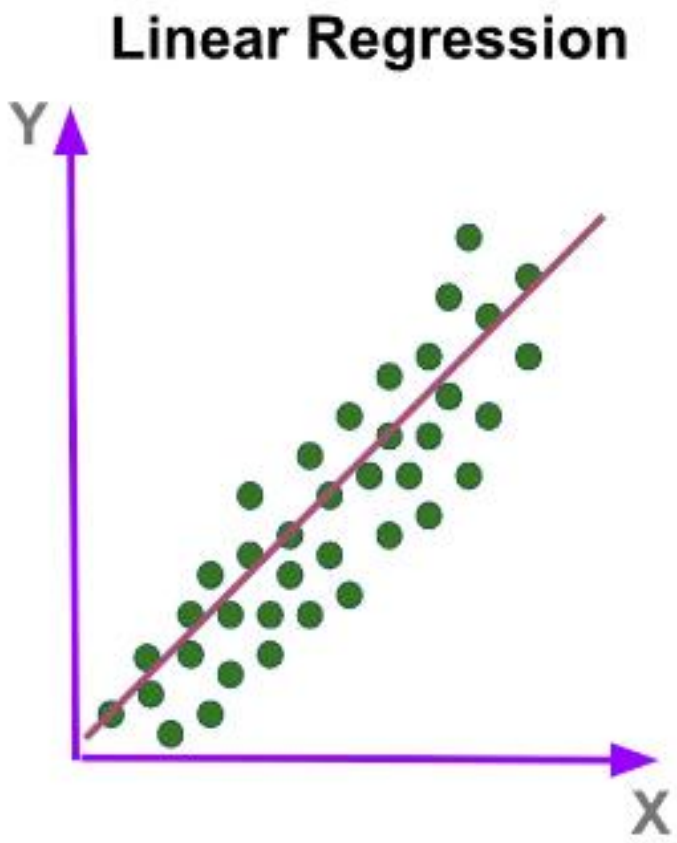
But…

- It is important to understand that linear models may not be the ideal choice for all types of data
- In cases where relationships are highly non-linear, more complex models such as *decision trees, neural networks or kernel methods (machine learning)* may be more appropriate

Linear Models

What is the difference between correlation and regression analysis?





Linear Models

What is Linear Regression?

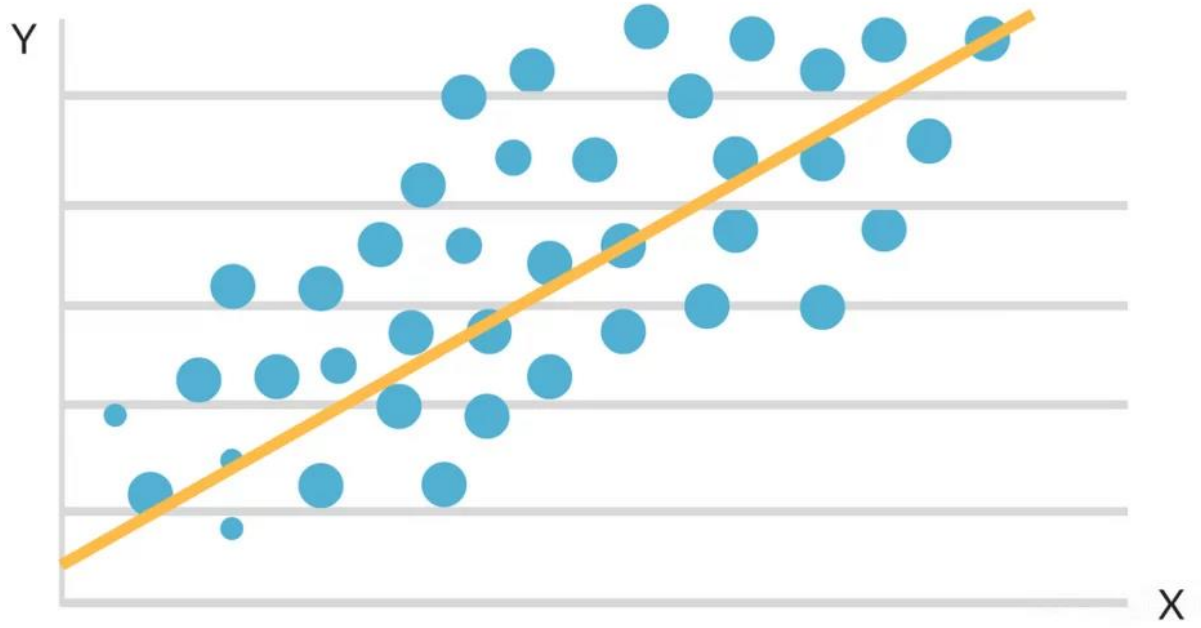
- Linear regression is a statistical technique used to model the relationship between a dependent variable (or response) and one or more independent variables (or predictors) linearly

- Linear regression analysis is used to predict the **value of a variable** based on the value of **another variable**

Linear Models

What is Linear Regression?

In resume, linear regression is to find the best line (or hyperplane, in cases of multiple independent variables) that fits the data to make predictions or inferences



Linear Models

Simple Linear Regression Models

- Simple Linear Regression Model describes the linear relationship between a dependent variable (y) and a single independent variable (X)

$$y = \beta_0 + \beta_1 X + \varepsilon$$

where:
y is the dependent or study variable
X is the independent or explanatory variable
β_0 is an intercept coefficient
β_1 is the slope coefficient
ε is the residuals $e \sim N(0, \sigma^2)$

Linear Models

Simple Linear Regression Models

- Simple Linear Regression Model describes the linear relationship between a dependent variable (y) and a single independent variable (X)

- Objective: Minimize the error term $\boldsymbol{\varepsilon}$, that is, try to get the predicted values as close as possible to the observed values $\boldsymbol{y}$

Multiple Linear Regression Models

- When the response y is often influenced by more than one predictor variable $(X_1, X_2, X_3 \ldots X_n)$

For example, the yield of a crop may depend on the amount of nitrogen, potash, and phosphate fertilizers used

Linear Models

Multiple Linear Regression Models

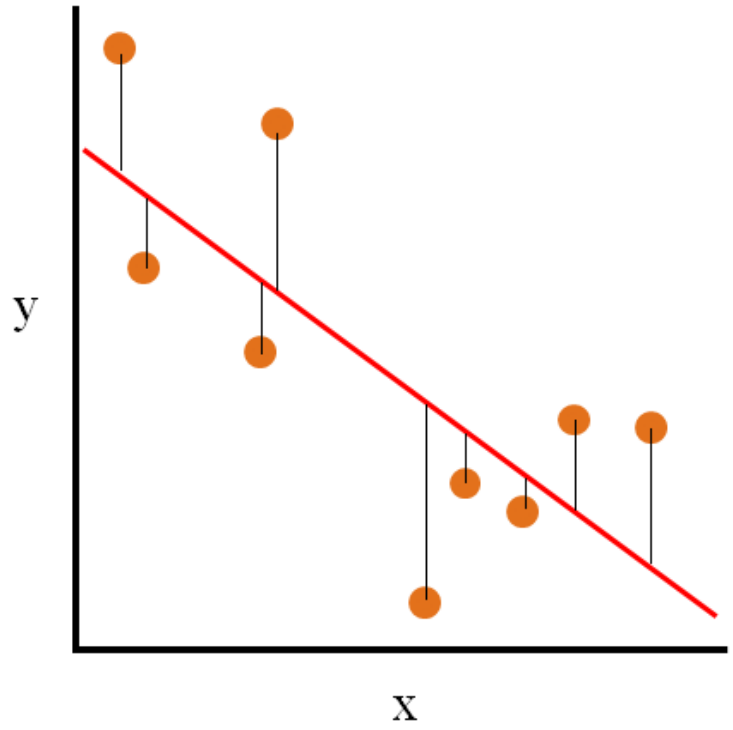
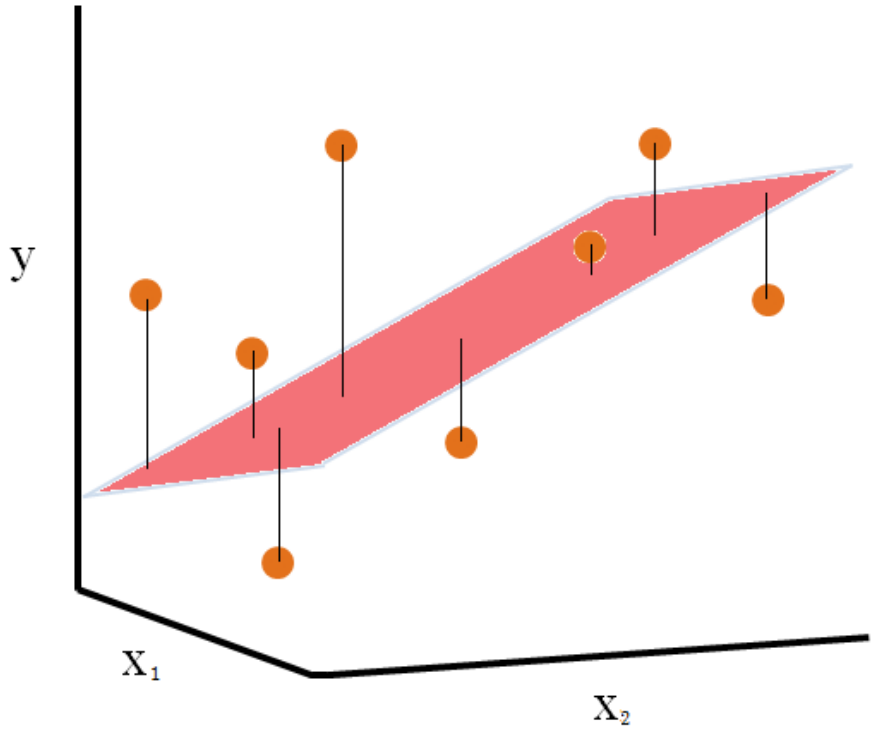
- Models the relationship between a dependent variable (y) and two or more independent variables ($X_1, X_2, X_3 \ldots X_n$)

