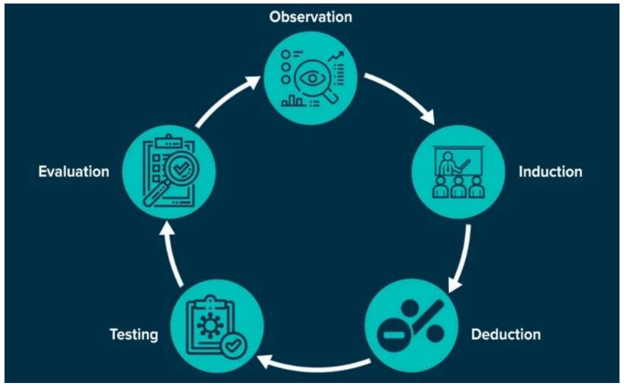
Empirical research in management and economics

Simple regression

Thorsten Pachur

Technical University of Munich
School of Management
Chair of Behavioral Research Methods



Recap from last session

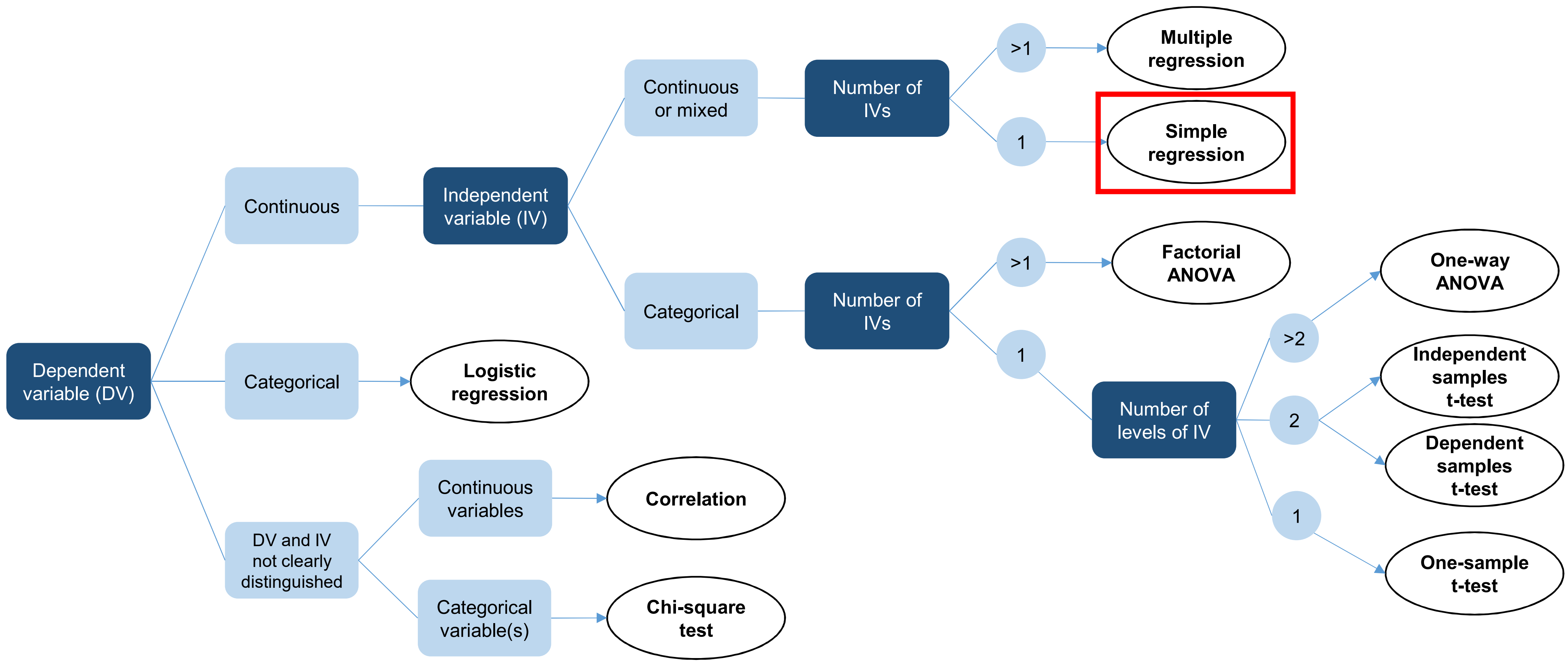
- Which statistical test is indicated for each of the following situations? For each test also give the test statistic that is used to compute a p-value.
  - Comparing the means of three or more groups across one or several factors
  - Association of two nominal-level variables
  - Comparing the means of two independent groups
- Give effect size measures for each of the tests
- Imagine that in a factorial ANOVA, you obtained a p-value of .02 for the interaction between two factors. How do you interpret this result?

