FACULTY OF HEALTH SCIENCES - SCHOOL OF MEDICINE
MSc Health Statistics and Data Analytics

ARISTOTLE UNIVERSITY OF THESSALONIKI

Linear Regression Models

KONSTANTINOS I. BOUGIOUKAS, PhD

PHYSICIST, BIOSTATISTICIAN AND RESEARCH METHODOLOGIST

THESSALONIKI 2022

MSc Health Statistics & Data Analytics

Statistical Models

What do we mean by a statistical model?

- A simplification or approximation of reality (Burnham, Anderson, 2002)
- Statistical models summarize patterns of the data available for analysis (Steyerberg, 2009)
- A powerful tool for developing and testing theories by way of causal explanation, prediction, and description (Shmueli, 2010)

Statistical models

Statistical models are simple mathematical rules derived from empirical data describing **the association between an outcome and several explanatory variables** (Dunkler et al, 2014)

- They should be **valid**: provide predictions with acceptable accuracy
- They should be **practically useful**: allow conclusions such as 'how large is the expected change in outcome if one of the explanatory variables changes by one unit'
- They should be **robust**.

Modeling for explanation

Describe and quantify the association between the outcome variable *Y* and a set of explanatory variables *X*’s.

- Identification of ‘important’ explanatory variables
- Understanding the effects of explanatory variables
- Adjustment for variables uncontrollable by experimental design

Modeling for prediction

When we want to predict an outcome variable *Y* based on the information contained in a set of predictor variables *X*’s.

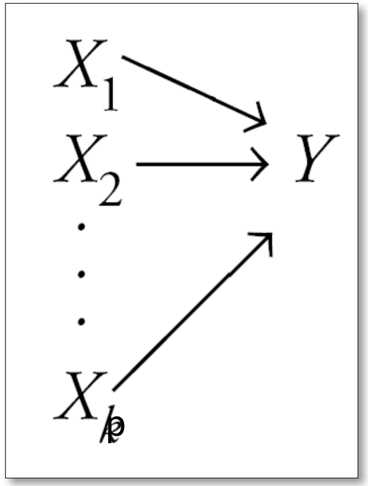
(Shmueli,2010)

Learning objectives

- Fitting a **single linear model** with a continuous, binary or categorical variable
- **Interpret the results** from examples of simple and multiple linear regression models
- Explain different **strategies** for picking a “final” multiple linear regression model

Linear Regression model

The effect of one or more (**continuous or categorical**) independent variables Xp on the values of a **continuous** dependent variable Y.



Example:

We would like to examine whether several variables (e.g., height, headc, gender, parity, education) have an effect on weight (in g) of infants at 1-month age.

Dataset

The data of 550 infants at 1 month age were collected (**BirthWeight**). The following variables were recorded:

- Body weight of the infant in kg (**weight**) (Note: we will transform this to g)
- Body height of the infant in cm (**height**),
- Head circumference in cm (**headc**),
- Gender of the infant (**gender**: Female, Male)
- Birth order in their family (**parity**: Singleton, One sibling, 2 or more siblings)
- Education of the mother (**education**: year10, year12, tertiary)