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

y is the dependent or study variable
X are the independent or explanatory variable
β_0 is an intercept coefficient (the estimated value of Y when X is equal to zero)
β_k is the slope coefficients for each explanatory variable (rate of change in Y for one unit of change in X)
ε is the residuals

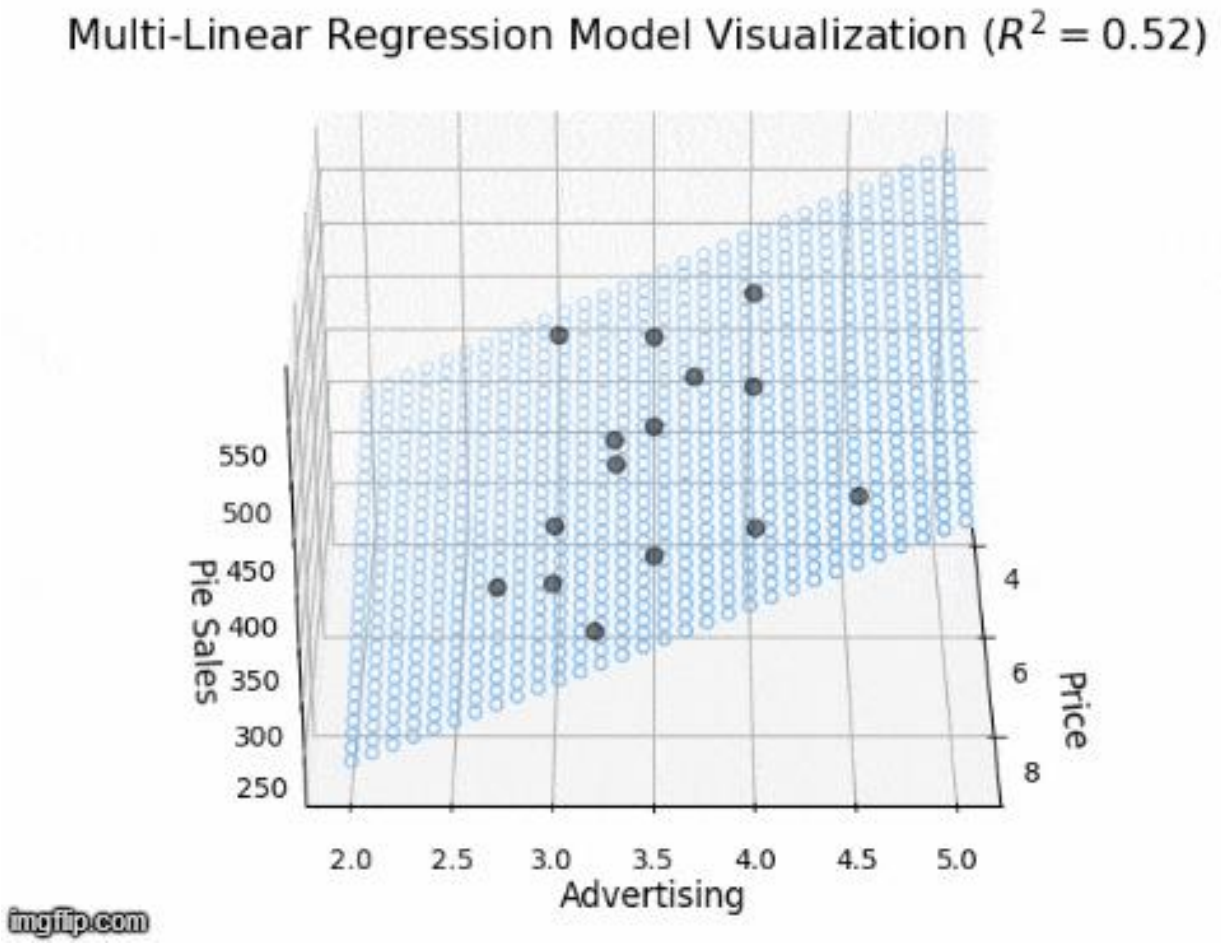
Linear Models

Simple x Multiple Linear Regression

Simple Linear Regression	Multiple Linear Regression (2 Independent Variables (x_1 , x_2))
	

Linear Models

Simple x Multiple Linear Regression



Linear Models

- Let´s Practice 01!

10 values of a variable X (temperature) were observed and the corresponding Y values (plant growth)

X	25	24.5	22	19.6	19	21	22.8	24	25.6	21.9
Y	5	2	3	3.8	4.9	5.1	3.5	4	5	2.3

#Define the model and parameter estimation

#use "lm ( ) function"

#Represent the points graphically (scatter plot) to see if there is apparent linear relationship between X and Y

#use "plot ( ) function"

Linear Models

- Let´s Practice 01!

```r
# Create a simple dataset
X = c(25, 25.4, 22, 19.6, 19, 21, 22.8, 24, 25.6, 21.9)
Y = c(5, 42, 3, 3.8, 4.9, 5.1, 3.5, 4, 5, 2.3)

# Perform simple linear regression
model <- lm(Y ~ X, data = data)

# Display the model summary
summary(model)

plot(X,Y)
```