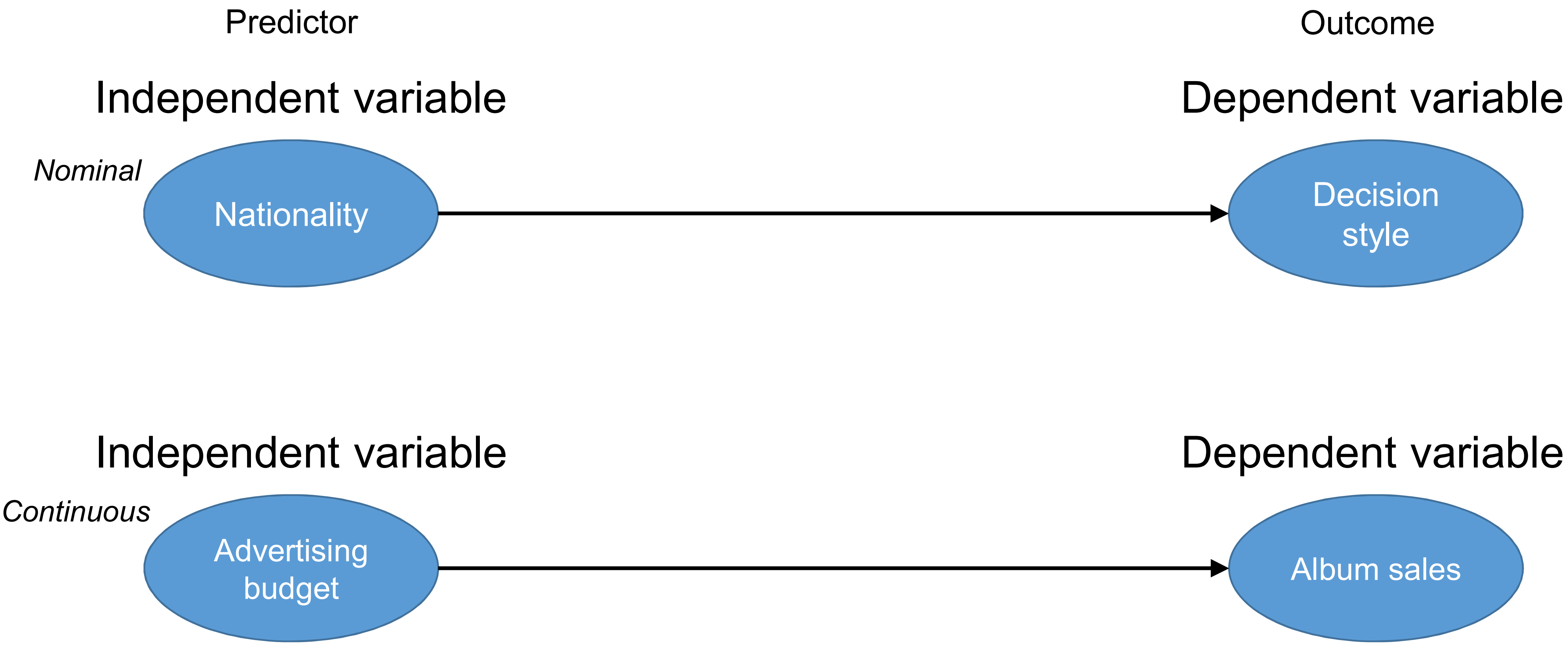
Agenda for the semester

Session	Date	Topic
1	13 October	Introduction
2	20 October	Descriptive data analysis
3	27 October	Hypothesis development and measurement
4	3 November	Inferential data analysis I
5	10 November	Inferential data analysis II
6	**17 November**	**Simple regression**
7	24 November	Multiple regression
8	1 December	Logistic regression
9	8 December	Factor analysis
10	15 December	Cluster analysis
11	12 January	Conjoint analysis
12	19 January	The replication crisis and open science
13	26 January	Summary and questions
	11 February	Exam

Goals for this week

- You know the purpose of conducting a regression analysis
- You have understood the parameters of a simple linear regression model and how the parameters are estimated
- You know how to evaluate the results of a regression model analysis statistically
- You know the assumptions underlying simple linear regression—and how to check whether they are fulfilled


Dependent variable (DV)
Continuous
Categorical
DV and IV not clearly distinguished
Independent variable (IV)
Continuous or mixed
Categorical
Number of IVs
Number of IVs
>1
1
>1
1
Multiple regression
Simple regression
Factorial ANOVA
Number of levels of IV
>2
2
1
One-way ANOVA
Independent samples t-test
Dependent samples t-test
One-sample t-test
Logistic regression
Continuous variables
Categorical variable(s)
Correlation
Chi-square test


Predictor
Outcome
Independent variable
Dependent variable
Nominal
Nationality
Decision style
Independent variable
Dependent variable
Continuous
Advertising budget
Album sales

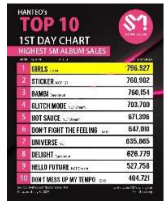
Goals of a regression analysis

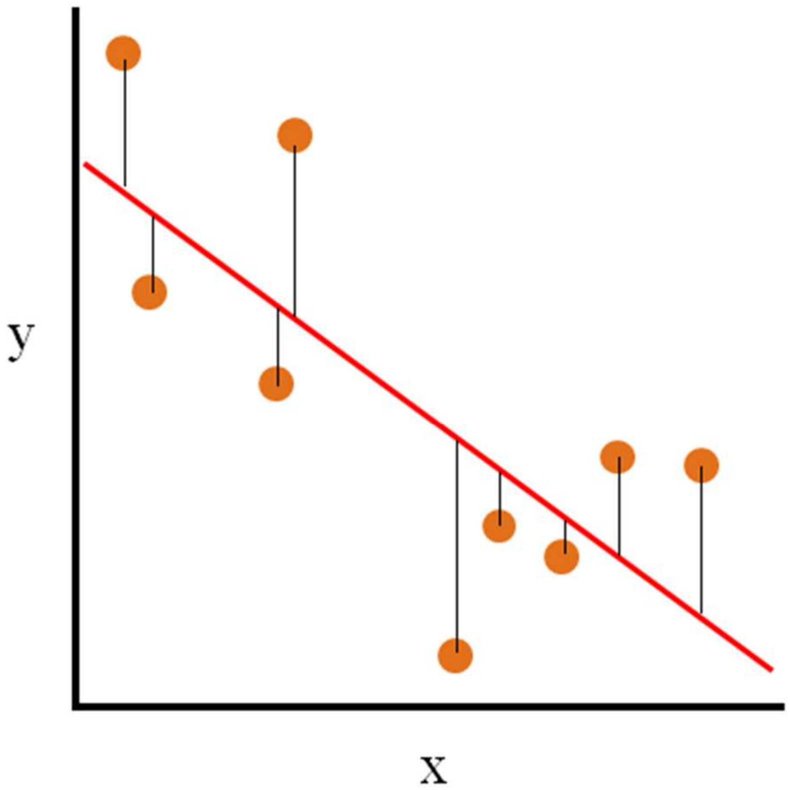
- **To describe** a relationship between a dependent variable (outcome variable) and a (set of) predictor(s) in a ***given*** set of observations
- **To predict** the dependent variable (*outcome variable*) from a (set of) predictor(s) for a ***new*** set of observations

Examples

- How is risk taking associated with a person‘s age, wealth, and affect?

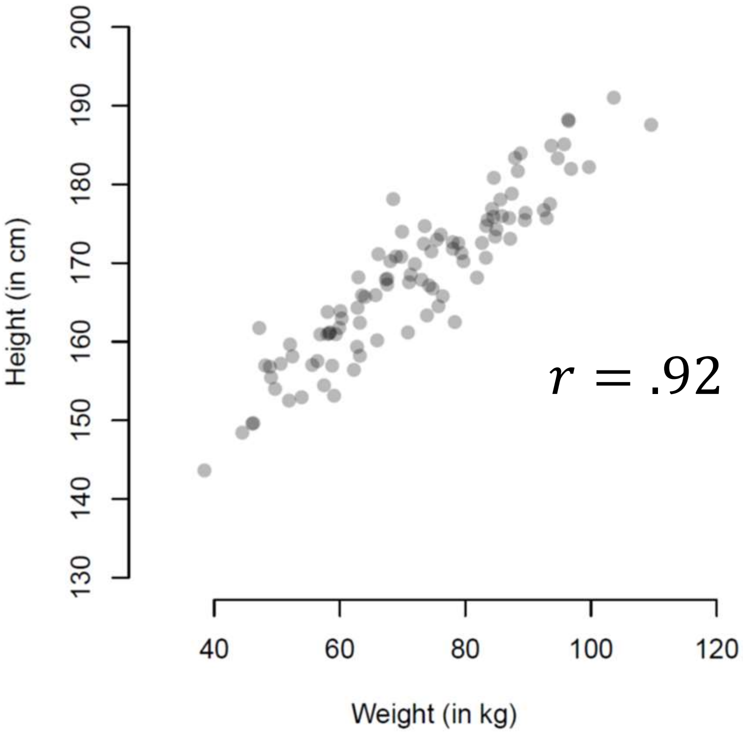
- How are sales of music albums associated with the size of the advertising budget and the amount of airplay?





