UNIVERSITÉ DE NEUCHÂTEL

Master of Cognitive Science

Data Science Course

Linear model I

Professor: Moran Steven

Lecturer: Maiolini Marco

Lecture 8: 13/April/2022

Outline

- Part 1: Book report & discussion (15 minutes)
- Part 2: Bases of linear models (45 minutes)

Break (15 minutes)

- Part 3: Practical

Bivariate relationship

When we have two numerical variables, we can distinguish:

Bivariate relationship

When we have two numerical variables, we can distinguish:

- *Response variable:* dependent variable, as known as Y

Bivariate relationship

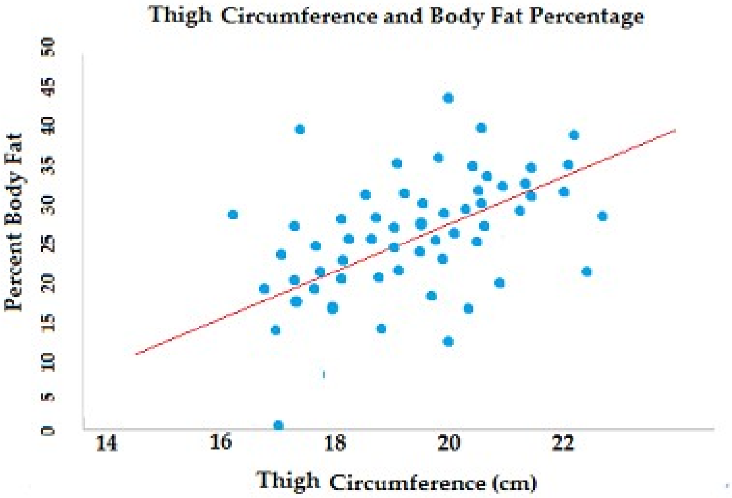
When we have two numerical variables, we can distinguish:

- *Response variable:* dependent variable, as known as Y
- *Explanatory variable:* independent variable, as known as X

Bivariate relationship

When we have two numerical variables, we can distinguish:

- *Response variable:* dependent variable, as known as Y
- *Explanatory variable:* independent variable, as known as X



Relationship between X and Y

Techniques based on fitting a straight line to the data:

Relationship between X and Y

Techniques based on fitting a straight line to the data:

Linear regression

Correlation analysis

Linear regression

Example

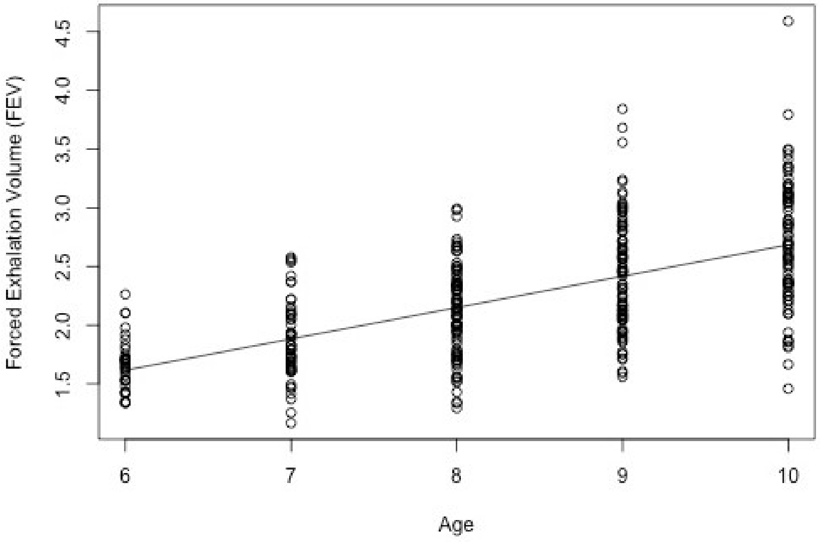
You want to test Lung Function in children. You measure the forced exhalation volume (FEV), the measure of how much air somebody can forcibly exhale from their lungs, from 6 to 10 year old children. You survey 345 children.

The data here are a part of dataset given in Kahn, Micheal (2005), “An Exhalent Problem for Teaching Statistics”, The Journal of Statistical Education, 13(2).

Linear regression

Example

You want to test Lung Function in children. You measure the forced exhalation volume (FEV), the measure of how much air somebody can forcibly exhale from their lungs, from 6 to 10 year old children. You survey 345 children.

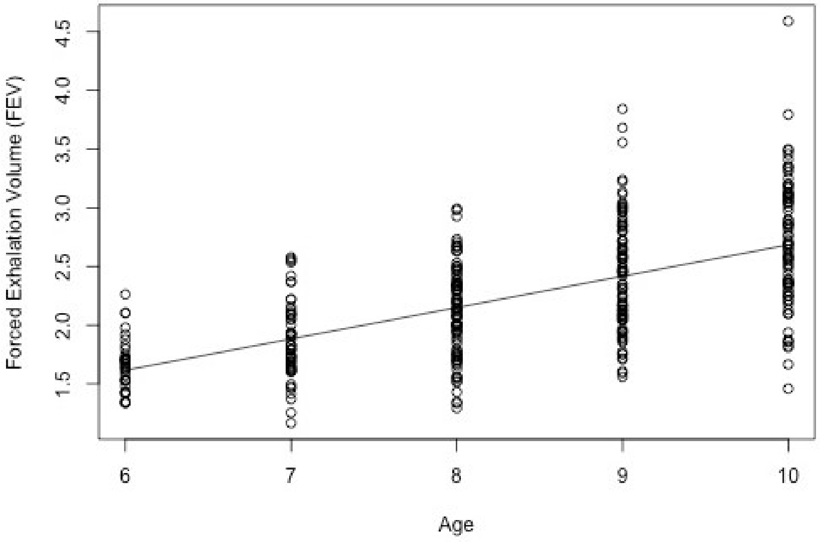


The data here are a part of dataset given in Kahn, Micheal (2005), “An Exhalent Problem for Teaching Statistics”, The Journal of Statistical Education, 13(2).