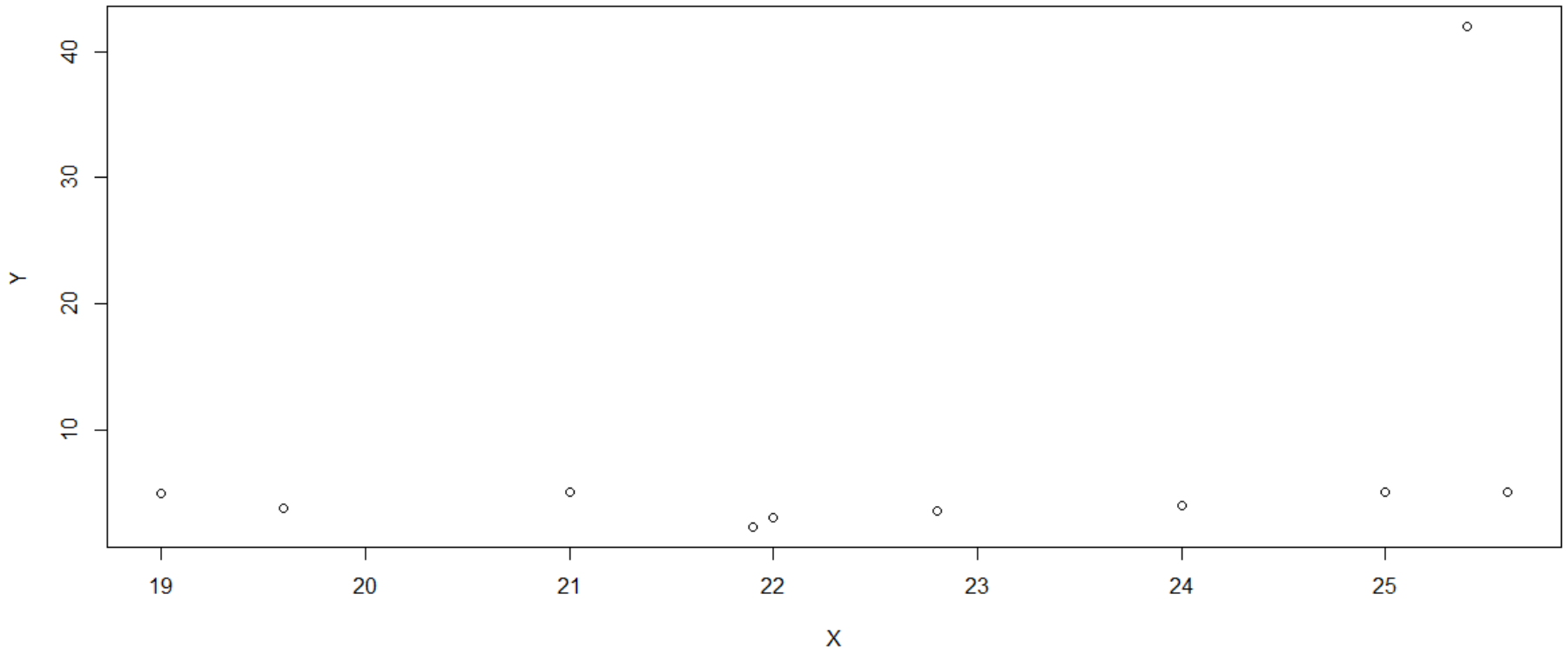
Linear Models

- Let´s Practice 01!

```
Call:
lm(formula = Y ~ X, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-9.276 -6.296 -3.741  2.055 28.156

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -41.028     37.208  -1.103    0.302
X              2.160      1.636   1.320    0.223

Residual standard error: 11.57 on 8 degrees of freedom
Multiple R-squared:  0.1789,    Adjusted R-squared:  0.07628
F-statistic: 1.743 on 1 and 8 DF,  p-value: 0.2232
```

Linear Models

- Let´s Practice 01!



Linear Models

Let´s Practice 02!

Considering an experiment with maize, the following were observed:

- 7 repetitions of variables:
- Y=Plant weight, X1=Dry Matter, X2=Average Diameter, X3=Average Height and X4=Number of sheets

Y	X1	X2	X3	X4
0,25	12	36	26	38
0,45	15	38	27	45
0,23	12.5	39	28	44
0,10	11	35.5	25	43.5
0,15	15	31	24	39
0,17	10	32	22	35
0,18	14	31.8	21	38

Linear Models

- Let´s Practice 02!

Considering an experiment with maize, the following were observed:

- 7 repetitions of variables:
- Y=Plant weight, X1=Dry Matter, X2=Average Diameter, X3=Average Height and X4=Number of sheets

Create a data frame with the data
#Define the model and parameter estimation
#use "lm ( ) function"

#Represent the points graphically (scatter plot) to see if there is apparent linear relationship between X and Y
#use "plot ( ) function"

Linear Models

- Let´s Practice 02!

```r
X1 = c(12,15,12.5,11,15,10,14)
X2 = c(36,38,39,35.5,31,32,31.8)
X3 = c(26,27,28,25,24,22,21)
X4 = c(38,45,44,43.5,39,35,38)
Y = c(0.25,0.45,0.23,0.10,0.15,0.17,0.18)

# Create a data frame with the data
data <- data.frame(X1, X2, X3, X4, Y)

# Perform multiple linear regression
model <- lm(Y ~ X1 + X2 + X3 + X4, data = data)

## Display the model summary
summary(model)
```

Linear Models

- Let´s Practice 02!

```
Call:
lm(formula = Y ~ X1 + X2 + X3 + X4, data = data)

Residuals:
       1        2        3        4        5        6
-0.020477  0.094573 -0.066446 -0.010990 -0.007082  0.065271
       7
-0.054849

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.93666    0.55360  -1.692    0.233
X1           0.04490    0.02663   1.686    0.234
X2           0.05372    0.03726   1.442    0.286
X3          -0.02003    0.04076  -0.492    0.672
X4          -0.01959    0.02117  -0.925    0.452

Residual standard error: 0.103 on 2 degrees of freedom
Multiple R-squared:  0.7255,    Adjusted R-squared:  0.1764
F-statistic: 1.321 on 4 and 2 DF,  p-value: 0.4737
```

Building Models

- Let´s Practice 03!

- Imagine you want to compare the yield of two particular genotypes
- For simplicity, assume you collect data from four field replicates for each genotype

Genotype A		Genotype B	
2.8	3.2	4.1	3.9
3.2	2.8	4.0	3.6

#Test an initial hypothesis
#Compare the means of each genotype
#Graph to observe the behavior of the 02 genotypes

Building Models

Initial Hypothesis Testing

- How can we compare the means of each genotype?
  - t- test

Building Models

Initial Hypothesis Testing

- How can we compare the means of each genotype?
  - t- test

```
yield_1 <- c(2.8, 3.2, 3.2, 2.8)
yield_2 <- c(4.1, 3.9, 4.0, 3.6)
t.test(yield_1, yield_2, var.equal = TRUE)
```

Building Models

Initial Hypothesis Testing

- How can we compare the means of each genotype?

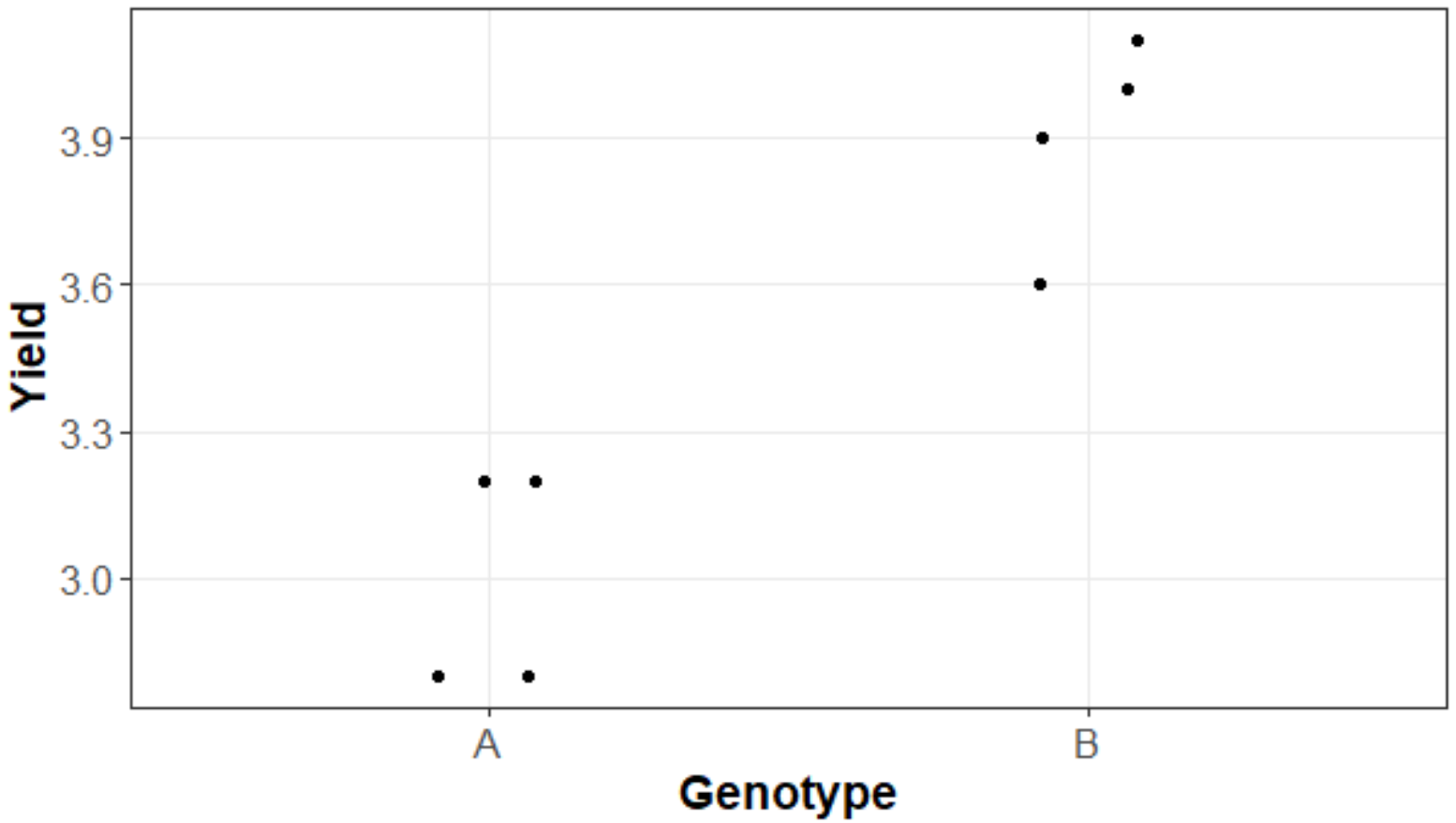
```
        Two Sample t-test

data:  yield_1 and yield_2
t = -5.6921, df = 6, p-value = 0.001269
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.2868907 -0.5131093
sample estimates:
mean of x mean of y
      3.0       3.9
```