Simple linear regression

The relationship between two variables

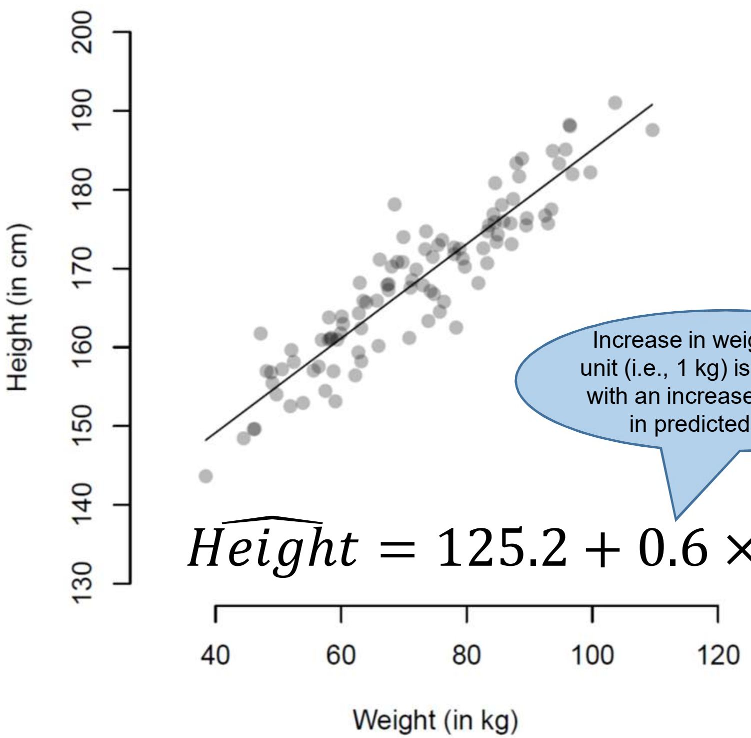


→ How to model the relationship between height and weight?

→ How to predict a new person‘s height from her weight?

Regression line

Dependent variable (outcome), Y



$$\widehat{Height} = 125.2 + 0.6 \times Weight$$

Increase in weight by one unit (i.e., 1 kg) is associated with an increase by 0.6 cm in predicted height

Independent variable (predictor), X

Predicted value of Y

$$\hat{Y} = b_0 + bX$$

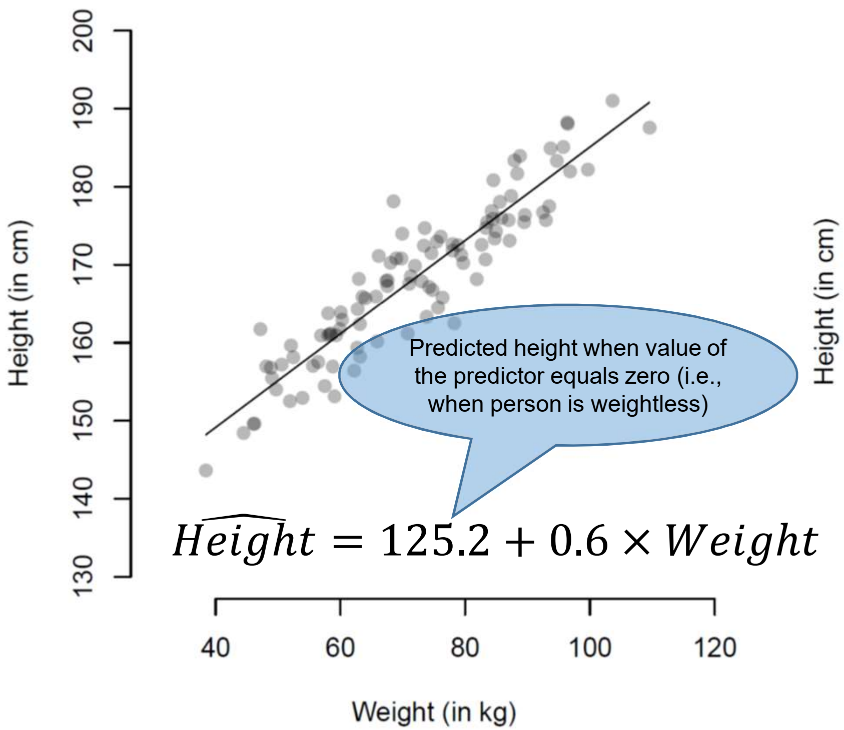
Slope
(the amount of change in $\hat{Y}$ associated with a one-unit change in X: $b = \frac{\Delta \hat{Y}}{\Delta X}$)
→ regression coefficient

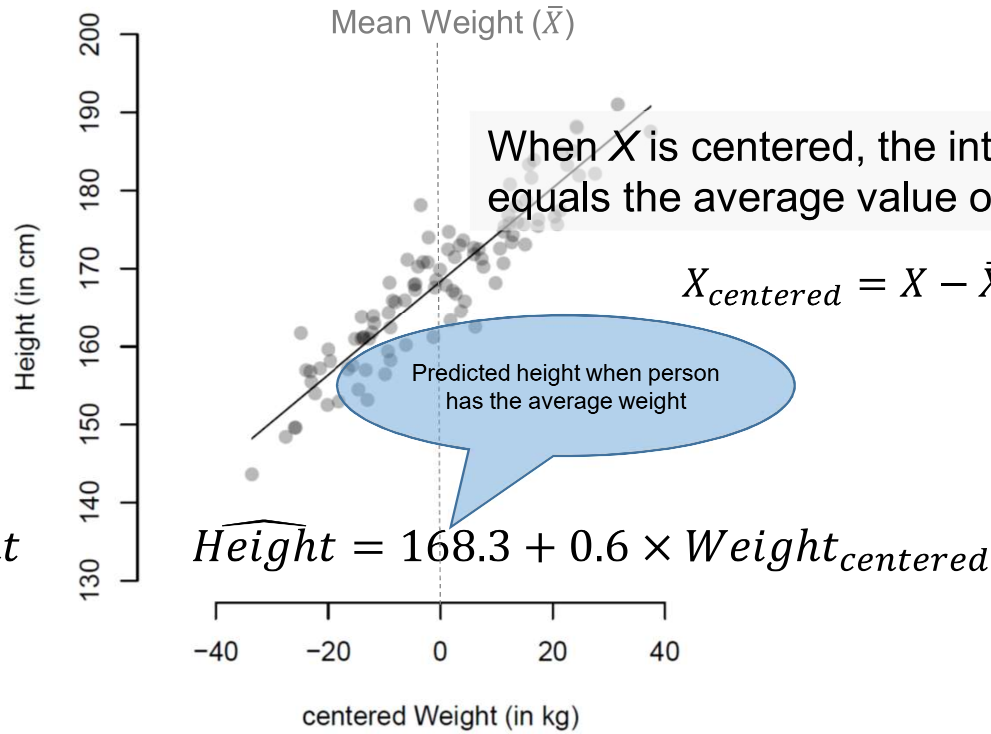
Value of the predictor variable

Intercept
(the value of $\hat{Y}$ when $X = 0$)

Centering the predictor

(to facilitate the interpretation of the intercept)