Simple Linear Regression

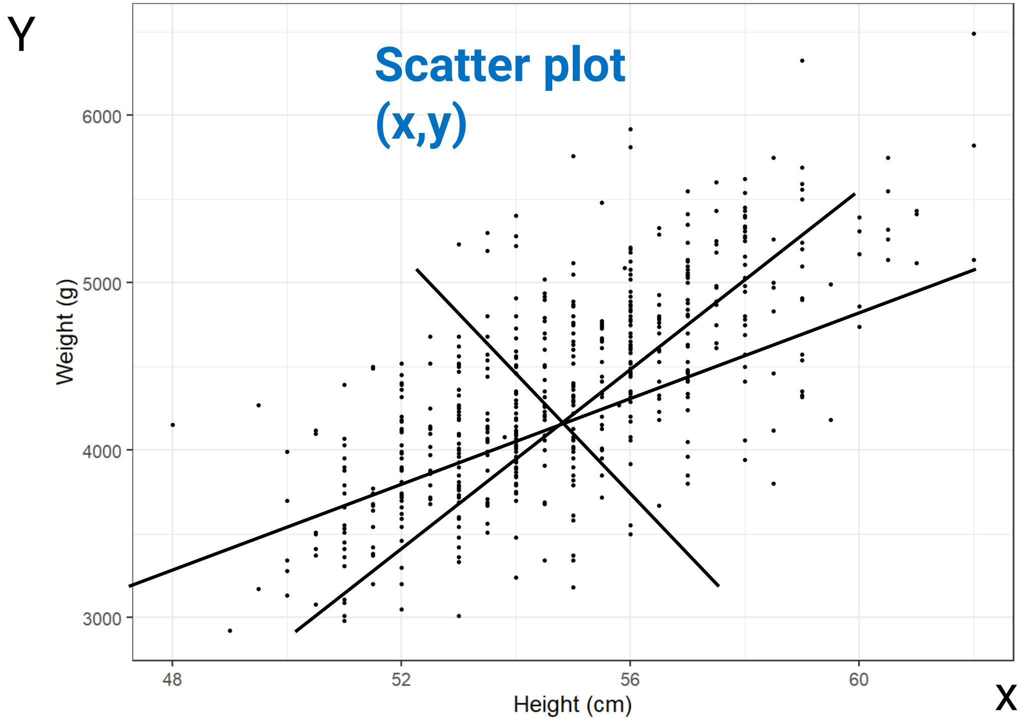
Simple Linear regression

Dependent variable

$$y_i = \boxed{\beta_o + \beta_1 * x_i} + \varepsilon_i \qquad i=1,2,\ldots,n$$

Systematic component

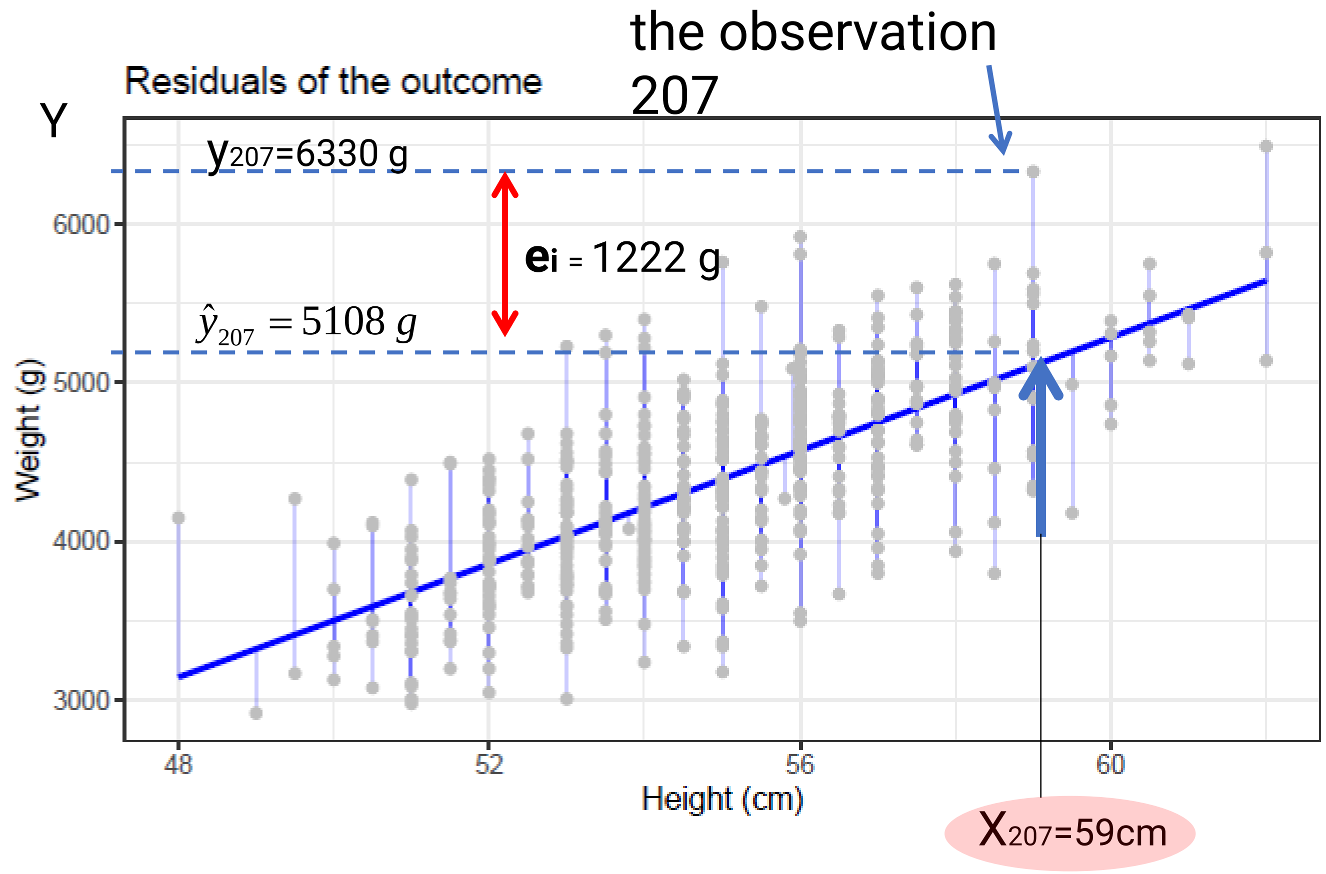
Random error



X: height
(independent or explanatory variable)

Y: weight
(response or dependent variable)

Line of best fit (direct regression)



Residuals (error)

$$\hat{e}_i = y_i - \hat{y}_i$$?

$$\sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_o - \hat{\beta}_1 * X_1)^2$$

least squares estimates

$$\hat{\beta}_o, \hat{\beta}_1$$

Best fitted line

$$\hat{y}_i = \hat{\beta}_o + \hat{\beta}_1 * x_i$$

Question:

What is the association between weight and height ?