Building Models

- Graph to observe the behavior of the 02 genotypes

```
yield_1 <- c(2.8, 3.2, 3.2, 2.8)
yield_2 <- c(4.1, 3.9, 4.0, 3.6)
yields <- c(yield_1, yield_2)
groups <- factor(rep(c("A", "B"), each = 4))
df <- data.frame(yield = yields, genotype = groups)


g <- ggplot(df, aes(genotype, yield)) +
  geom_jitter(width = 0.1, height = 0) +
  ylab("Yield") +
  xlab("Genotype") +
  theme_bw() +
  theme(axis.text=element_text(size=12),
        axis.title=element_text(size=14,face="bold"),
        legend.title=element_text(size=14,face="bold"),
        legend.text=element_text(size=12)) +
  theme(panel.grid.minor = element_blank())
g
```

Building Models



Building Models

- Can we simply model the average yields?
    - Let y_1 represent the yield of a plant with genotype **A** and y_2 the yield of a plant with genotype **B**

Building Models

- How can we express the models?

$$y_1 = \mu_1 + \varepsilon_1 \qquad y_2 = \mu_2 + \varepsilon_2$$

where μ_1 is the average yield of genotype A; μ_2 is the average of genotype B; ε_1 and ε_2 are random error terms

Building Models

- Can we model deviations from a common mean?

  - Denote a common intercept by μ and the effects of genotypes A and B by and τ_1 and τ_2, respectively

The model can then be expressed as:

$$y_1 = \mu_1 + \tau_1 + \varepsilon_1 \qquad y_2 = \mu_2 + \tau_2 + \varepsilon_2$$

where the values are as previously defined

Building Models

Because we have observations from four replicates of each genotype, we can write a model for each of the observations as:

$$y_{11} = \mu_1 + \tau_1 + \varepsilon_{11} \qquad y_{12} = \mu_1 + \tau_1 + \varepsilon_{12}$$

$$y_{13} = \mu_1 + \tau_1 + \varepsilon_{13} \qquad y_{14} = \mu_1 + \tau_1 + \varepsilon_{14}$$

$$y_{21} = \mu_1 + \tau_2 + \varepsilon_{21} \qquad y_{22} = \mu_1 + \tau_2 + \varepsilon_{22}$$

$$y_{23} = \mu_1 + \tau_3 + \varepsilon_{23} \qquad y_{24} = \mu_1 + \tau_3 + \varepsilon_{24}$$

Building Models

Linear Model

Equivalently, we can write:

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij} \qquad i = 1,2, \qquad j = 1,2,3,4$$

where y_{ij} is the observed yield of the j_{th} plant of the i_{th} genotype and ε_{ij} is the associated random error

Building Models

- How to represent the eight equations in matrix form?

Building Models

- How to represent the eight equations in matrix form?

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{24} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{14} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \\ \varepsilon_{24} \end{bmatrix}$$

Verify!

Linear Models

Exploring more… Linear Model!

Linear Models

Tradicional Linear Model

$$y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$i = 1, \ldots, n$$

where:
y_i is the dependent or study variable (is the value of the response variable)
X_i is the independent or explanatory variable (a known constant)
β_0 is an intercept coefficient (unknown parameter)
β_1 is the slope coefficient (unknown parameter)
ε_i is the residuals

Linear Models

Tradicional Linear Model

We assume that y_i and ε_i are random variables and that the values of x_i are known constants, which means that the same values of $x_1, x_2, \ldots, x_n$ would be used in repeated sampling

Linear Model in Matrix Form

- In general, we can write a linear model in matrix form as:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

where:

$\boldsymbol{y}$ is an $n\ x\ 1$ vector of observed values (is the dependent or study variable)
$\boldsymbol{X}$ is an $n\ x\ p$ **design matrix** (independent variable)
$\boldsymbol{\beta}$ is a $p\ x\ 1$ vector of **unknown** parameters
$\boldsymbol{\varepsilon}$ is an $n\ x\ 1$ vector of **unknown** errors

Building Models

Residues Assumptions

$i)\ E(\varepsilon_{ij}) = 0$ for all ij (presupposition of conditional expectation of residues)

$ii)\ var(\varepsilon_{ij}) = \sigma^2$ for all ij

$iii)\ cov\left(\varepsilon_{ij}, \varepsilon_{i'j'}\right) = 0$ for all $i \neq i'$ and $j \neq j'$

$iv)\ \varepsilon_{ij} \sim N\ (0, \sigma^2)$ for all i, j

Any of these assumptions may not be valid with real data

Building Models

$i)\ E(\varepsilon_{ij}) = 0$ for all ij

- Implying that y_i depends only on x_i and that all other variation in y_i is random
- Means that the expected value (mean) of the residuals (ε) for all observations (i) and all independent variables (j) is equal to zero

Building Models

$ii)\ var(\varepsilon_{ij}) = \sigma^2$ for all ij

- The variance of ε or y does not depend on the values of x_i (is also known as the assumption of **homoscedasticity**, homogeneous variance or constant variance)
- The dispersion or variability of the residuals must be the same for all combinations of values of the independent variables

Building Models

$iii)\ cov\left(\varepsilon_{ij}, \varepsilon_{i'j'}\right) = 0$ for all $i \neq i'$ and $j \neq j'$

- Indicates that the covariance between the residuals (ε) for all different observations ($i \neq i'$) and all different independent variables ($j \neq j'$) must be equal to zero

Building Models

$iv)\ \varepsilon_{ij} \sim N\ (0, \sigma^2)$ for all i, j

- Errors are normally distributed (normal distribution with **mean equal to zero and constant variance**)

- This is one of the fundamental assumptions of linear regression, known as the assumption of normality of residuals

Building Models

- Each assumption has been stated in terms of the ε´s or the y´s

For example, $var(\varepsilon_{ij})=\sigma^2$, then

$$var(y_i)=E[y_i - E(y_i)]^2 = E(y_i - \beta_0 - \beta_1 x_i)^2 = E(\varepsilon_i^2) = \sigma^2$$

Building Models

- Any of these assumptions may not be valid with real data
  - There are techniques for checking on the assumptions

Building Models

- What are the techniques for checking residual assumptions?