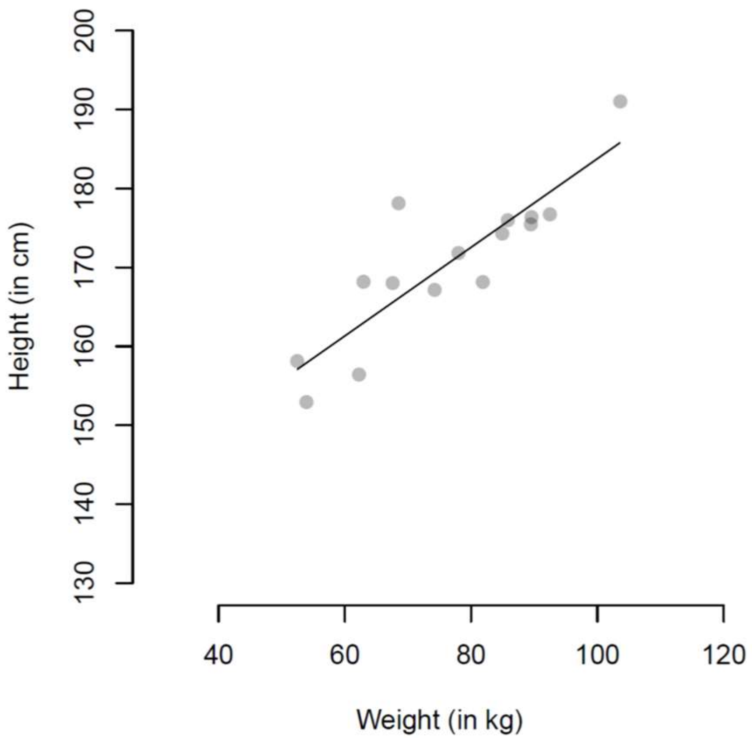


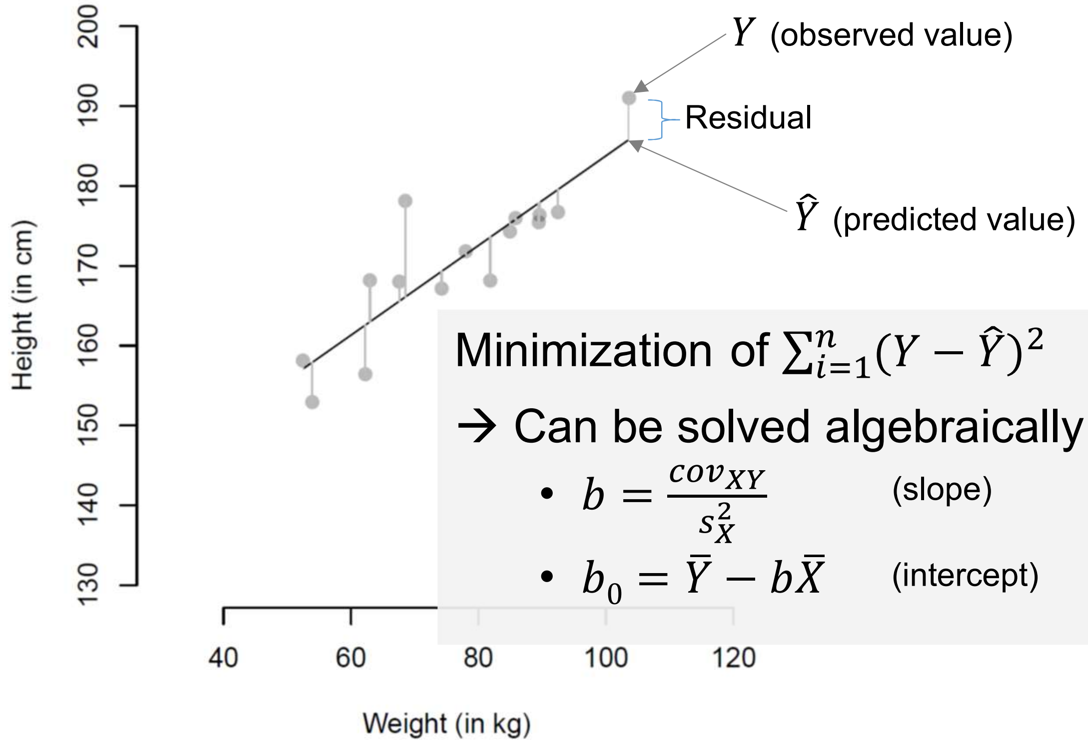
When X is centered, the intercept equals the average value of Y

$$X_{centered} = X - \bar{X}$$

Method of least squares

"Ordinary least squares" (OLS)



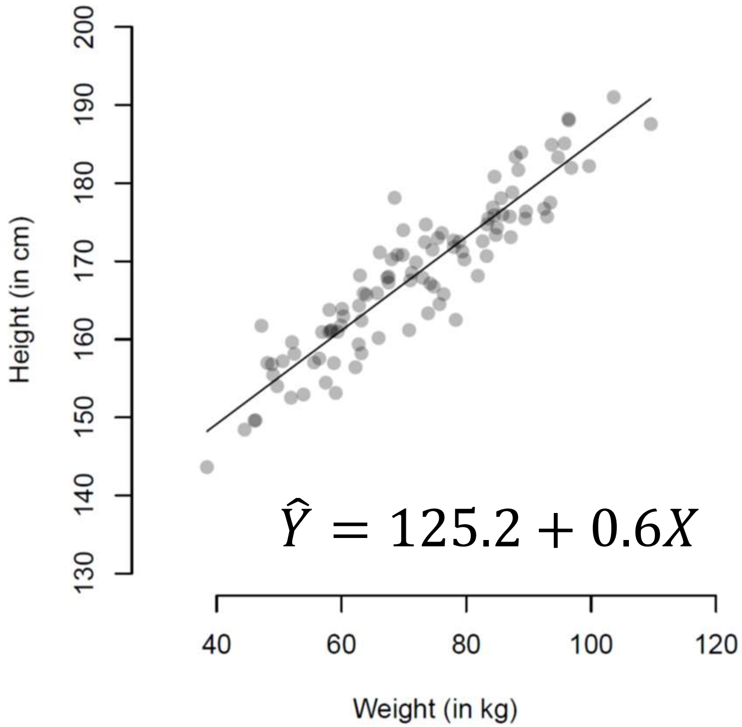


Minimization of $\sum_{i=1}^{n}(Y - \hat{Y})^2$

→ Can be solved algebraically

- $b = \frac{cov_{XY}}{s_X^2}$ (slope)
- $b_0 = \bar{Y} - b\bar{X}$ (intercept)