Linear regression

Example

You want to test Lung Function in children. You measure the forced exhalation volume (FEV), the measure of how much air somebody can forcibly exhale from their lungs, from 6 to 10 year old children. You survey 345 children.



The scatter plot suggest a definite age-relationship, with larger X tending to be associated with bigger values of Y

The data here are a part of dataset given in Kahn, Micheal (2005), “An Exhalent Problem for Teaching Statistics”, The Journal of Statistical Education, 13(2).

Correlation analysis

Example

You investigate whether standardized scores from high school (SAT) are related to academic grades in college (GPA). You predict that there’s a positive correlation: higher SAT scores are associated with higher college GPAs while lower SAT scores are associated with lower college GPAs.

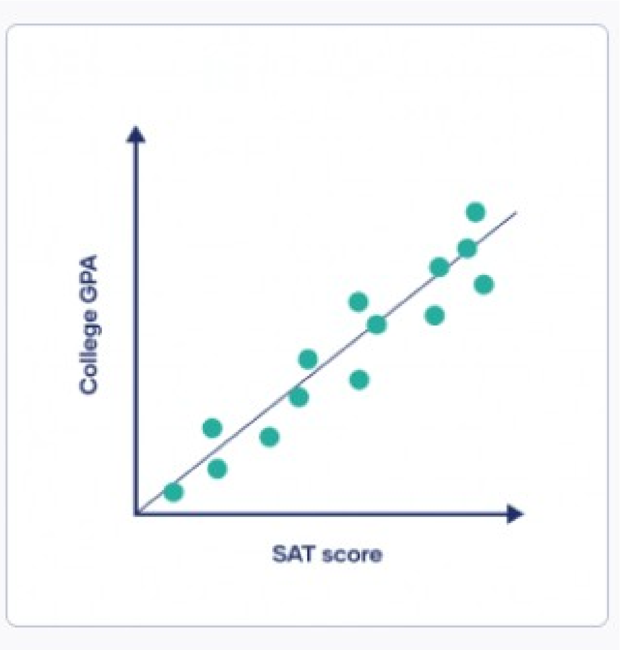
You gather a sample of 5000 college graduates and survey them on their high school SAT scores and college GPAs.

Correlation analysis

Example

You investigate whether standardized scores from high school (SAT) are related to academic grades in college (GPA). You predict that there's a positive correlation: higher SAT scores are associated with higher college GPAs while lower SAT scores are associated with lower college GPAs.

You gather a sample of 5000 college graduates and survey them on their high school SAT scores and college GPAs.

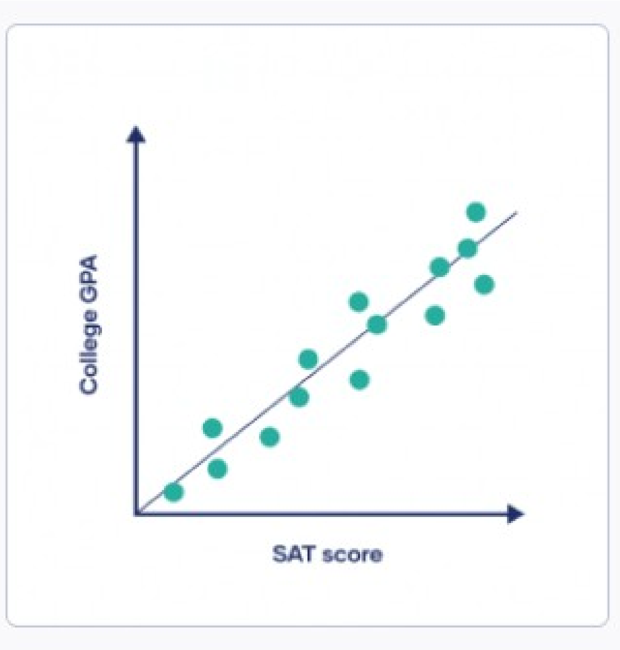


Correlation analysis

Example

You investigate whether standardized scores from high school (SAT) are related to academic grades in college (GPA). You predict that there's a positive correlation: higher SAT scores are associated with higher college GPAs while lower SAT scores are associated with lower college GPAs.

You gather a sample of 5000 college graduates and survey them on their high school SAT scores and college GPAs.