Building Models

- What are the techniques for checking residual assumptions?
  - **Shapiro-Wilk test:** Tests the normality of the residuals. A low W statistic suggests that the residuals do not follow a normal distribution

#If the p-value is greater than the significance level (usually 0.05), there is not enough evidence to reject the null hypothesis that the data follows a normal distribution

Building Models

#Data: Practice 03

Genotype A		Genotype B	
2.8	3.2	4.1	3.9
3.2	2.8	4.0	3.6

```
yield_1 <- c(2.8, 3.2, 3.2, 2.8)
yield_2 <- c(4.1, 3.9, 4.0, 3.6)

check_residual <- shapiro.test(yield_1)
check_residual

check_residual <- shapiro.test(yield_2)
check_residual


        Shapiro-Wilk normality test

data:  yield_1
W = 0.72863, p-value = 0.02386

> check_residual <- shapiro.test(yield_2)
> check_residual

        Shapiro-Wilk normality test

data:  yield_2
W = 0.92708, p-value = 0.5774
```

#Use shapiro.test ( ) function to test

#If the p-value is greater than the significance level (usually 0.05), there is not enough evidence to reject the null hypothesis that the data follows a normal distribution

Building Models

- What are the techniques for checking residual assumptions?
  - **Durbin-Watson test:** Evaluates the autocorrelation of residuals. Values close to 2 indicate independence of the residuals

Building Models

#install.packages("lmtest")

#Data: Practice 02

```
X1 = c(12,15,12.5,11,15,10,14)
X2 = c(36,38,39,35.5,31,32,31.8)
X3 = c(26,27,28,25,24,22,21)
X4 = c(38,45,44,43.5,39,35,38)
Y = c(0.25,0.45,0.23,0.10,0.15,0.17,0.18)
```

#Create a data frame with the data

```
data <- data.frame(X1, X2, X3, X4, Y)
```

Perform multiple linear regression

```
model <- lm(Y ~ X1 + X2 + X3 + X4, data = data)

result <- dwtest(model)
result

        Durbin-Watson test

data:  model
DW = 2.9182, p-value = 0.8967
alternative hypothesis: true autocorrelation is greater than 0
```

#A p-value smaller than your chosen significance level (usually 0.05) may suggest the presence of first-order autocorrelation

Building Models

- What are the tests for checking residual assumptions?
  - Box-Cox or logarithmic **transformations** can be applied to the data to make residuals more normally distributed or to improve homoscedasticity

And others…

Linear Models

- We say that there is a functional relationship between two variables x and y if there is a function f such that $y = f(x)$

Linear Models

- There is a probability distribution for y for each value of the variable x, since we assume that there is a statistical relationship between x and y
- The mean of y varies systematically with respect to x

Linear Models

- To fit a simple linear regression we need to have at least 3 observations:
    - If we only have 2 observations (2 points), the determination of the line is a problem of analytic geometry
        - It is not possible, in this case, to make any statistical analysis

Linear Models

- We must also check whether the number of **available observations** is greater than **the number of parameters** in the regression equation

- The general rule of is that **n** should be at least **10 to 20 times** larger than **p** to obtain robust results and avoid overfitting problems

#What is overfitting?
The model is overfitted to the training data, resulting in unstable and unreliable coefficient estimates

Linear Models

- Let´s Practice 04!

  - Create a vector of “X” values
  - Apply the quadratic function: $y = f(x) = x^2$ and to get the y values
  - Create a scatter plot

Linear Models

- Let´s Practice 04!

```
> # Create a vector of x values
> x <- seq(-10, 10, by = 0.1)  # Values from -10 to 10 with an interv
al of 0.1
> x
  [1] -10.0  -9.9  -9.8  -9.7  -9.6  -9.5  -9.4  -9.3  -9.2  -9.1
 [11]  -9.0  -8.9  -8.8  -8.7  -8.6  -8.5  -8.4  -8.3  -8.2  -8.1
 [21]  -8.0  -7.9  -7.8  -7.7  -7.6  -7.5  -7.4  -7.3  -7.2  -7.1
 [31]  -7.0  -6.9  -6.8  -6.7  -6.6  -6.5  -6.4  -6.3  -6.2  -6.1
 [41]  -6.0  -5.9  -5.8  -5.7  -5.6  -5.5  -5.4  -5.3  -5.2  -5.1
 [51]  -5.0  -4.9  -4.8  -4.7  -4.6  -4.5  -4.4  -4.3  -4.2  -4.1
 [61]  -4.0  -3.9  -3.8  -3.7  -3.6  -3.5  -3.4  -3.3  -3.2  -3.1
 [71]  -3.0  -2.9  -2.8  -2.7  -2.6  -2.5  -2.4  -2.3  -2.2  -2.1
 [81]  -2.0  -1.9  -1.8  -1.7  -1.6  -1.5  -1.4  -1.3  -1.2  -1.1
 [91]  -1.0  -0.9  -0.8  -0.7  -0.6  -0.5  -0.4  -0.3  -0.2  -0.1
[101]   0.0   0.1   0.2   0.3   0.4   0.5   0.6   0.7   0.8   0.9
[111]   1.0   1.1   1.2   1.3   1.4   1.5   1.6   1.7   1.8   1.9
[121]   2.0   2.1   2.2   2.3   2.4   2.5   2.6   2.7   2.8   2.9
[131]   3.0   3.1   3.2   3.3   3.4   3.5   3.6   3.7   3.8   3.9
[141]   4.0   4.1   4.2   4.3   4.4   4.5   4.6   4.7   4.8   4.9
[151]   5.0   5.1   5.2   5.3   5.4   5.5   5.6   5.7   5.8   5.9
[161]   6.0   6.1   6.2   6.3   6.4   6.5   6.6   6.7   6.8   6.9
[171]   7.0   7.1   7.2   7.3   7.4   7.5   7.6   7.7   7.8   7.9
[181]   8.0   8.1   8.2   8.3   8.4   8.5   8.6   8.7   8.8   8.9
[191]   9.0   9.1   9.2   9.3   9.4   9.5   9.6   9.7   9.8   9.9
[201]  10.0
```