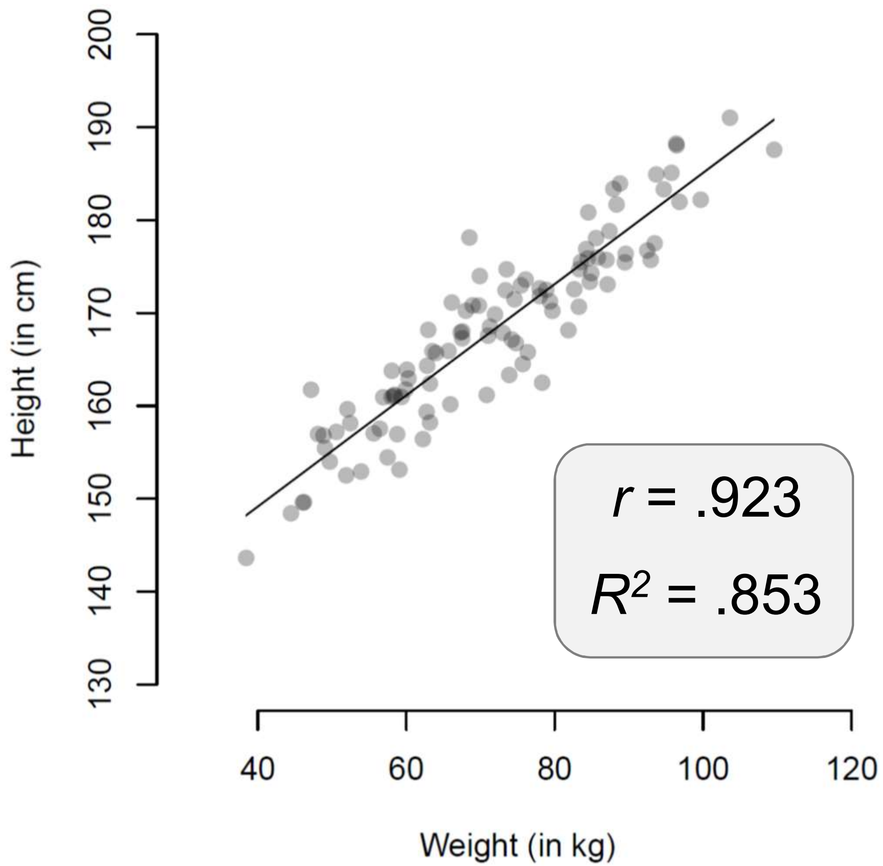
Statistical evaluation of a regression model



$$\hat{Y} = 125.2 + 0.6X$$

- How much variance in the outcome variable is explained by the predictor?
- Is the value of the regression coefficient *b* significantly different from zero?

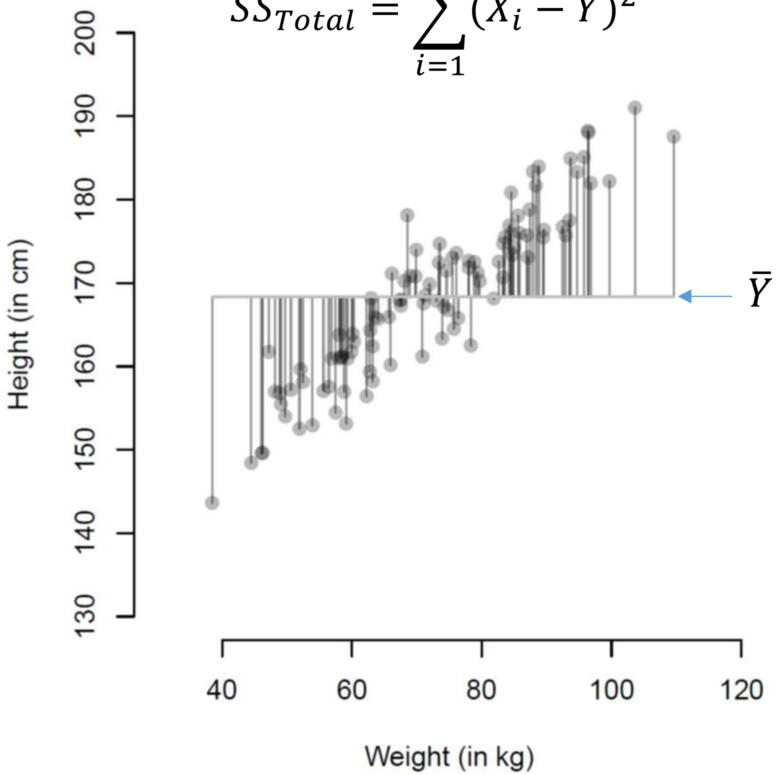
Amount of explained variance



Amount of explained variation

Total variation

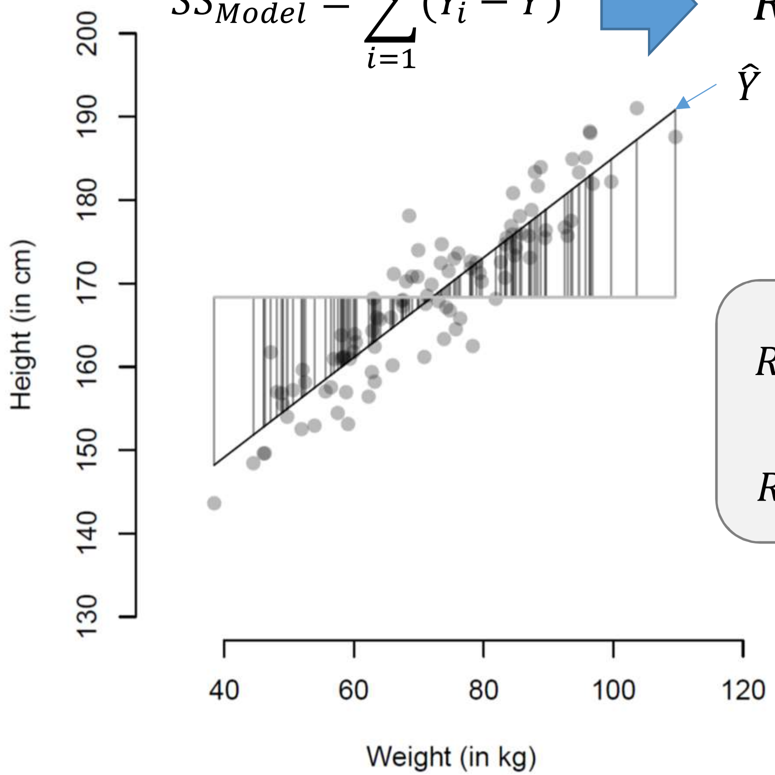
$$SS_{Total} = \sum_{i=1}^{N} (X_i - \bar{Y})^2$$



