Linear Regression Models

Multi-variable models

Konstantinos I. Bougioukas, MSc, PhD

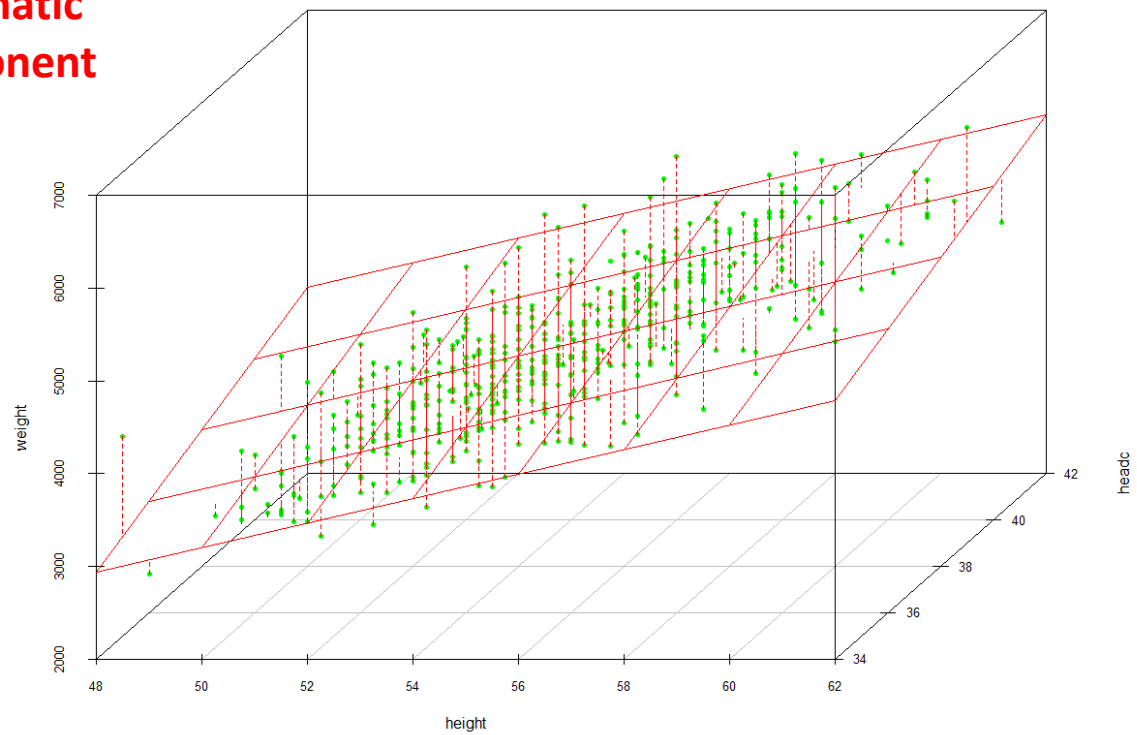
School of Psychology
Faculty of Philosophy
Aristotle University of Thessaloniki

Multiple Linear regression

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

Systematic component

Random error



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

Model building: Possible strategies

– **sample size calculation** (at least 10-15 participants per variable; 50 + 8k, where k is the number of variables; 104 + k)