Hypothesis Testing

$$\hat{weight} = \hat{\beta}_0 + \hat{\beta}_1 * height$$

- **Ho: β_1=0** (no association)

- **H1: β_1≠0** (there is association)

Results and interpretation

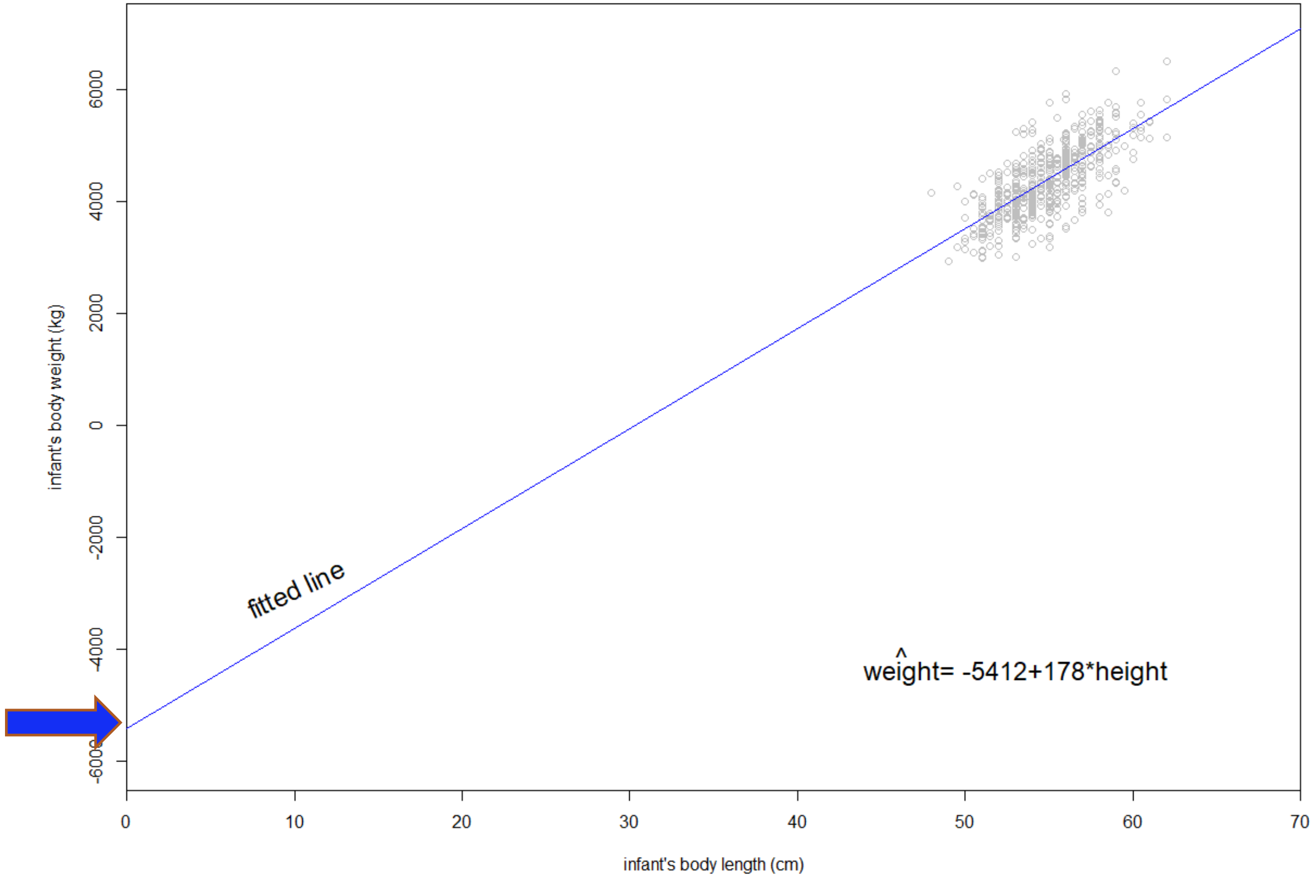
$$\hat{weight} = -5412 + 178 * height$$

On average, there's an expected increase of 178 g of weight for every 1 cm increase in height (95%CI: 164 to 193, P<0.001)

The intercept

$$\hat{weight} = -5412 + 178 * height$$

Plot the fitted line crossing the y-axis (weight):



The fitted line crosses the y-axis roughly at -5400. This value is the estimate of the intercept β_0. Not physical interpretation.