Variation explained by the regression model

$$SS_{Model} = \sum_{i=1}^{N} (\hat{Y}_i - \bar{Y})^2$$

$$R^2 = \frac{SS_{Model}}{SS_{Total}}$$



$$R^2 = \frac{8433.8}{9884.1}$$

$$R^2 = .853$$

Statistical evaluation of a regression model

- Amount of explained variance

SS: Sum of squares

$$R^2 = \frac{\sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2} = \frac{SS_{\hat{Y}}}{SS_Y}$$

$$R^2_{adjusted} = 1 - \frac{(1 - R^2)(N - 1)}{N - k - 1}$$

N: Sample size k: Number of predictors

$$R^2 = .853$$

$N = 100$ $k = 1$

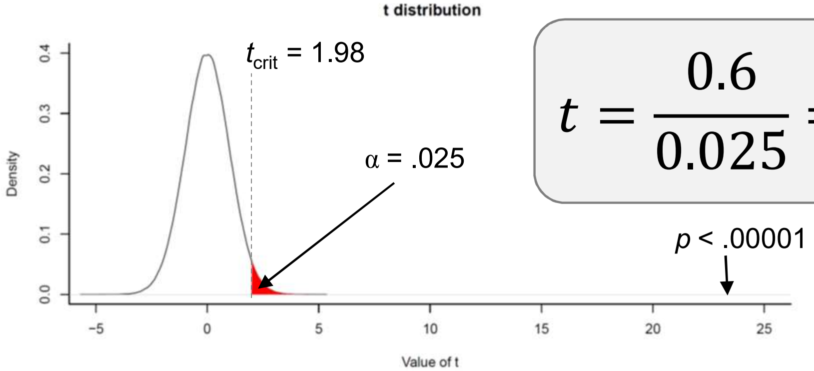
$$R^2_{adjusted} = .852$$

- Evaluating the regression coefficient *b*

$$t = \frac{b}{SE_b}$$

$$SE_b = \frac{SD_{Y-\hat{Y}}}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2}}$$

$$df = N - (k + 1)$$



$$t = \frac{0.6}{0.025} = 23.87$$

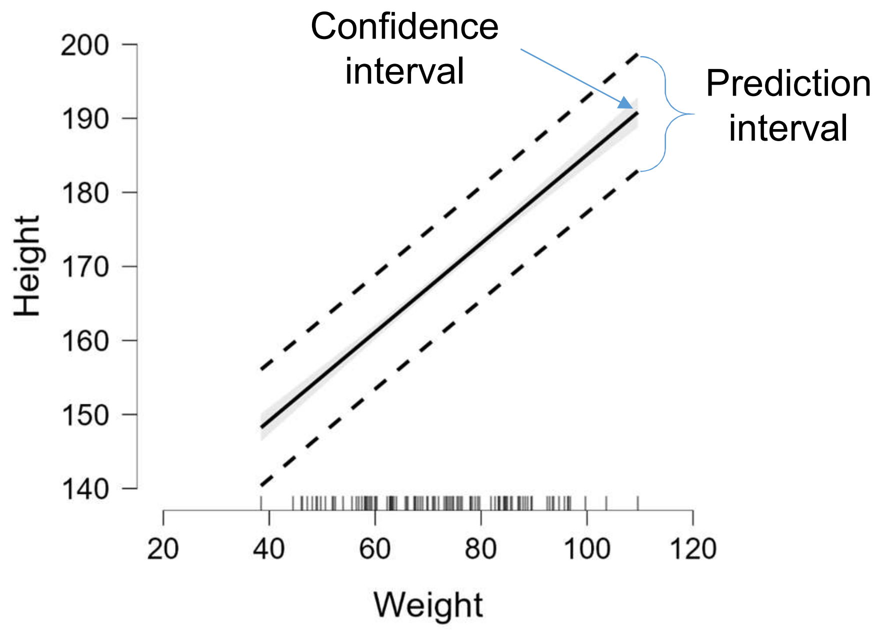
Confidence limits on the prediction

- **Confidence interval**
  → Precision of the estimate of the average *Y* for people with a given *X*

$$\hat{Y}_i \pm t_{\alpha/2}\sqrt{\frac{\sum_{i=1}^{N}(Y_i-\hat{Y}_i)^2}{N-2}} \times \sqrt{\frac{1}{N}+\frac{\sum_{i=1}^{N}(X_i-\bar{X})^2}{(N-1)s_X^2}}$$

- **Prediction interval**
  → Precision of the estimate of an individual person's Y with a given *X*

$$\hat{Y}_i \pm t_{\alpha/2}\sqrt{\frac{\sum_{i=1}^{N}(Y_i-\hat{Y}_i)^2}{N-2}} \times \sqrt{1+\frac{1}{N}+\frac{\sum_{i=1}^{N}(X_i-\bar{X})^2}{(N-1)s_X^2}}$$



Assumptions in regression analysis

- *Linearity*: The relationship between outcome variable and the predictor variable(s) is linear

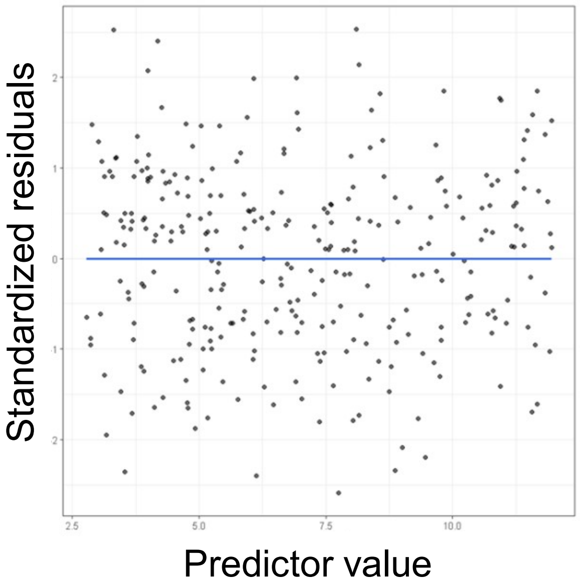
- *Homoscedasticity*: At each level of the predictor variable(s), the variance of the residuals is the same

- The residuals are *normally distributed*

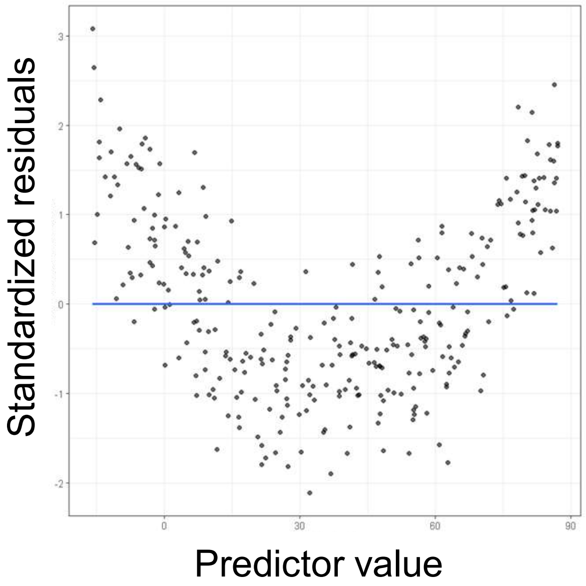
Checking for linearity

→ Plot residuals for different levels of the predictor variable:
*Is the **average value** of the residuals similar across different levels of the predictor?*