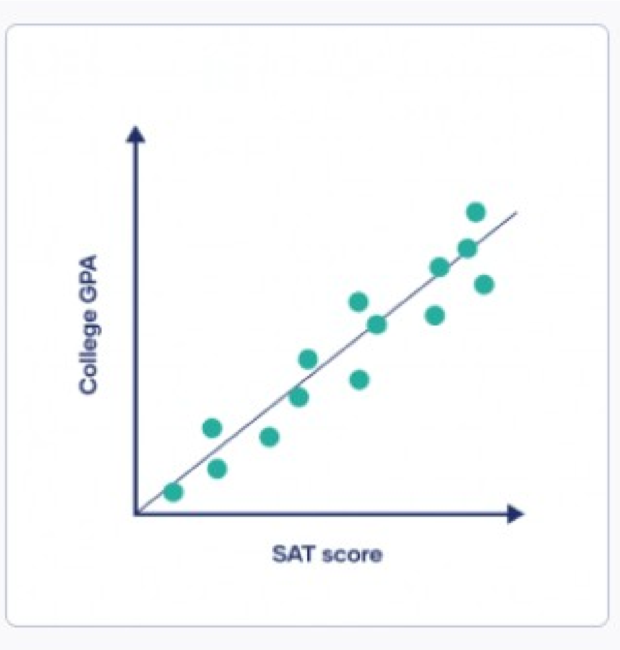
Correlation coefficient = 0.58

Correlation analysis

Example

You investigate whether standardized scores from high school (SAT) are related to academic grades in college (GPA). You predict that there's a positive correlation: higher SAT scores are associated with higher college GPAs while lower SAT scores are associated with lower college GPAs.

You gather a sample of 5000 college graduates and survey them on their high school SAT scores and college GPAs.



Correlation coefficient = 0.58

The scatter plot seems to confirm our prediction, with higher SAT scores associated with higher GPA values.

The Correlation Coefficient

The Correlation Coefficient measure the strength of linear association between the two variables.

The Correlation Coefficient

The Correlation Coefficient measure the strength of linear association between the two variables.

$$\text{The Correlation Coefficient} \left\{ \begin{array}{l} +1 \\ \\ \\ -1 \end{array} \right.$$

The Correlation Coefficient

The Correlation Coefficient measure the strength of linear association between the two variables.

$$\text{The Correlation Coefficient} \begin{cases} +1 \longrightarrow \text{Strong Positive Relationship} \\ \\ -1 \end{cases}$$

The Correlation Coefficient

The Correlation Coefficient measure the strength of linear association between the two variables.

The Correlation Coefficient

+1 → Strong Positive Relationship

-1 → Strong Negative Relationship

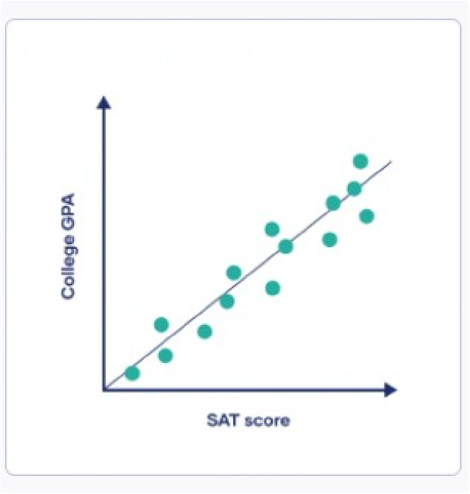
The Correlation Coefficient

The Correlation Coefficient measure the strength of linear association between the two variables.

The Correlation Coefficient

+1 → Strong Positive Relationship

-1 → Strong Negative Relationship



The Pearson's r correlation test:

- Variables are quantitative
- Variables normally distributed

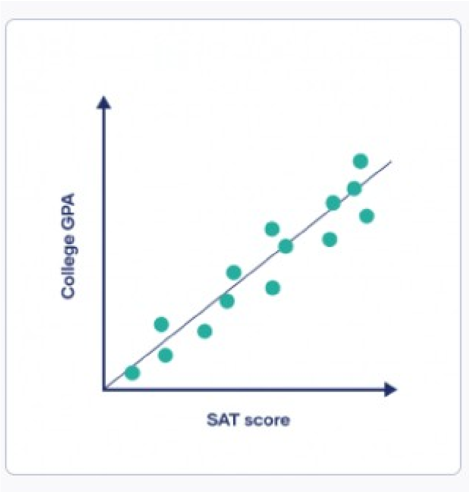
The Correlation Coefficient

The Correlation Coefficient measure the strength of linear association between the two variables.

The Correlation Coefficient

+1 → Strong Positive Relationship

-1 → Strong Negative Relationship



The Pearson’s r correlation test:

- Variables are quantitative
- Variables normally distributed

In R you see if two variable are correlated:

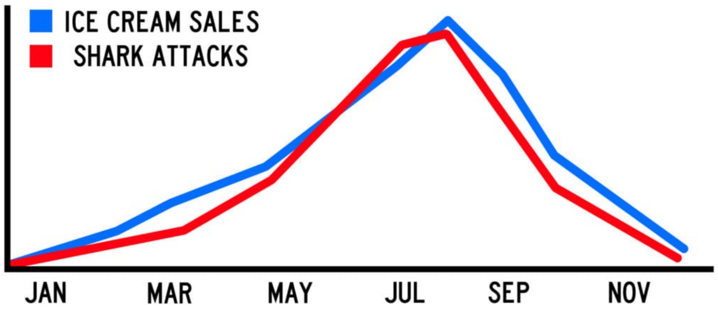
cor(x, y)

The Correlation Coefficient

A strong correlation between two variables does not indicate any causal connection between them. It is important to remember this concept when interpreting correlation.

The Correlation Coefficient

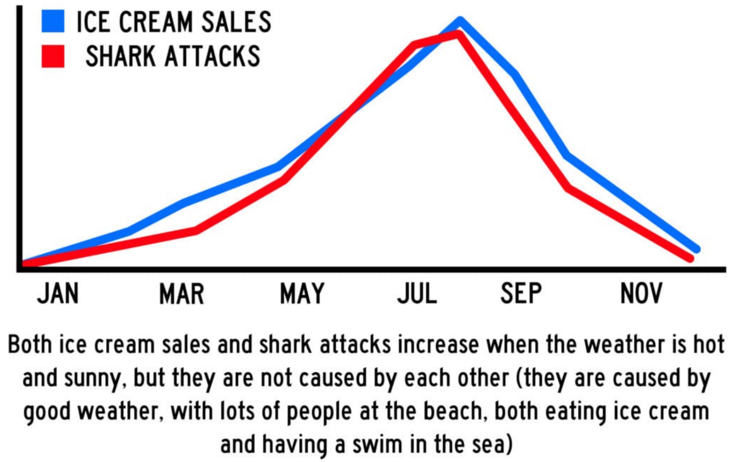
A strong correlation between two variables does not indicate any causal connection between them. It is important to remember this concept when interpreting correlation.