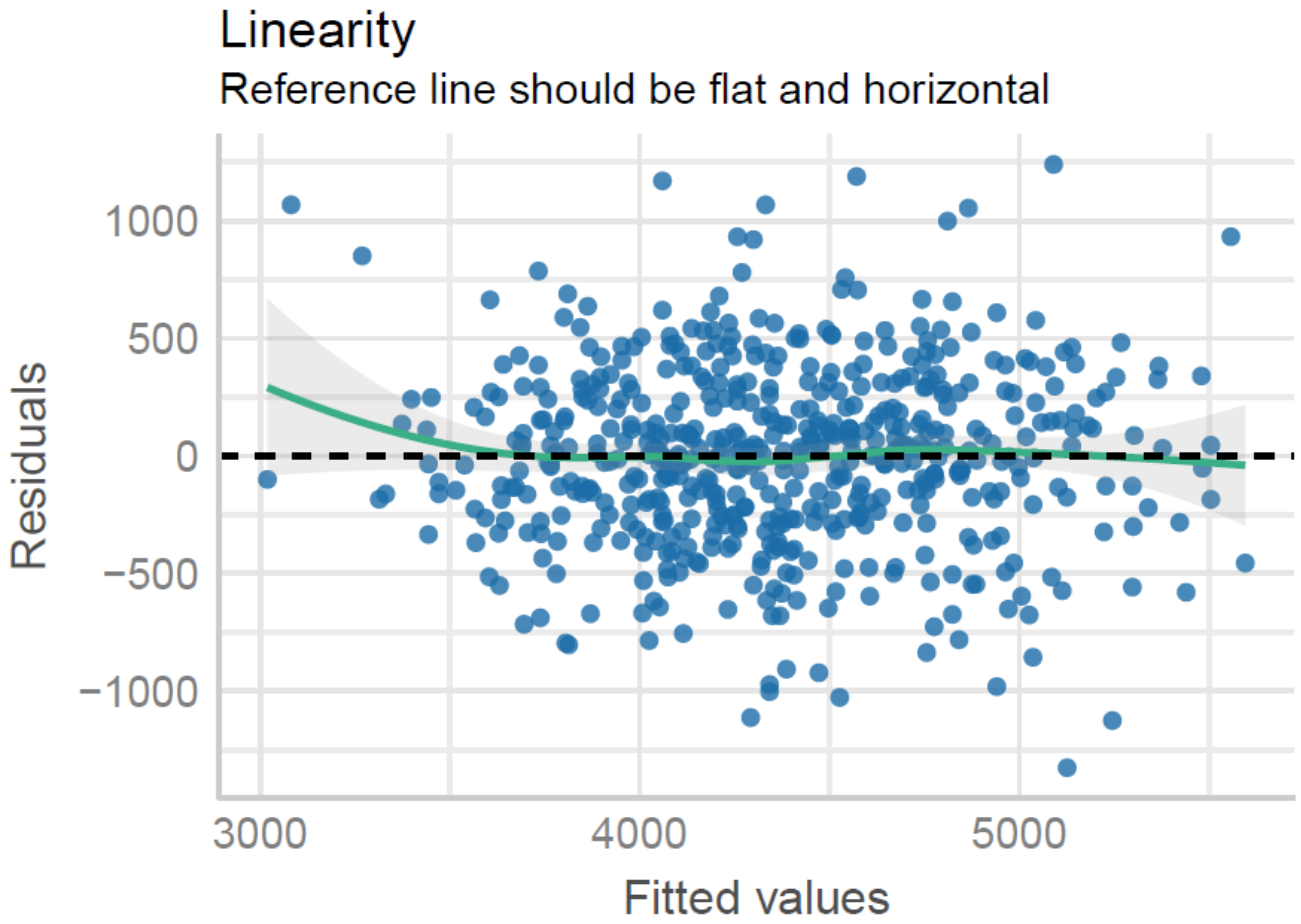
– **existing knowledge** (e.g., confounders, moderation effects) should be used