The slope

The slope β_1 from two points of the fitted line is:

$$\hat{\beta}_1 = \frac{\Delta y}{\Delta x} = \frac{5270 - 4560}{60 - 56} = \frac{710}{4} \approx 178 \text{ (g/cm)}$$

Question:

What is the association between weight and gender of the infant?

$$gender = \begin{cases} 1, & Male \\ 0, & Female\ (ref.) \end{cases}$$

$$\hat{y} = \widehat{\text{weight}} = b_0 + b_1 \cdot \ genderMale$$

Results and interpretation

$$genderMale = \begin{cases} 1 & \text{if infant is Male} \\ 0 & \text{otherwise (ref.)} \end{cases}$$

$$\hat{weight} = 4140.5 + 451.8 * genderMale$$

- For **females**:

  Weight = **4140.5 + 451.8*0 = 4140.5 g**

  The intercept is the mean body weight (in g) for a female infant which is the **reference category**.

- For **males**:

  Weight = **4140.5 + 451.8*1 = 4140.5 + 451.8 = 4592.3 g**

  The coefficient value 451.8 is **the difference** (4592.3 – 4140.5) in the **mean** weight (in g) for a male infant **relative** to a female infant.

Conclusion

The mean weight of a male infant is 4592 g which is **significantly higher about 452 g** relative to a female infant 4141 kg (95%CI: 358 to 545, $p<0.001$)

The above analysis is equivalent to perform a **two-sample t-test**!

Categorical explanatory variable (>2 categories)

Question:

What is the association between weight and birth order in the family (parity) of the infant?

$$parity = \begin{cases} Singleton\ (ref.) \\ One\ sibling \\ 2\ or\ more\ siblings \end{cases}$$

A categorical explanatory variable with k-levels or categories requires (k-1) ***dummy variables*** to represent it.

The explanatory variable here has three categories, so we need to create **two dummy variables**.

Considering the **Singleton** as the reference group:

$$parityOne\ sibling = \begin{cases} 1 & \text{if infant has one sibling} \\ 0 & \text{otherwise} \end{cases}$$

$$parity\ \geqslant 2\text{siblings} = \begin{cases} 1 & \text{if infant has 2 or more siblings} \\ 0 & \text{otherwise} \end{cases}$$

parity	One sibling	2 or more siblings
Singleton (ref.)	0	0
One sibling	1	0
2 or more siblings	0	1

We are including all the categories to the linear regression model **except one which is going to be used as the reference group** (here the Singleton category).

Results and interpretation

$$\hat{y} = \widehat{\text{weight}} = b_0 + b_1 \cdot \; parityOne\ sibling + b_2 \cdot \; parity \geqslant 2 \text{ siblings}$$

$$\hat{weight} = 4259 + 130 * parityOneSibling + 192 * parity \geq 2siblings$$

- For a singleton infant:

  Weight = 4259 + 130*0 + 192*0 = 4259 g

The **intercept** equals to the mean weight in g for a singleton infant **which is the reference category**.

- For an infant with **one sibling:**

  Weight = 4259 + 130*1 + 192*0 = 4259 + 130 = 4389 g

The coefficient for “One sibling” dummy variable is **130** and represents the difference in the mean weight in grams for an infant with **one sibling relative to a singleton infant.**

Results and interpretation

- For an infant with **2 or more siblings:**

  Weight = 4259 + 130*0 + 192*1 = 4259 + 192= 4451 g

The coefficient for “2 or more siblings” dummy variable is **192** and represents the difference in the mean weight in grams for an infant with **2 or more** siblings **relative to a singleton infant**.

Headc and Education variables

Headc

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	-6059.866	560.364	-10.814	0	-7160.591	-4959.142
headc	275.134	14.778	18.618	0	246.106	304.162

Education

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	4352.929	42.821	101.653	0.000	4268.815	4437.044
educationyear12	57.980	74.169	0.782	0.435	-87.711	203.671
educationtertiary	6.636	57.173	0.116	0.908	-105.669	118.941