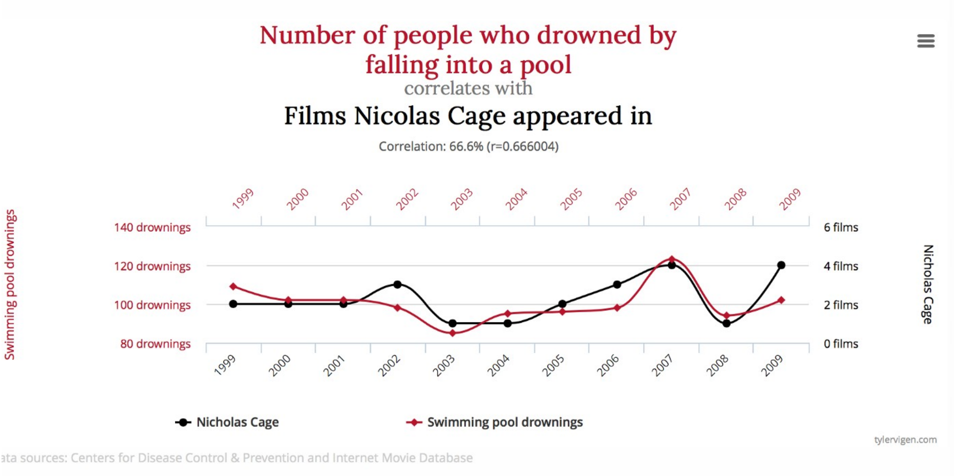


Both ice cream sales and shark attacks increase when the weather is hot and sunny, but they are not caused by each other (they are caused by good weather, with lots of people at the beach, both eating ice cream and having a swim in the sea)

The Correlation Coefficient

A strong correlation between two variables does not indicate any causal connection between them. It is important to remember this concept when interpreting correlation.





The Correlation Coefficient

A strong correlation between two variables does not indicate any causal connection between them. It is important to remember this concept when interpreting correlation.

The Correlation Coefficient

A strong correlation between two variables does not indicate any causal connection between them. It is important to remember this concept when interpreting correlation.

The cat didn’t crush the awning

The Regression Line

In perfect linear relationships the line that fits exactly the data have slope Sy/Sx and passes through the point (x,y) or SD line.

The Regression Line

In perfect linear relationships the line that fits exactly the data have slope Sy/Sx and passes through the point (x,y) or SD line. When there is not linear relationship:

$$\mathbf{Y = b_0 + b_1 x}$$

The Regression Line

In perfect linear relationships the line that fits exactly the data have slope Sy/Sx and passes through the point (x,y) or SD line. When there is not linear relationship:

Slope

$$\mathbf{Y = b_0 + b_1 x}$$

Intercept

The Regression Line

In perfect linear relationships the line that fits exactly the data have slope Sy/Sx and passes through the point (x,y) or SD line. When there is not linear relationship:

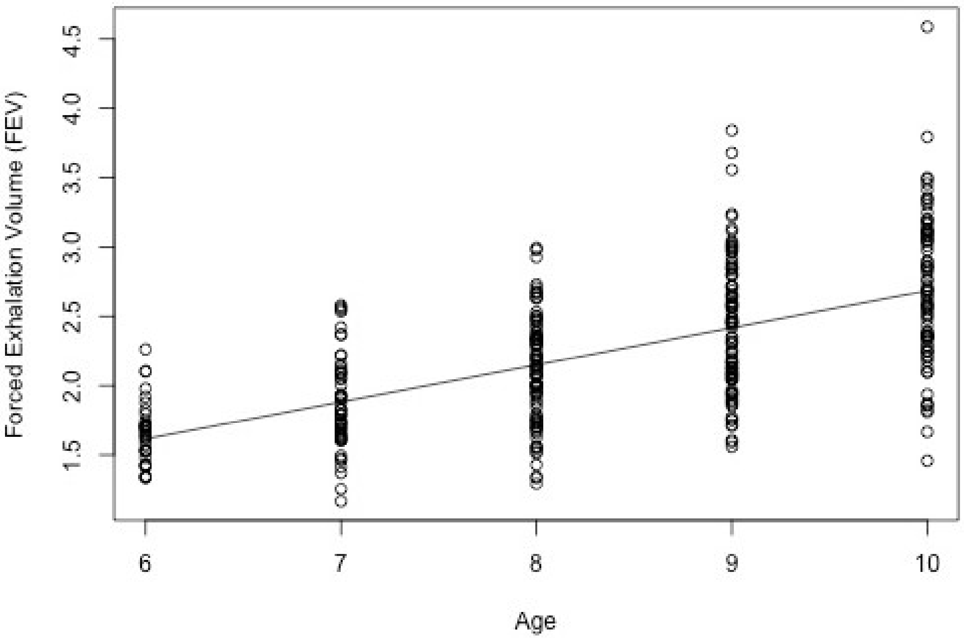
Slope

$$Y = b_0 + b_1x$$

Intercept

Intercept = 0.01165

Slope = 0.26721



The Regression Line

In perfect linear relationships the line that fits exactly the data have slope Sy/Sx and passes through the point (x,y) or SD line. When there is not linear relationship:

Slope

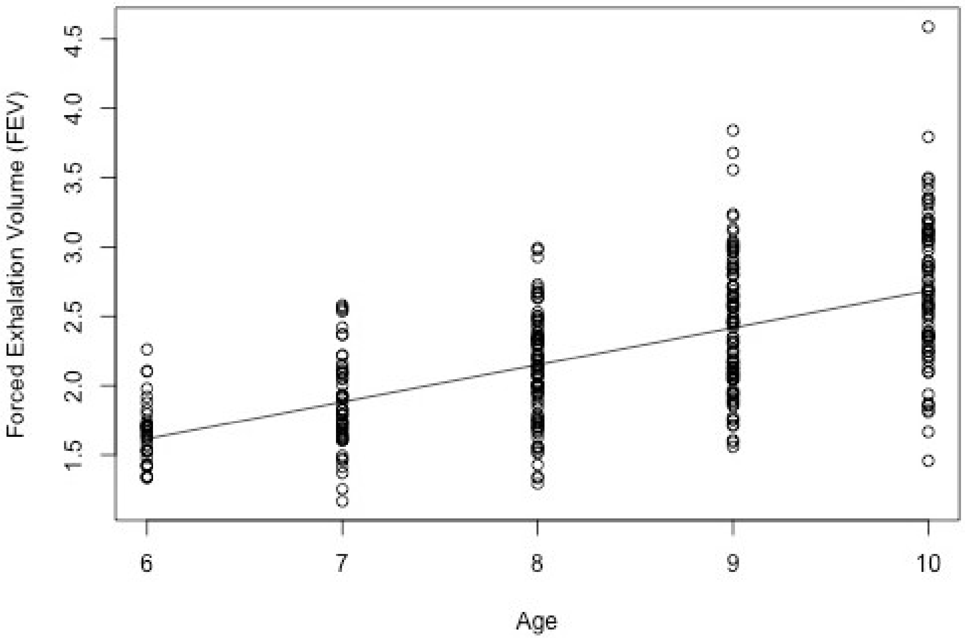
$$\mathbf{Y = b_0 + b_1x}$$

Intercept

Intercept = 0.01165

Slope = 0.26721

FEV = 0.01165 + 0.26721*Age



The Regression Line

In perfect linear relationships the line that fits exactly the data have slope Sy/Sx and passes through the point (x,y) or SD line. When there is not linear relationship:

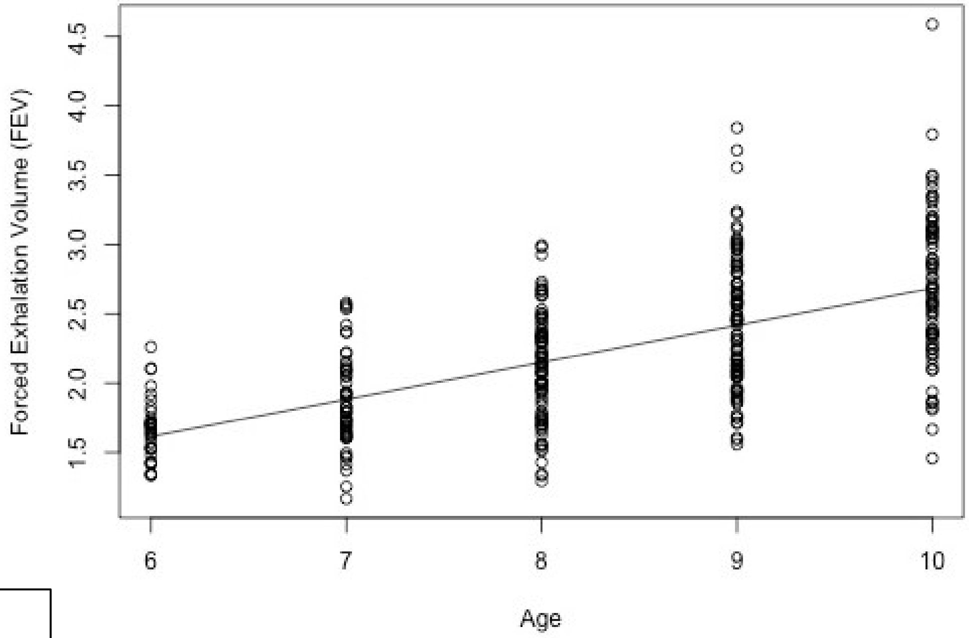
Slope

$$Y = b_0 + b_1x$$

Intercept

Intercept = 0.01165 Slope = 0.26721

FEV = 0.01165 + 0.26721*Age



In R you can estimate slope & intercept:

lm(formula = Response ~ Explanatory, data = dataset)

(General) Linear Models

The General Linear Models are used to predict one <u>Response variable</u> from one or more <u>Explanatory variables</u>

Image taken from “Mixed Effects Models and Extensions in Ecology with R” Alain F. Zuur, Elena N. Ieno, Neil Walker, Anatoly A. Saveliev, Graham M. Smith - (2009, Springer-Verlag New York)

(General) Linear Models

The General Linear Models are used to predict one <u>Response variable</u> from one or more <u>Explanatory variables</u>

- Simple Regression $\longrightarrow$ $\mathbf{Y = b_0 + b_1 x}$

Image taken from “Mixed Effects Models and Extensions in Ecology with R” Alain F. Zuur, Elena N. Ieno, Neil Walker, Anatoly A. Saveliev, Graham M. Smith - (2009, Springer-Verlag New York)