– models should be **interpretable**

Possible strategies

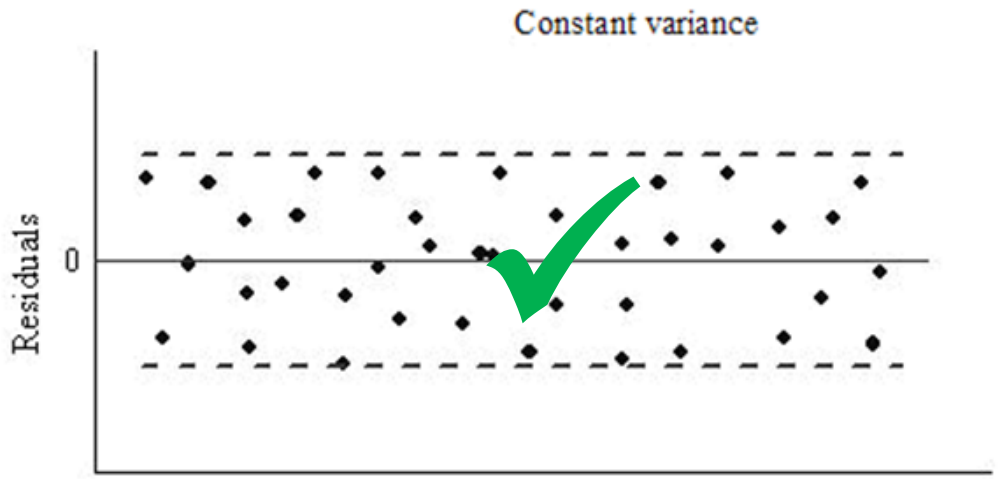
- Simultaneous Regression
- Hierarchical Regression (blocks)
- Purposeful selection process (e.g., the variables that have a p-value <0.2 in the univariable analysis are candidate variables for the model)
- Automated regression methods (using backward elimination, forward selection, or both[stepwise selection])

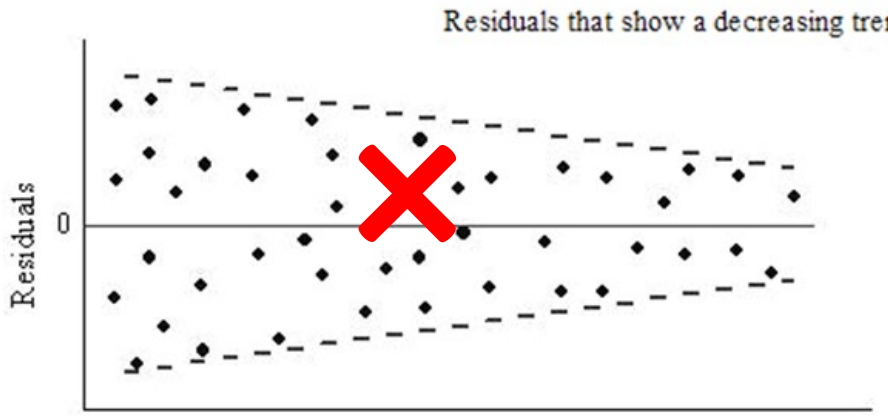
Checking assumptions: Linearity

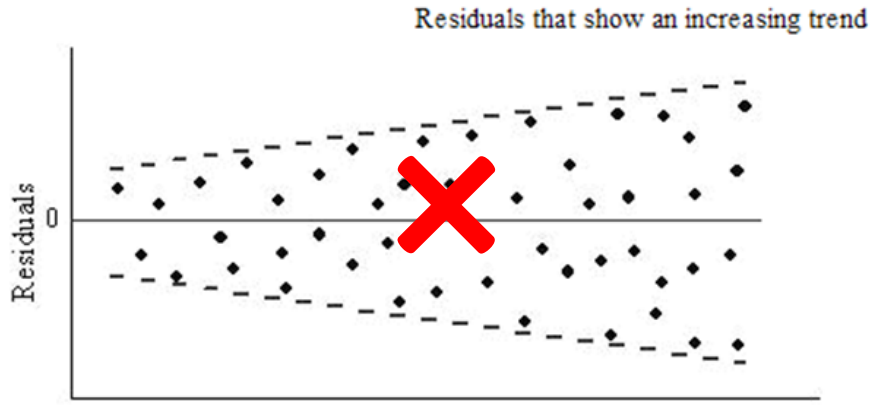


If you find equally spread residuals around a **horizontal line** without distinct **patterns**, that is a good indication for linear association.