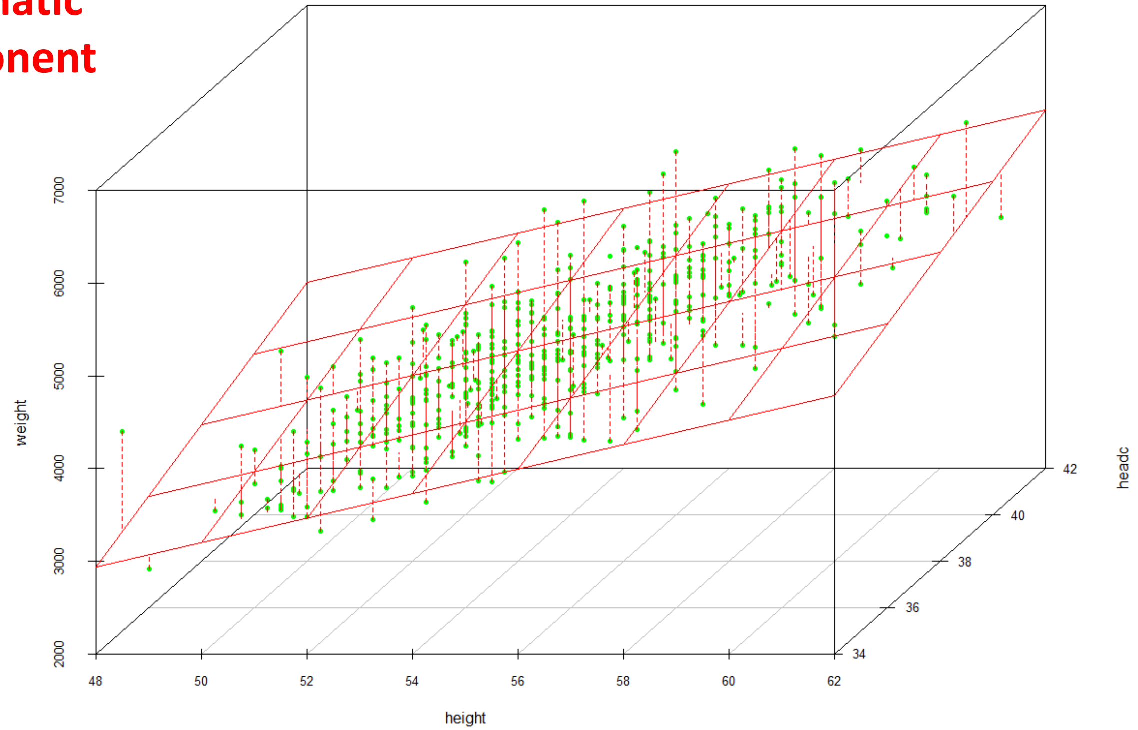
Multiple Linear Regression

Multiple Linear regression

$$y_i = \boxed{\beta_o + \beta_1 * x_{1i} + \beta_2 * x_{2i} + \beta_3 * x_{3i} + \ldots + \beta_p * x_{pi}} + \varepsilon_i$$

Systematic component

Random error



Two explanatory variables: a *plane* in three-dimensional space

Criteria for selecting variables

- **Significance criteria** (e.g., forward selection, backward elimination, p-values)

- **Information criteria** (e.g., AIC, BIC)   *the **smaller** the value of AIC the better the model*

- **Background knowledge** (directed acyclic graph “DAG”)