(General) Linear Models

The General Linear Models are used to predict one Response variable from one or more Explanatory variables

- Simple Regression → $Y = b_0 + b_1x$
- Multiple Regression → $Y = b_0 + b_1x_1 + b_2x_2 + b_3x_3...$

Image taken from "Mixed Effects Models and Extensions in Ecology with R" Alain F. Zuur, Elena N. Ieno, Neil Walker, Anatoly A. Saveliev, Graham M. Smith - (2009, Springer-Verlag New York)

(General) Linear Models

The General Linear Models are used to predict one <u>Response variable</u> from one or more <u>Explanatory variables</u>

- Simple Regression ⟶ $\mathbf{Y = b_0 + b_1x}$
- Multiple Regression ⟶ $\mathbf{Y = b_0 + b_1x_1 + b_2x_2 + b_3x_3...}$

<u>Assumption</u>

- Linearity
- Normality of residuals
- Homoscedasticity
(Homogeneity of variance)

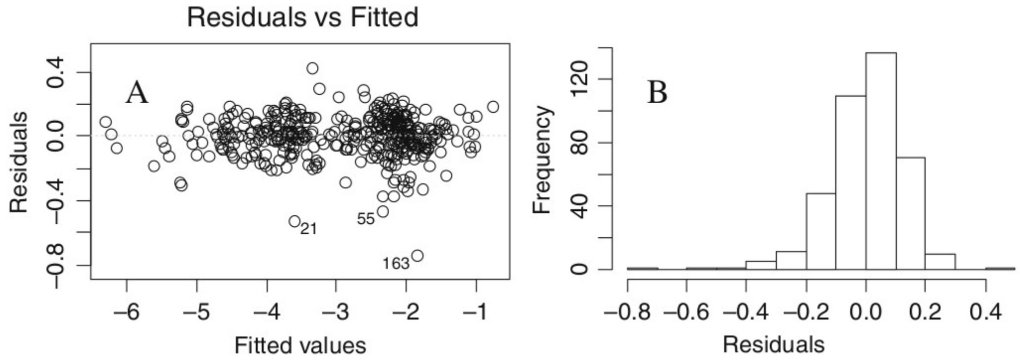


Image taken from “Mixed Effects Models and Extensions in Ecology with R” Alain F. Zuur, Elena N. Ieno, Neil Walker, Anatoly A. Saveliev, Graham M. Smith - (2009, Springer-Verlag New York)

(General) Linear Models

The General Linear Models are used to predict one Response variable from one or more Explanatory variables

Simple regression

(General) Linear Models

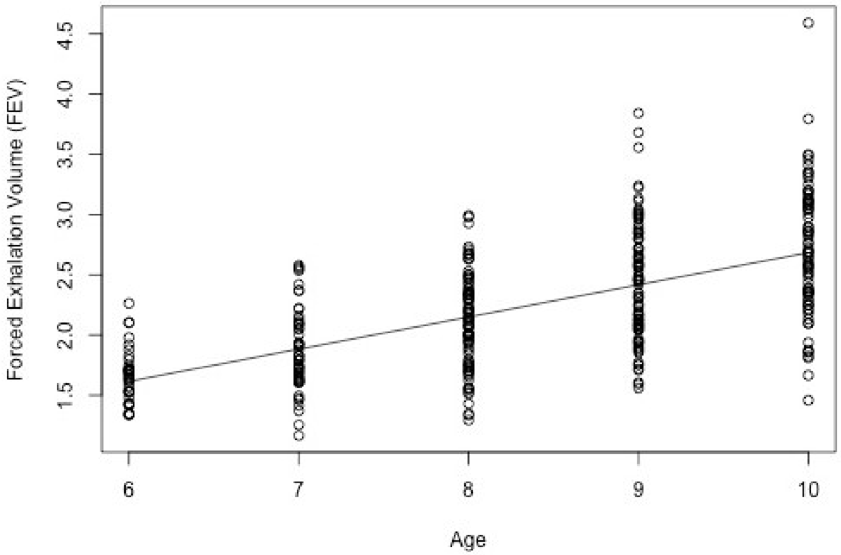
The General Linear Models are used to predict one Response variable from one or more Explanatory variables

Simple regression → One explanatory variables $Y = b_1X + b_0$

(General) Linear Models

The General Linear Models are used to predict one Response variable from one or more Explanatory variables

Simple regression → One explanatory variables $Y = b_1X + b_0$



(General) Linear Models

The General Linear Models are used to predict one <u>Response variable</u> from one or more <u>Explanatory variables</u>