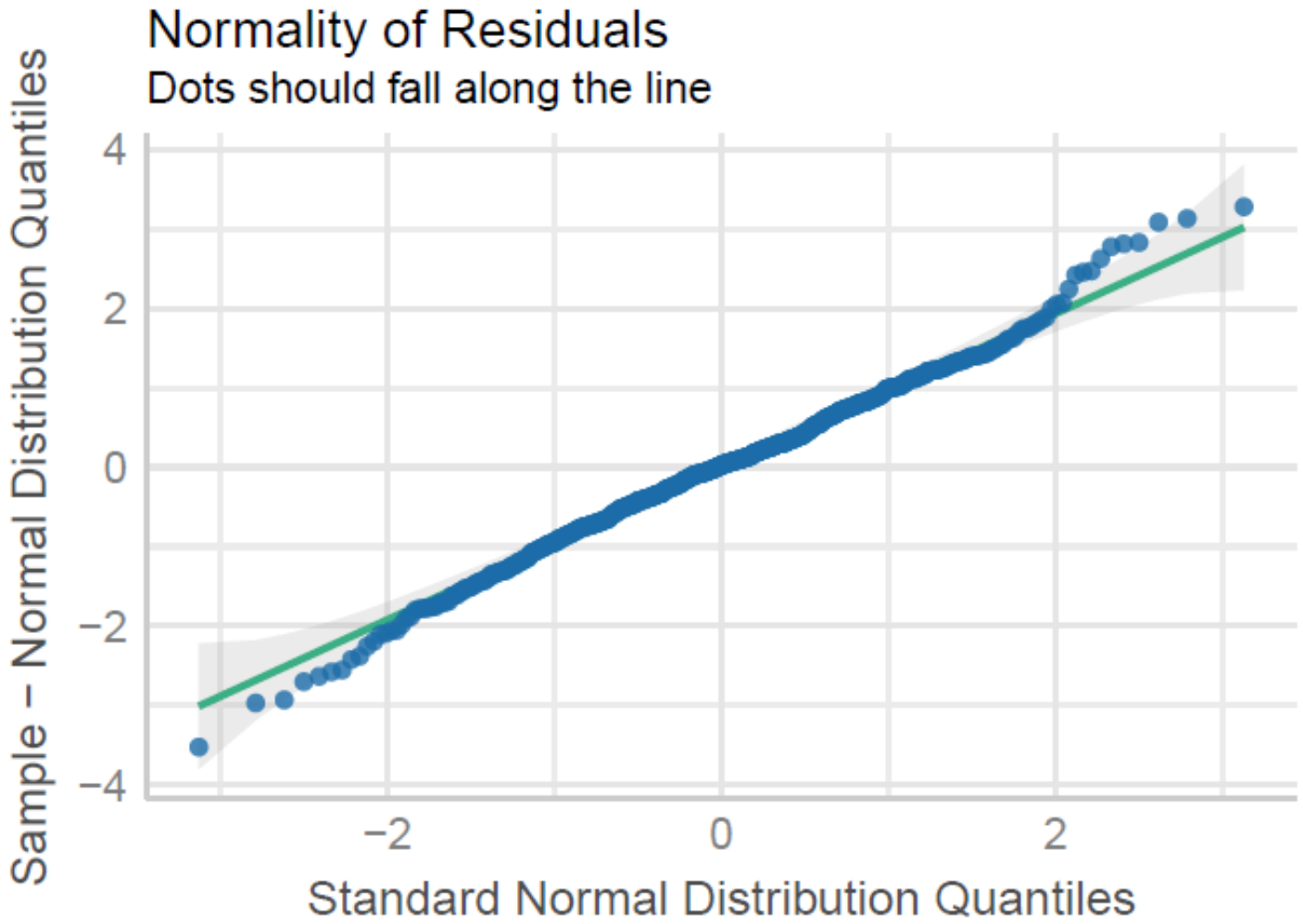
Homoscedasticity assumption







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 multicollinearity in a model can be estimated by the variance inflation factor (VIF).

VIF< 5 ➲ no-multicollinearity

Collinearity Statistics

	VIF	Tolerance
height	1.26	0.79
headc	1.29	0.77
gender	1.07	0.94
parity	1.02	0.98
education	1.02	0.98

The smallest possible value for VIF is 1, which indicates the complete absence of collinearity. A VIF less than 5 indicates a low correlation of that explanatory variable with other variables (in practice no-multicollinearity).

Assessing a model and compare models

Coefficient of determination:

$$R^2 = \frac{\exp lained\ \text{var}\, iation}{total\ \text{var}\, iation}$$

(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