Linear Models

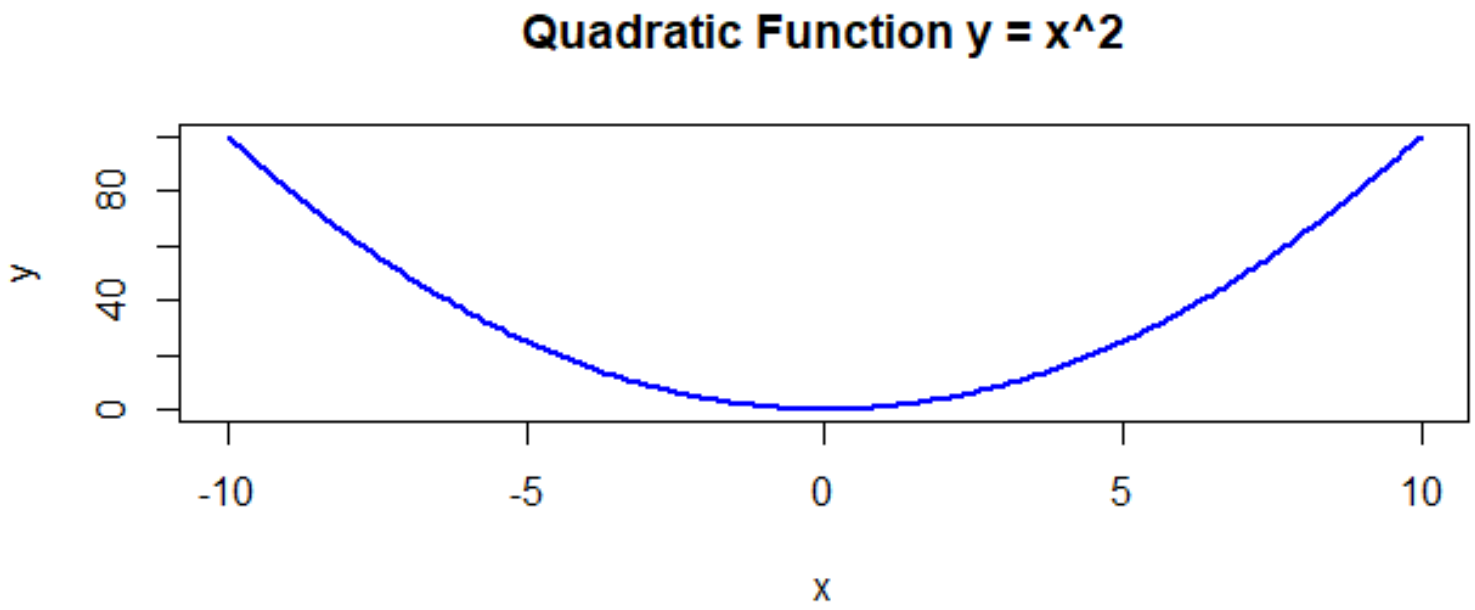
- Let´s Practice 04!

```
> # Apply the function f(x) to get the y values
> y <- x
> y
  [1] -10.0  -9.9  -9.8  -9.7  -9.6  -9.5  -9.4  -9.3  -9.2  -9.1
 [11]  -9.0  -8.9  -8.8  -8.7  -8.6  -8.5  -8.4  -8.3  -8.2  -8.1
 [21]  -8.0  -7.9  -7.8  -7.7  -7.6  -7.5  -7.4  -7.3  -7.2  -7.1
 [31]  -7.0  -6.9  -6.8  -6.7  -6.6  -6.5  -6.4  -6.3  -6.2  -6.1
 [41]  -6.0  -5.9  -5.8  -5.7  -5.6  -5.5  -5.4  -5.3  -5.2  -5.1
 [51]  -5.0  -4.9  -4.8  -4.7  -4.6  -4.5  -4.4  -4.3  -4.2  -4.1
 [61]  -4.0  -3.9  -3.8  -3.7  -3.6  -3.5  -3.4  -3.3  -3.2  -3.1
 [71]  -3.0  -2.9  -2.8  -2.7  -2.6  -2.5  -2.4  -2.3  -2.2  -2.1
 [81]  -2.0  -1.9  -1.8  -1.7  -1.6  -1.5  -1.4  -1.3  -1.2  -1.1
 [91]  -1.0  -0.9  -0.8  -0.7  -0.6  -0.5  -0.4  -0.3  -0.2  -0.1
[101]   0.0   0.1   0.2   0.3   0.4   0.5   0.6   0.7   0.8   0.9
[111]   1.0   1.1   1.2   1.3   1.4   1.5   1.6   1.7   1.8   1.9
[121]   2.0   2.1   2.2   2.3   2.4   2.5   2.6   2.7   2.8   2.9
[131]   3.0   3.1   3.2   3.3   3.4   3.5   3.6   3.7   3.8   3.9
[141]   4.0   4.1   4.2   4.3   4.4   4.5   4.6   4.7   4.8   4.9
[151]   5.0   5.1   5.2   5.3   5.4   5.5   5.6   5.7   5.8   5.9
[161]   6.0   6.1   6.2   6.3   6.4   6.5   6.6   6.7   6.8   6.9
[171]   7.0   7.1   7.2   7.3   7.4   7.5   7.6   7.7   7.8   7.9
[181]   8.0   8.1   8.2   8.3   8.4   8.5   8.6   8.7   8.8   8.9
[191]   9.0   9.1   9.2   9.3   9.4   9.5   9.6   9.7   9.8   9.9
[201]  10.0
```

Linear Models

- Let´s Practice 04!

```r
# Create a scatter plot
plot(x, y, type = "l", col = "blue", lwd = 2, xlab = "x", ylab = "y", main = "Quadratic Function y = x^2")
```



Parameter Estimation: Method of Least Squares

Using a random sample of n observations $y_1, y_2, \ldots, y_n$ and the accompanying fixed values $x_1, x_2, \ldots, x_n$, we can estimate the parameters β_0, β_1, and σ^2

Method of Least Squares

- The main objective of the method is to find the best estimates β_0 and β_1 that minimize the sum of the squares of the residuals (errors) between the observed values and the values predicted by the model

Method of Least Squares

- In the least-squares method, we seek estimators β_0 and β_1 that minimize the sum of squares of the deviations $(y_i - \hat{y}_i)$ of the n observed $y_i's$ from their predicted values, then $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

$$\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Method of Least Squares

- To find the values of β_0 and β_1 that minimize the sum of squares of the deviations, we derivatives with respect to β_0 and β_1

Method of Least Squares

- First, to find the derivatives $\hat{\varepsilon}'\hat{\varepsilon}$ (sum of the squares of the errors) with respect to β_0
  - and set the results equal to 0

$$\frac{\partial \hat{\boldsymbol{\varepsilon}}' \hat{\boldsymbol{\varepsilon}}}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0,$$

Method of Least Squares

- And, to find the derivatives $\hat{\varepsilon}'\hat{\varepsilon}$ (the sum of the squares of the errors) with respect to β_1
  - and set the results equal to 0

$$\frac{\partial \hat{\boldsymbol{\varepsilon}}' \hat{\boldsymbol{\varepsilon}}}{\partial \hat{\beta}_1} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0$$

Method of Least Squares

- The solution to β_0 and β_1 and is given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

x_i and y_i are the individual values of x and y in your data
$\bar{x}$ is the average of the x values
$\bar{y}$ is the average of the y values

Linear Models

- Let´s Practice 05!

Students in a statistics class claimed that doing the homework had no helped prepare them for the exam (y). The **exam score y and homewc score x** for the 18 students in the class were as follows:

y	x	y	x	y	x
95	96	72	89	35	0
80	77	66	47	50	30
0	0	98	90	72	59
0	0	90	93	55	77
79	78	0	18	75	74
77	64	95	86	66	67

#Find the values of β_0 and β_1: use coef( ) or lm ( ) functions
#Define the prediction equation
Create the regression graph (use plot function)

Linear Models

#Find the values of β_0 and β_1: use coef( ) or lm ( ) functions

```
# Example data loading
data <- data.frame(x = c(96,77,0,0,78,64,89,47,90,93,18,86,0,30,59,77,74,67),
                   y = c(95,80,0,0,79,77,72,66,98,90,0,95,35,50,72,55,75,66))
data

# Fit the linear regression model
model <- lm(y ~ x, data = data)
model

# Coefficients of the model
beta_0 <- coef(model)[1]  # Intercept (β0)
beta_1 <- coef(model)[2]  # Slope coefficient (β1)

beta_0
beta_1
```

Linear Models

#Find the values of β_0 and β_1: use coef( ) or lm ( ) functions

```
> # Fit the linear regression model
> model <- lm(y ~ x, data = data)
> model

Call:
lm(formula = y ~ x, data = data)

Coefficients:
(Intercept)            x
    10.7269       0.8726


> # Coefficients of the model
> beta_0 <- coef(model)[1]  # Intercept (β₀)
> beta_1 <- coef(model)[2]  # Slope coefficient (β₁)
> beta_0
(Intercept)
   10.72691
> beta_1
        x
0.8726465
```

Linear Models

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

$$= \frac{81{,}195 - 18(58.056)(61.389)}{80{,}199 - 18(58.056)^2} = .8726,$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 61.389 - .8726(58.056) = 10.73$$