Simple regression → One explanatory variables $Y = b_1X + b_0$

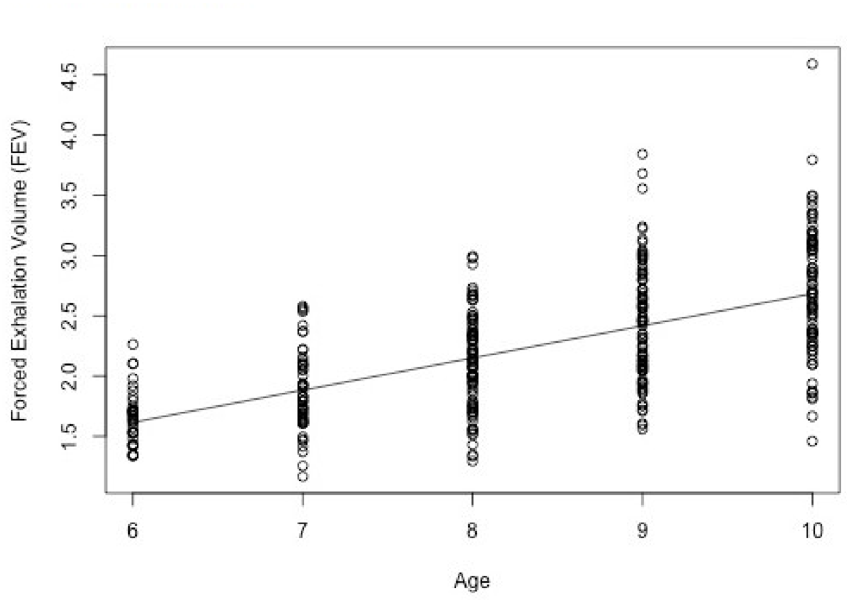
Linear Regression

X

Y

X explain Y

X ~ Y



(General) Linear Models

The General Linear Models are used to predict one <u>Response variable</u> from one or more <u>Explanatory variables</u>

Multiple regression

(General) Linear Models

The General Linear Models are used to predict one Response variable from one or more Explanatory variables

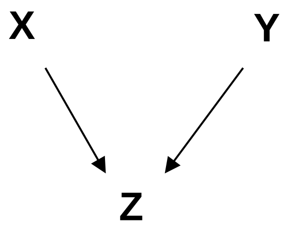
Multiple regression → Have multiple explanatory variables $Y = b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_0$

(General) Linear Models

The General Linear Models are used to predict one <u>Response variable</u> from one or more <u>Explanatory variables</u>

Multiple regression $\longrightarrow$ Have multiple explanatory variables $Y = b_1X_1 + b_2X_2 + b_3X_3 + \ldots. + b_0$

Additive independent effects



X and Y explain the variation in Z independently

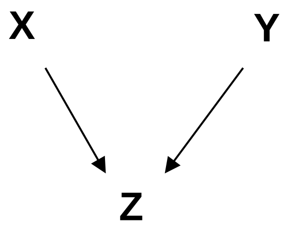
Z ~ X + Y

(General) Linear Models

The General Linear Models are used to predict one <u>Response variable</u> from one or more <u>Explanatory variables</u>

Multiple regression $\longrightarrow$ Have multiple explanatory variables $Y = b_1X_1 + b_2X_2 + b_3X_3 + \ldots. + b_0$