Model building: A possible strategy

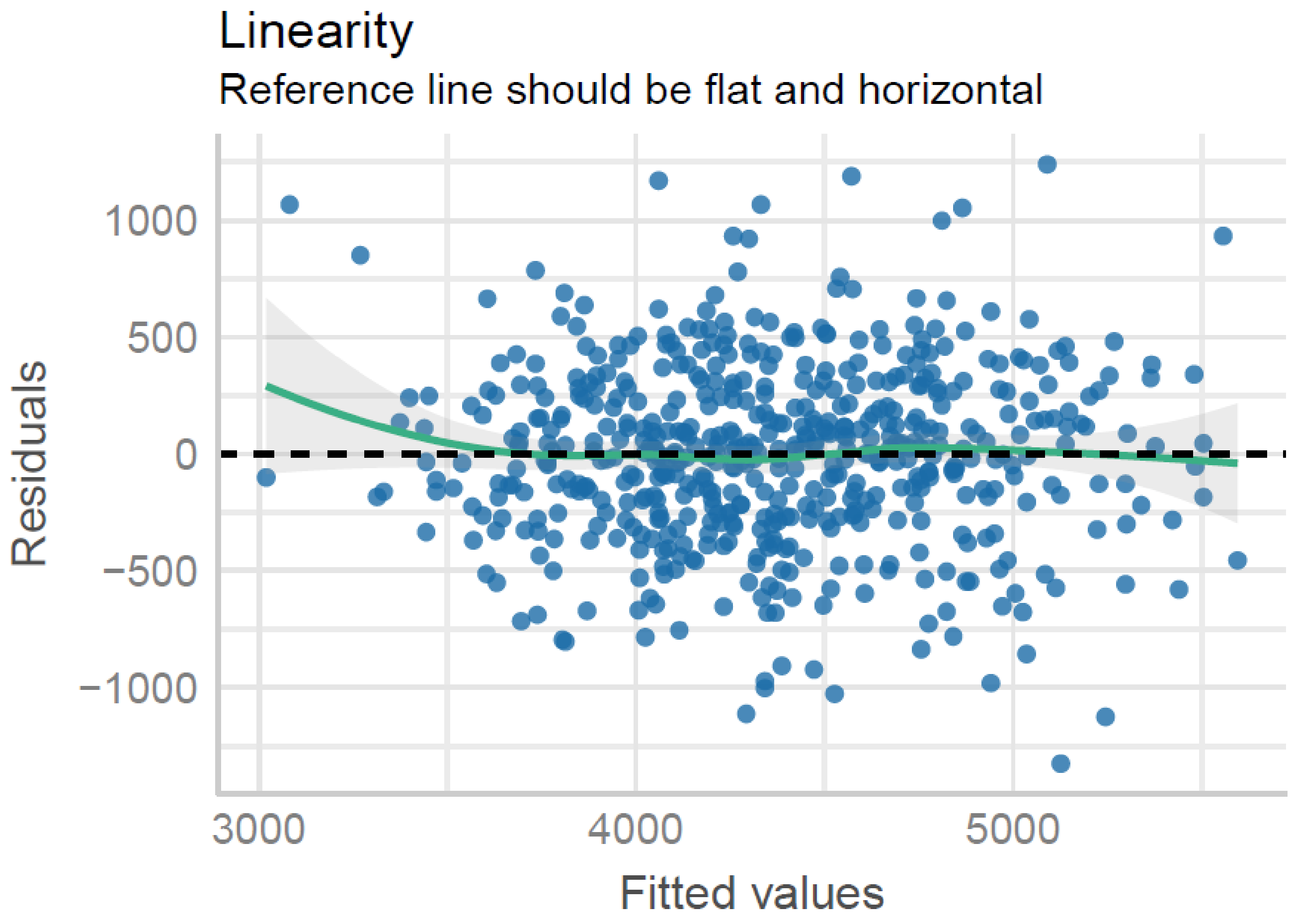
– sample size calculation

– the variables that have a **p-value <0.2** in the univariable analysis are candidate variables for the model