#Define the prediction equation

$$\hat{y} = 10.73 + .8726x$$

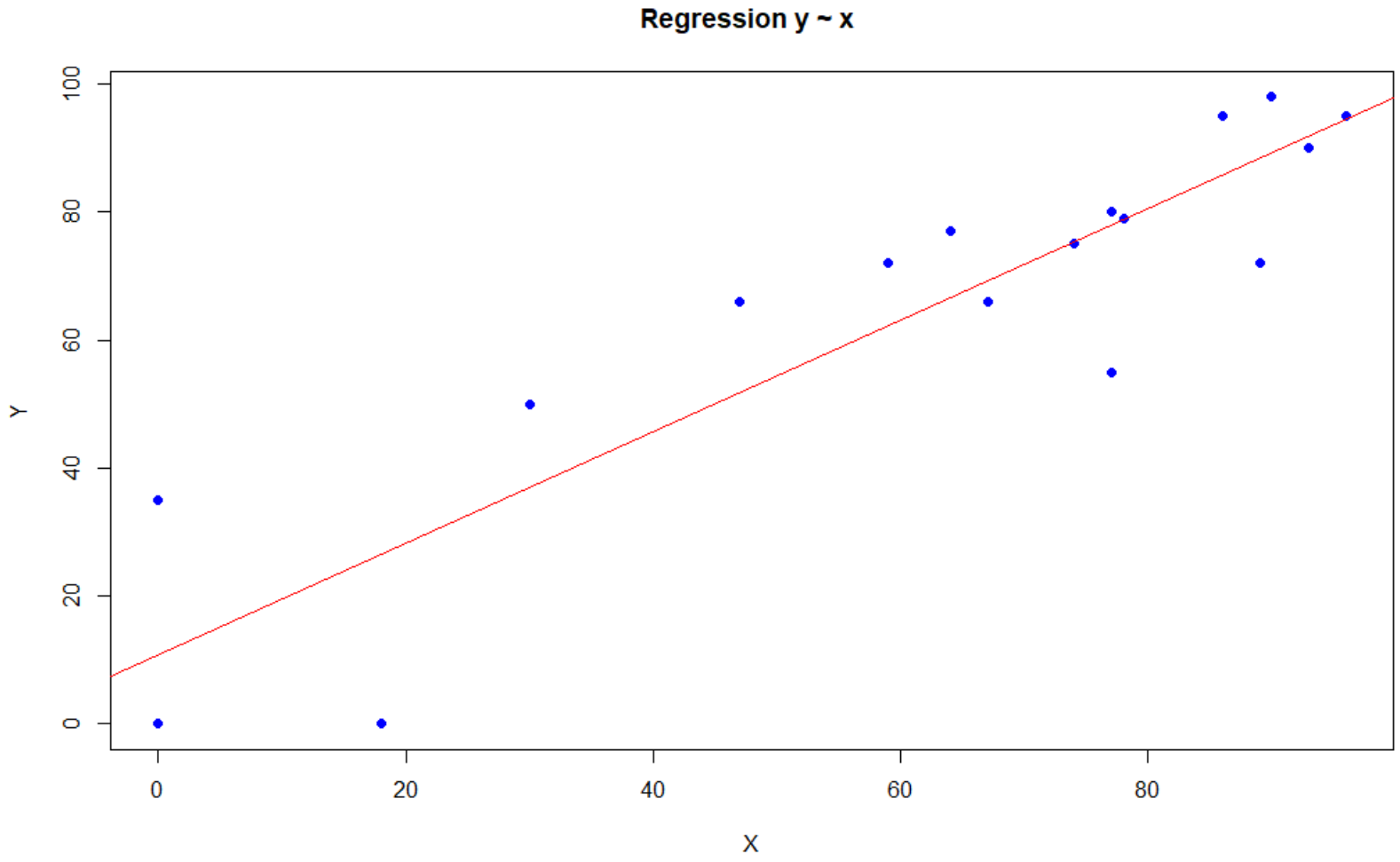
Linear Models

#Create the regression graph

```
# Graphic
plot(data$x, data$y, pch = 16, col = "blue", xlab = "X", ylab = "Y", main = "Regression y ~ x ")
abline(model, col = "red")
```

Linear Models

#Create the regression graph



$$\hat{y} = 10.73 + .8726x$$

Let´s Practice 06!

Remember the Practice 03:

#Calculate the matrix "X"
#Calculate $X´X$ and $X´Y$
#Calculate the coefficients in a regression model
#Estimate μ, τ_1 and τ_2

Genotype A		Genotype B	
2.8	3.2	4.1	3.9
3.2	2.8	4.0	3.6

Linear Models

#Calculate the matrix "X"

```
> y <- c(2.8, 3.2, 3.2, 2.8,4.1, 3.9, 4.0, 3.6)
> X <- matrix(c(1,1,1,1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,1,1,1,1), ncol=3)
> X
     [,1] [,2] [,3]
[1,]    1    1    0
[2,]    1    1    0
[3,]    1    1    0
[4,]    1    1    0
[5,]    1    0    1
[6,]    1    0    1
[7,]    1    0    1
[8,]    1    0    1
```

Linear Models

- Remember that:

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{24} \end{bmatrix} = \begin{bmatrix} 2.8 \\ 3.2 \\ 3.2 \\ 2.8 \\ 4.1 \\ 3.9 \\ 4.0 \\ 3.6 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{14} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \\ \varepsilon_{24} \end{bmatrix}$$

Linear Models

#Calculate $X'X$ and $X'Y$

```
> xlx <- t(X) %*% X
> xlx
     [,1] [,2] [,3]
[1,]    8    4    4
[2,]    4    4    0
[3,]    4    0    4
>
> xly <- t(X) %*% y
> xly
     [,1]
[1,] 27.6
[2,] 12.0
[3,] 15.6
```

%*% is used to perform matrix multiplication in R

Linear Models

#Calculate the coefficients in a regression model

#Calculate $\hat{\beta} = (X'X)^{-1}X'Y$

```
beta_hat <- solve(XlX)%*%XlY
```

the solve( ) function to calculate the inverse matrix of XlX

```
> beta_hat <- solve(XlX)%*%XlY
Error in solve.default(XlX) :
  Lapack routine dgesv: system is exactly singular: U[3,3] = 0
```

#the matrix has no inverse

Restrictions

Non-Full-Rank Models

- As currently defined, X is not of full rank
- Notice that the first column is equal to the sum of the other two

- We need **restrictions** to make $X'X$ nonsingular

  #nonsingular: is a square matrix in which its determinant is non-zero (no inverse)

Restrictions

- There are multiple possible restrictions

Restrictions

- There are multiple possible constraints
  - One such possibility is to set $\tau_1 = 0$ (effects of genotypes A)
  - The model is then expressed as:

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{24} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \tau_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{14} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \\ \varepsilon_{24} \end{bmatrix}$$

Restrictions

Genotype A		Genotype B	
2.8	3.2	4.1	3.9
3.2	2.8	4.0	3.6

Linear Models

Parameter Estimation with Restrictions: $\tau_1 = 0$ (effects of genotypes A)

#Calculate the matrix "X"

```
> y <- c(2.8, 3.2, 3.2, 2.8,4.1, 3.9, 4.0, 3.6)
> X <- matrix(c(1, 1, 1, 1, 1, 1, 1, 1,0, 0, 0, 0, 1, 1, 1, 1), ncol=2)
> X
     [,1] [,2]
[1,]    1    0
[2,]    1    0
[3,]    1    0
[4,]    1    0
[5,]    1    1
[6,]    1    1
[7,]    1    1
[8,]    1    1
```

Linear Models

#Calculate $X'X$ and $X'Y$

```
> xlx <- t(x) %*% x
> xlx
     [,1] [,2]
[1,]    8    4
[2,]    4    4
> xly <- t(x) %*% y
> xly
     [,1]
[1,] 27.6
[2,] 15.6
```

```
> beta_hat <- solve(xlx)%*%xly
> beta_hat
     [,1]
[1,]  3.0
[2,]  0.9
```

Linear Models

#Calculate the coefficients in a regression model

#Calculate $\hat{\beta} = (X'X)^{-1}X'Y$

```
> beta_hat <- solve(xlx)%*%xly
> beta_hat
     [,1]
[1,]  3.0
[2,]  0.9
```

Linear Models

#Calculate the coefficients in a regression model

- Or use the function lm

Linear Models

#Calculate the coefficients in a regression model

```
y <- c(2.8, 3.2, 3.2, 2.8,4.1, 3.9, 4.0, 3.6)

genotype <-factor(rep(c("A", "B"), each = 4))
genotype

model <- lm(y ~ genotype)
model
```

Linear Models

#Calculate the coefficients in a regression model

```
Call:
lm(formula = y ~ genotype)

Coefficients:
(Intercept)    genotypeB
        3.0          0.9
```

Linear Models

Parameter Estimation with Restrictions: Model without the intercept μ

```
model <-lm(y ~ 0 + genotype)
model
```

```
Call:
lm(formula = y ~ 0 + genotype)

Coefficients:
genotypeA  genotypeB
      3.0        3.9
```

Linear Models

Parameter Estimation with Restrictions: Model with the sum-to-zero

#The coefficients (or parameters) associated with the independent variables are estimated so that the sum of these coefficients is equal to zero
#to facilitate the interpretation of coefficients

```
contrasts(genotype) <- contr.sum
model <- lm(y ~ genotype)
model
```

```
Call:
lm(formula = y ~ genotype)

Coefficients:
(Intercept)    genotype1
       3.45        -0.45
```

Reduced and Full Models

- Continuing with the previous example, let us begin with the following **reduced model**:

$$y_{ij} = \mu + \varepsilon_{ij,} \qquad i = 1,2 \qquad j = 1,2,3,4$$

Reduced and Full Models

$$y_{ij} = \mu + \varepsilon_{ij,} \qquad i = 1,2 \qquad j = 1,2,3,4$$

- In matrix form

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{24} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} [\mu] + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{14} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \\ \varepsilon_{24} \end{bmatrix}$$