No problem

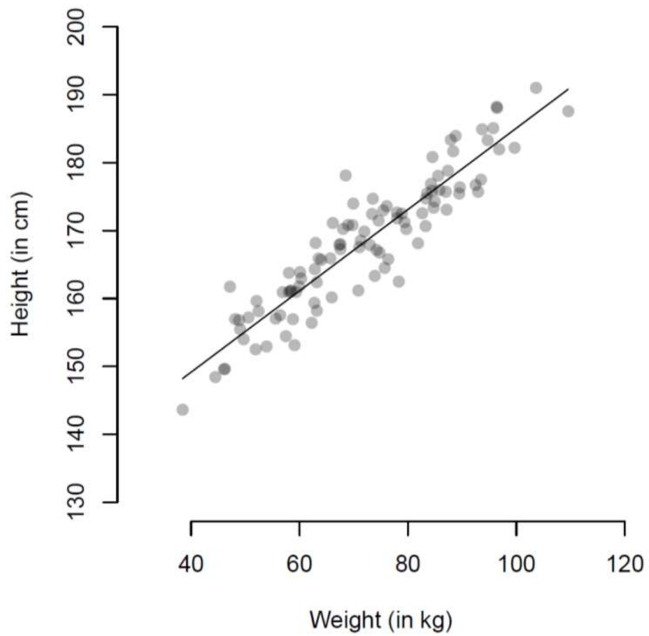


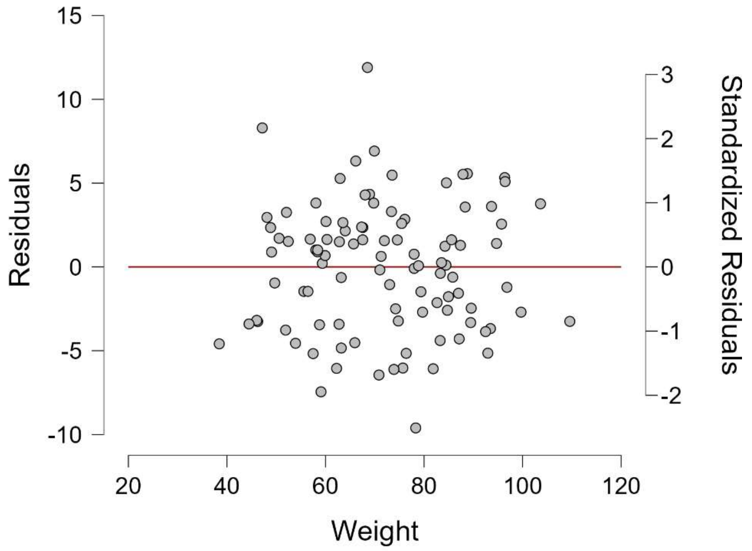
Nonlinear association between predictor(s) and outcome variable



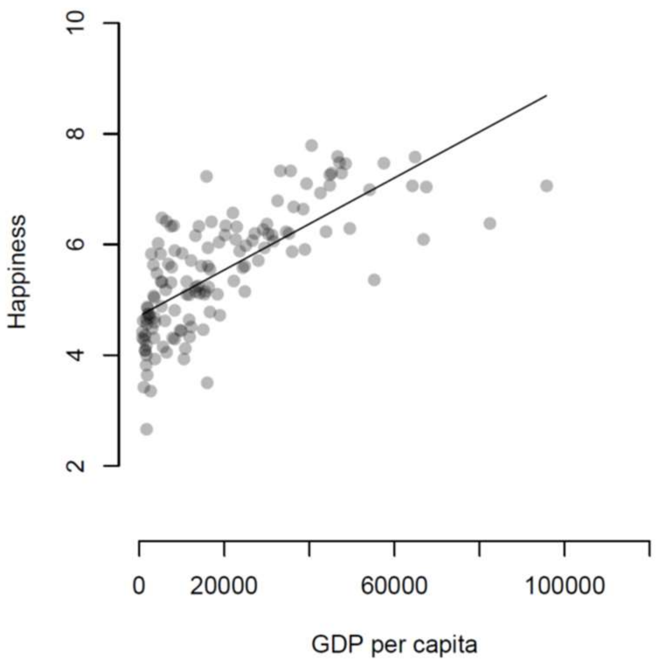
Checking for linearity

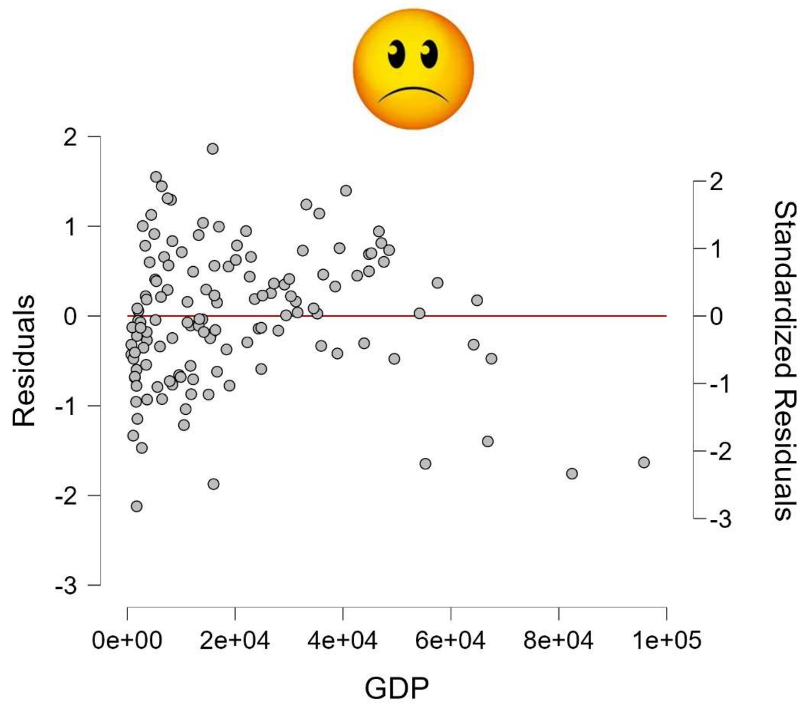
→ Plot residuals for different levels of the predictor variable