– **existing knowledge** (e.g., confounders) should be used

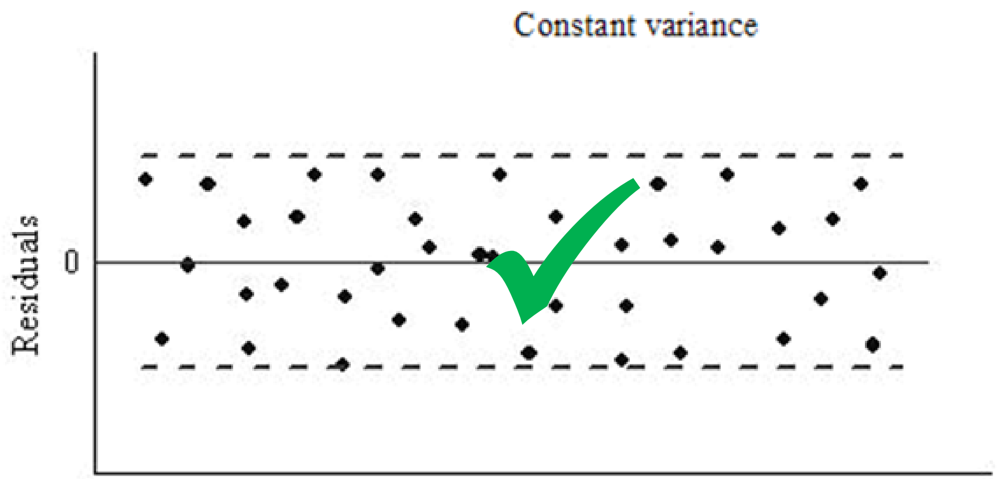
– models should be **interpretable**

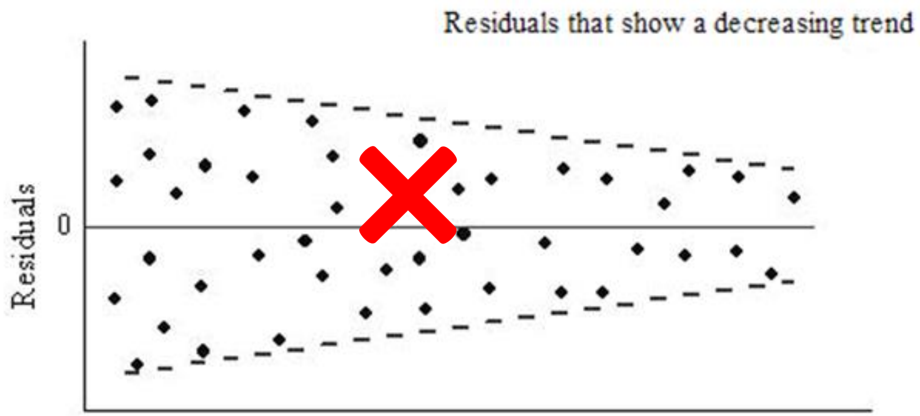
Checking assumptions: Linearity

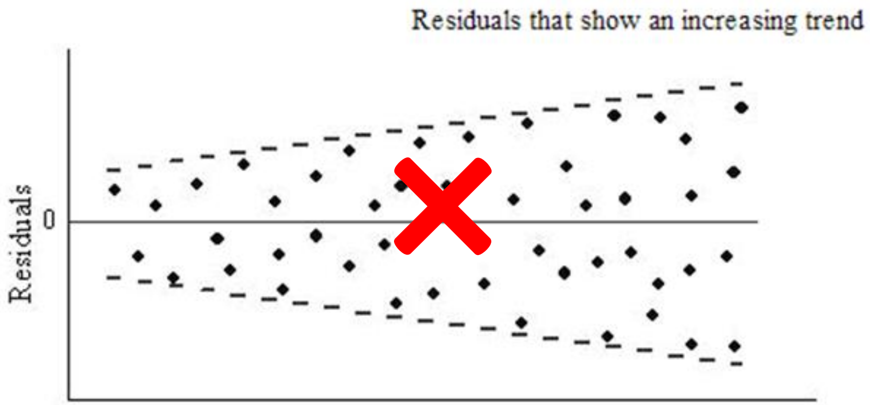


If you find equally spread residuals around a **horizontal line** without distinct **patterns**, that is a good indication for linear association.