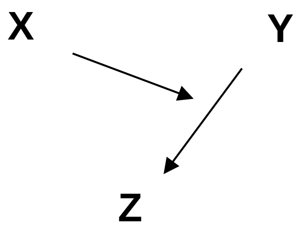
Additive independent effects



X and Y explain the variation in Z independently

Z ~ X + Y

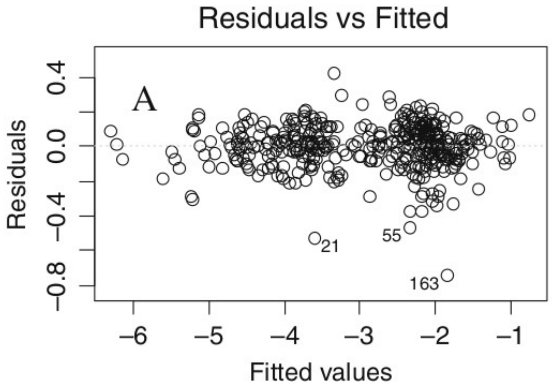
Interaction among variable



X modifies how Y affects Z

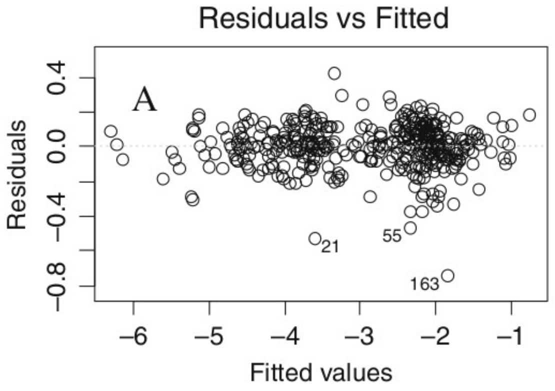
Z ~ X + Y + X*Y

Residuals

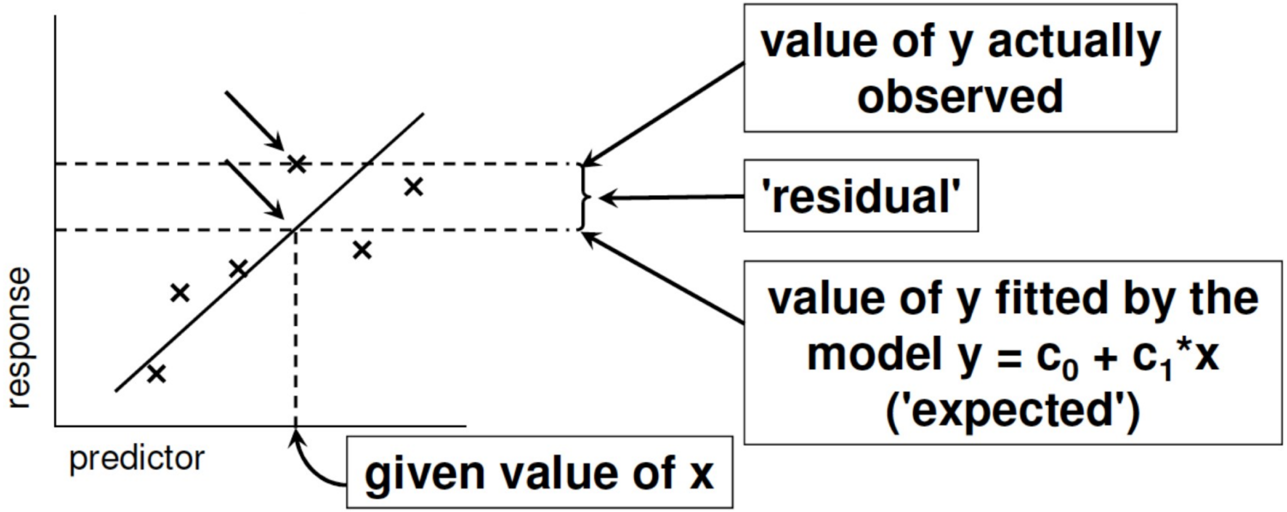


- *Residuals:* Difference between observation and fitted values
- *Fitted values: Estimation of an observation using all previous ones*

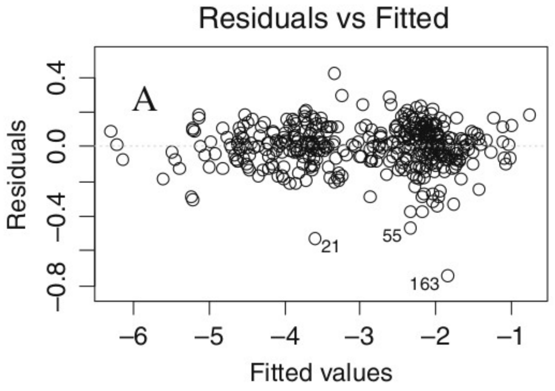
Residuals



- *Residuals:* Difference between observation and fitted values
- *Fitted values: Estimation of an observation using all previous ones*



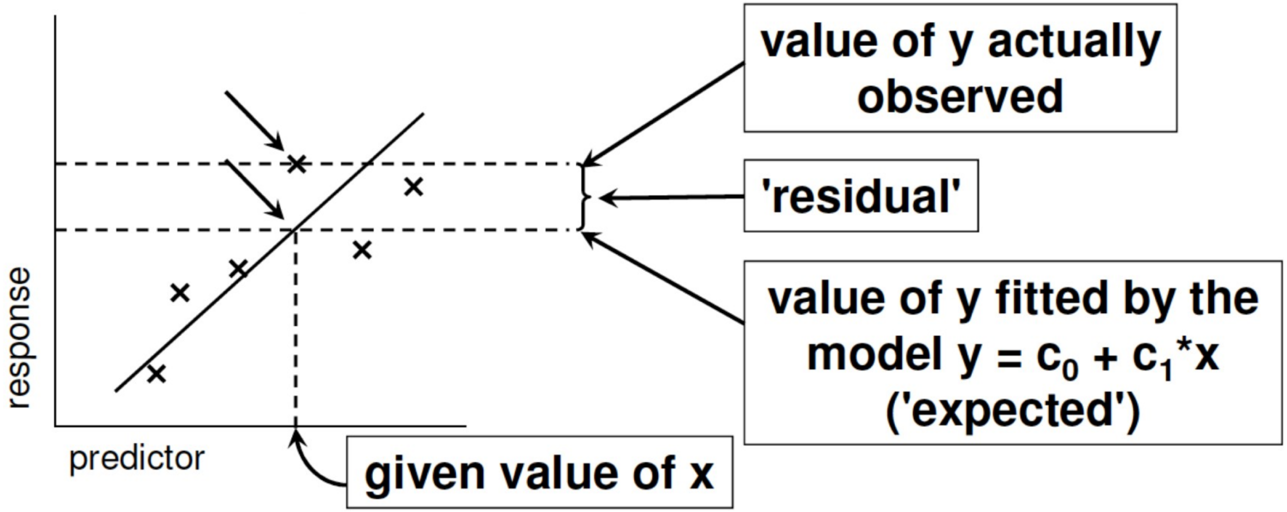
Residuals



- *Residuals:* Difference between observation and fitted values
- *Fitted values: Estimation of an observation using all previous ones*

Error (so called Residual)

$Y = b_1X + b_0 + e$



Generalized Linear Models

If we don't have the normality of residuals, we can use the Generalized Linear Models (GLM).

Generalized Linear Models

If we don’t have the normality of residuals, we can use the Generalized Linear Models (GLM).

- Can be used with residuals with distribution normal, binomial, poisson...

- Have the same features of General Linear Models

Generalized Linear Models

If we don't have the normality of residuals, we can use the Generalized Linear Models (GLM).

- Can be used with residuals with distribution normal, binomial, poisson...
- Have the same features of General Linear Models

In R you can fit your data in a General Linear Model:

*lm(formula = Response ~ Explanatory + Z + Z*Y, data = dataset)*

In R you can fit your data in a Generalized Linear Model (GLM):

*glm(formula = Response ~ Explanatory + Z + Z*Y, family = binomial, data = dataset)*

Summary

Model	Variables	Distribution	R code
Linear Regression	$\mathbf{Y = b_0 + b_1x}$	Normal	*lm(formula, data)*
General Linear Models	$\mathbf{Y = b_0 + b_1x_1 + b_2x_2 + \ldots}$	Normal	*lm(formula, data)*
Generalized Linear Models (GLM)	$\mathbf{Y = b_0 + b_1x_1 + b_2x_2 + \ldots}$	Any	*glm(formula, family, data)*

datacamp

R