Reduced and Full Models

- For this reduced model, the only source of variation is the random error

Reduced and Full Models

- In that case, the **error sum of squares** (SSE) is equal to the **total sum of squares** (SST)
  - It has $n - 1$ associated degrees of freedom (correction for the mean)
  - Dividing the sum of squares by its corresponding degrees of freedom yields the **mean square**

Linear Models

Reduced Models

- In R

```
y <- c(2.8, 3.2, 3.2, 2.8,4.1, 3.9, 4.0, 3.6)
y
reduced_model <- lm(y ~ 1)
reduced_model
```

```
Call:
lm(formula = y ~ 1)

Coefficients:
(Intercept)
       3.45
```

#y is a function of 1
#models only the intercept, without any effect

Reduced Models

- The ANOVA table:

```
> anova(reduced_model)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals  7   1.92 0.27429
```

Full Models

- Next, we fit the **full model**, in this case including the genotype factor

Full Models

- Next, we fit the **full model**, in this case including the genotype factor
  - This causes a reduction in the SSE (**error sum of squares**)
    - It explains part of the variation
- The difference in SSE between the full and reduced models is the **treatment sum of squares**
  - It has *t* degrees of freedom, where *t* is the number of parameters

Linear Models

Full Models

- Fit the full model

```
full_model <- lm(y ~ genotype)
full_model
```

```
Call:
lm(formula = y ~ genotype)

Coefficients:
(Intercept)    genotype1
       3.45        -0.45
```

Linear Models

Full Models

- Finally, build the ANOVA table:

```
> anova(full_model)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value   Pr(>F)
genotype   1   1.62    1.62    32.4 0.001269 **
Residuals  6   0.30    0.05
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> pf(32.4, df1 = 1, df2 = 6, lower.tail = FALSE)
[1] 0.001269296
```

Linear Models

Partitioning Variability

- How can we quantify the proportion of the total variability explained by the model?

Partitioning Variability

- How can we quantify the proportion of the total variability explained by the model?
    - Coefficient of determination, often denoted as R^2 (R-squared)

Linear Models

Coefficient of Determination (R^2)

- R^2 is a measure that ranges from 0 to 1
- Represents the proportion of the total variability in the dependent variable (Y) that is explained by the independent variables (X)

$$R^2 = 1 - \frac{SSE}{SST}$$

SSE (Sum of Squared Errors)
SST (Total Sum of Squares)

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \qquad SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

#Therefore, a higher R^2 implies a better fit of the model to the data

Coefficient of Determination (R^2)

- The total sum of squares can be partitioned into SST = SSR + SSE, that is

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

Linear Models

Coefficient of Determination (R^2)

#use summary( ) function
"Multiple R-squared"

Genotype A		Genotype B	
2.8	3.2	4.1	3.9
3.2	2.8	4.0	3.6

Linear Models

Coefficient of Determination (R^2)

```
y <- c(2.8, 3.2, 3.2, 2.8,4.1, 3.9, 4.0, 3.6)

genotype <-factor(rep(c("A", "B"), each = 4))
genotype

model <- lm(y ~ genotype)
model
summary(model)
```

```
Call:
lm(formula = y ~ genotype)

Residuals:
   Min     1Q Median     3Q    Max
 -0.30  -0.20   0.05   0.20   0.20

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0000     0.1118  26.833 1.77e-07 ***
genotypeB     0.9000     0.1581   5.692  0.00127 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2236 on 6 degrees of freedom
Multiple R-squared:  0.8437,    Adjusted R-squared:  0.8177
F-statistic:  32.4 on 1 and 6 DF,  p-value: 0.001269
```

#Multiple R-squared is a measure of the proportion of the total variability in the dependent variable that is explained by the regression model

#Adjusted R-squared is a modified version of Multiple R-squared that takes into account the number of independent variables in the model

#Penalizes the inclusion of unnecessary or irrelevant independent variables in the model

Let´s Practice 07

- Now we can work through a more realistic example
- We will use a subset of the data from the maize drought and nitrogen stress trials conducted at the CIMMYT breeding program[2]
- The subset includes yield data for (progenies of) 25 genotypes from an F2 population, obtained from a biparental cross between drought tolerant and susceptible maize plants
- Data from four different water stress trials are available, including no stress, intermediate stress and severe stress environments (in two years)

Let´s Practice 07

- Now there are two factors of interest, namely **genotypes** and **environments**
  - The genotype factor has 25 levels and the environment factor has four levels
  - There is only one observation per genotype and environment combination

Linear Models

Let´s Practice

- Now there are two factors of interest, namely **genotypes** and **environments**
  - The genotype factor has 25 levels and the environment factor has four levels
  - There is only one observation per genotype and environment combination

- Let:
  y_{ij} represent the yield of the ith genotype at the jth environment
  g_i represent the effect of the ith genotype
  e_j represent the effect of the jth environment

Linear Models

Let´s Practice

- We want to test hypotheses such as $H_0 = g_1 = g_2 = \cdots = g_{25}$ and $H_0 = e_1 = e_2 = e_3 = e_4$

Linear Models

Let´s Practice

- Use the R function read.csv to import the data
- Fit the following models:

$$y_{ij} = \mu + \varepsilon_{ij}$$
$$y_{ij} = \mu + g_i + \varepsilon_{ij}$$
$$y_{ij} = \mu + e_j + \varepsilon_{ij}$$
$$y_{ij} = \mu + g_i + e_j + \varepsilon_{ij}$$

- Investigate the sums of squares and F statistics
  - Make sure to use the correct reduced model in each case!

Linear Models

Let´s Practice

$$y_{ij} = \mu + \varepsilon_{ij}$$

```
reduced_model <- lm(yield ~ 1, data =data)
anova(reduced_model)
```

```
> reduced_model <- lm(yield ~ 1, data =data)
> anova(reduced_model)
Analysis of Variance Table

Response: yield
           Df    Sum Sq Mean Sq F value Pr(>F)
Residuals 99 326820441 3301217
```

#y is a function of 1
#models only the intercept, without any effect

Linear Models

Let´s Practice

$$y_{ij} = \mu + g_i + \varepsilon_{ij}$$

```
genotype_model <- lm(yield ~ genotype, data = data)
anova(genotype_model)
```

```
> genotype_model <- lm(yield ~ genotype, data = data)
> anova(genotype_model)
Analysis of Variance Table

Response: yield
           Df    Sum Sq Mean Sq F value   Pr(>F)
genotype   24 131813635 5492235  2.1123 0.007587 **
Residuals  75 195006805 2600091
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Linear Models

Let´s Practice

$$y_{ij} = \mu + e_j + \varepsilon_{ij}$$

```
environment_model <- lm(yield ~ environment, data = data)
anova(environment_model)
```

```
> environment_model <- lm(yield ~ environment, data = data)
> anova(environment_model)
Analysis of Variance Table

Response: yield
             Df    Sum Sq Mean Sq F value Pr(>F)
environment   3  10301198 3433733  1.0414 0.3779
Residuals    96 316519243 3297075
```

Linear Models

Let´s Practice

$$y_{ij} = \mu + g_i + e_j + \varepsilon_{ij}$$

```
genotype_environment_model <- lm(yield ~ genotype + environment, data = data)
anova(genotype_environment_model)
```

```
> genotype_environment_model <- lm(yield ~ genotype + environment, data = data)
> anova(genotype_environment_model)
Analysis of Variance Table

Response: yield
             Df    Sum Sq Mean Sq F value   Pr(>F)
genotype     24 131813635 5492235  2.1409 0.007068 **
environment   3  10301198 3433733  1.3385 0.268637
Residuals    72 184705608 2565356
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

References

Chapters 6, 7[3].

1. Box, G., Hunter, J. & Hunter, W. Statistics for Experimenters: Design, Innovation, and Discovery, 2nd Edition. (2005).

2. Malosetti, M., Ribaut, J.-M. & van Eeuwijk, F. A. The statistical analysis of multi-environment data: modeling genotype-by-environment interaction and its genetic basis. *Frontiers in Physiology* 4, 44 (2013).

3. Rencher, A. Linear Models in Statistics. (1999).