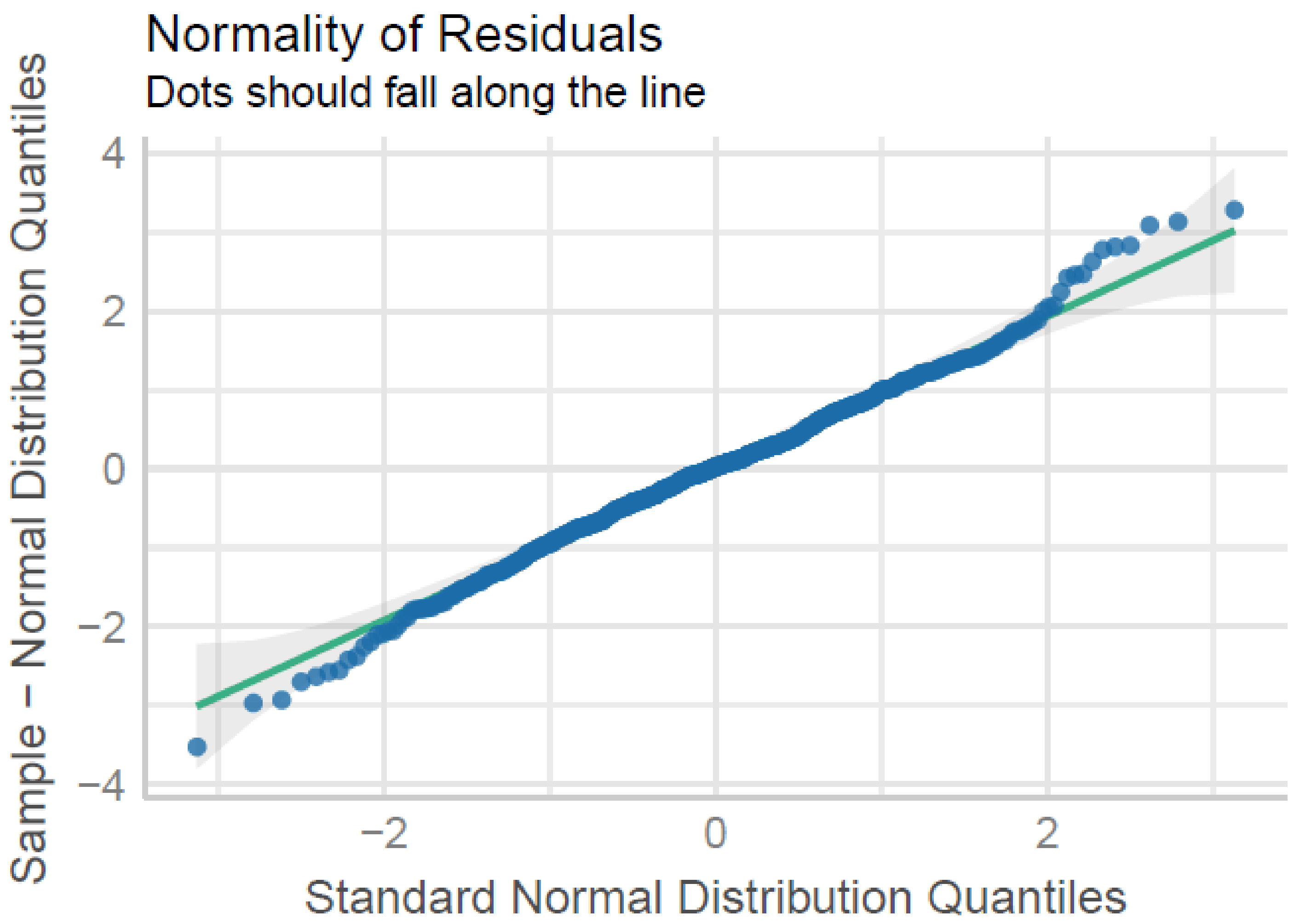
Homoscedasticity assumption







Normality of the residuals



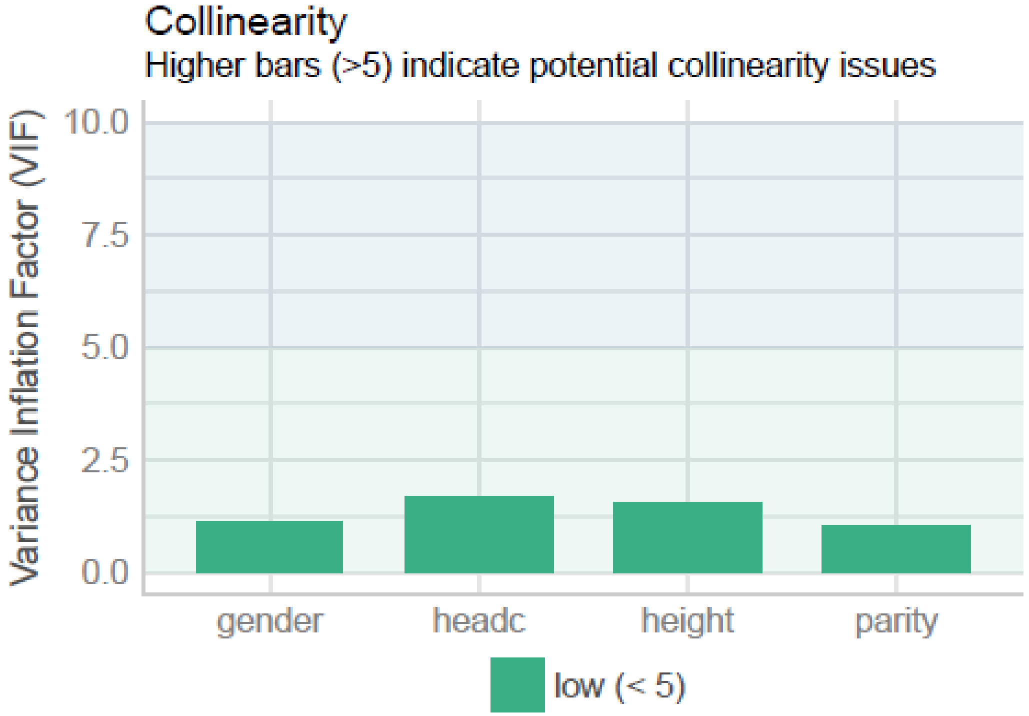
Multicollinearity (between explanatory variables)

- Two or more **explanatory** variables are significantly related to one other, conditional on the other explanatory variables

- The amount of mulitcollinearity in a model can be estimated by the variance inflation factor (VIF).

VIF< 5 ➲ no-multicollinearity

variables	VIF
(Intercept)	NA
height	1.568948
headc	1.673260
genderMale	1.135022
parityOne sibling	1.023174
parity2 or more siblings	1.023174



Coefficient of determination:

$$R^2 = \frac{\text{explained variation}}{\text{total variation}}$$

(R^2 : 0 to 1)

a measure of '**goodness of fit**' of the regression line to the data

Close to 1 ⇨ a large proportion of the variability in the response has been explained by the regression.

The **adjusted R** square is the R square value adjusted for the number of explanatory variables included in the model. In our example, adjusted $R^2 = 0.59$:

59% of the variation in infant's weight can be explained by the variables in the model.

AIC: compare different models

the smaller value of AIC the better the model

Presentation of the results

Variables	Univariable Analysis			Multivariable Analysis		
	Unadjusted β	95%CI	p-value	Adjusted β	95% CI	p-value
height (cm)	178	(164, 193)	<0.001	130	(113, 147)	<0.001
gender						
male/female	452	(358, 545)	<0.001	197	(128, 265)	<0.001
parity						
1 sibling/Singleton	130	(8, 252)	0.037	82	(3, 161)	0.041
2 or more siblings/ Singleton	192	(68, 316)	0.002	105	(24, 185)	0.011
head circumference (cm)	275	(246, 304)	<0.001	110	(79, 140)	<0.001
education						
year12/year10	58	(-88, 203)	0.44			
tertiary/year10	6.6	(-106, 119)	0.91			

β: coefficient of the explanatory variable, CI: Confidence Interval

Thank you!