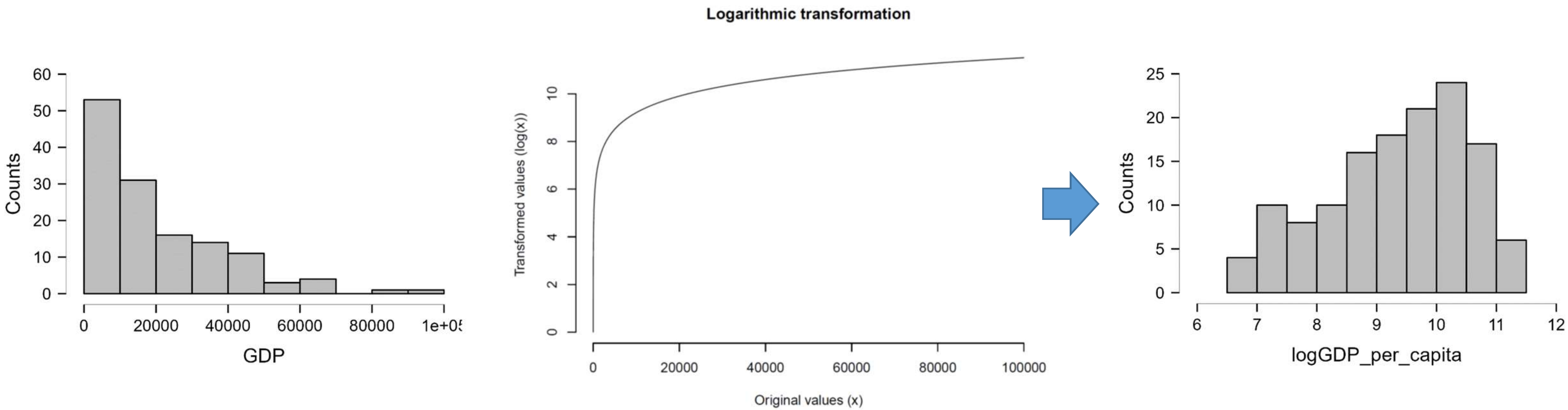


Checking for linearity

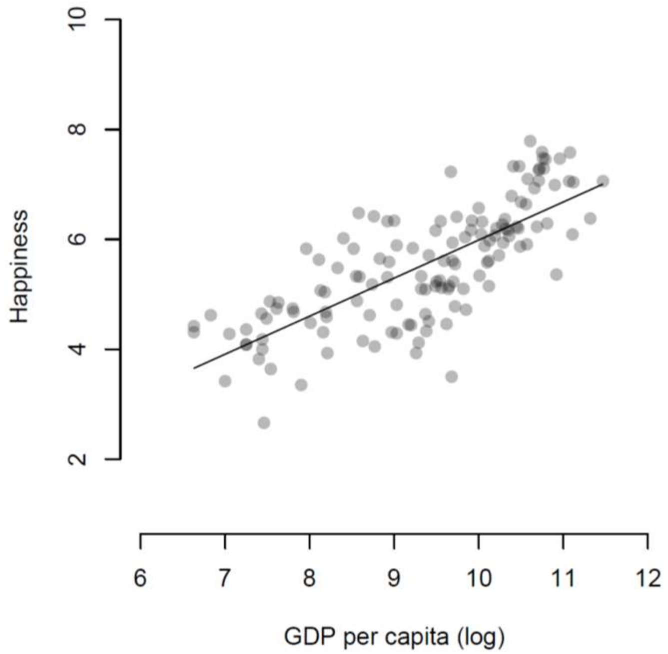


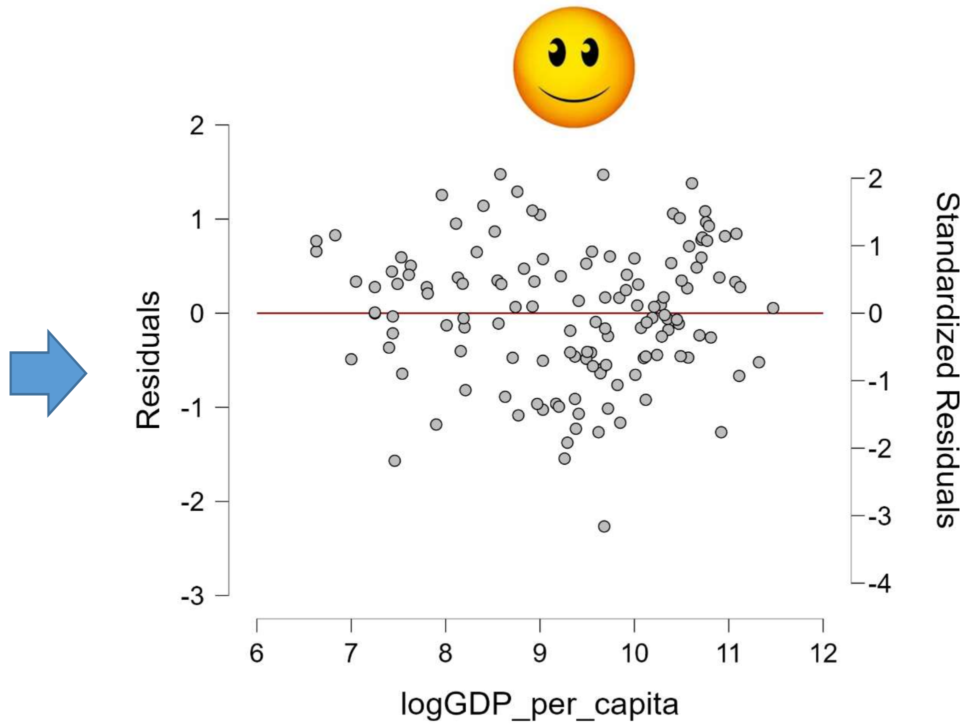


Logarithmic transformation



Checking for linearity



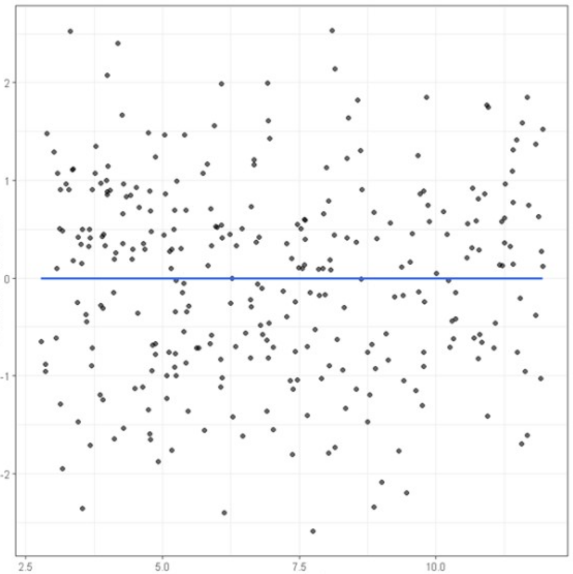


Checking for homoscedasticity

→ Plot residuals for different levels of the predictor variable:
*Is the **variability** of the residuals similar across different levels of the predictor?*

Homoscedasticity fullfilled

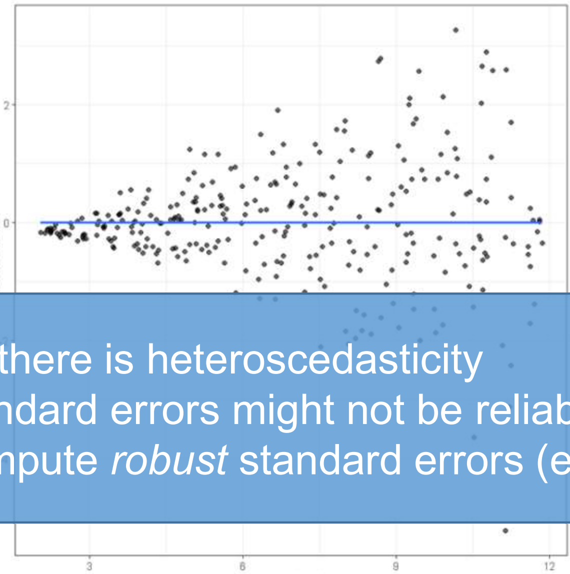
Standardized residuals



Predictor values

Heteroscedasticity

Standardized residuals



Predictor values

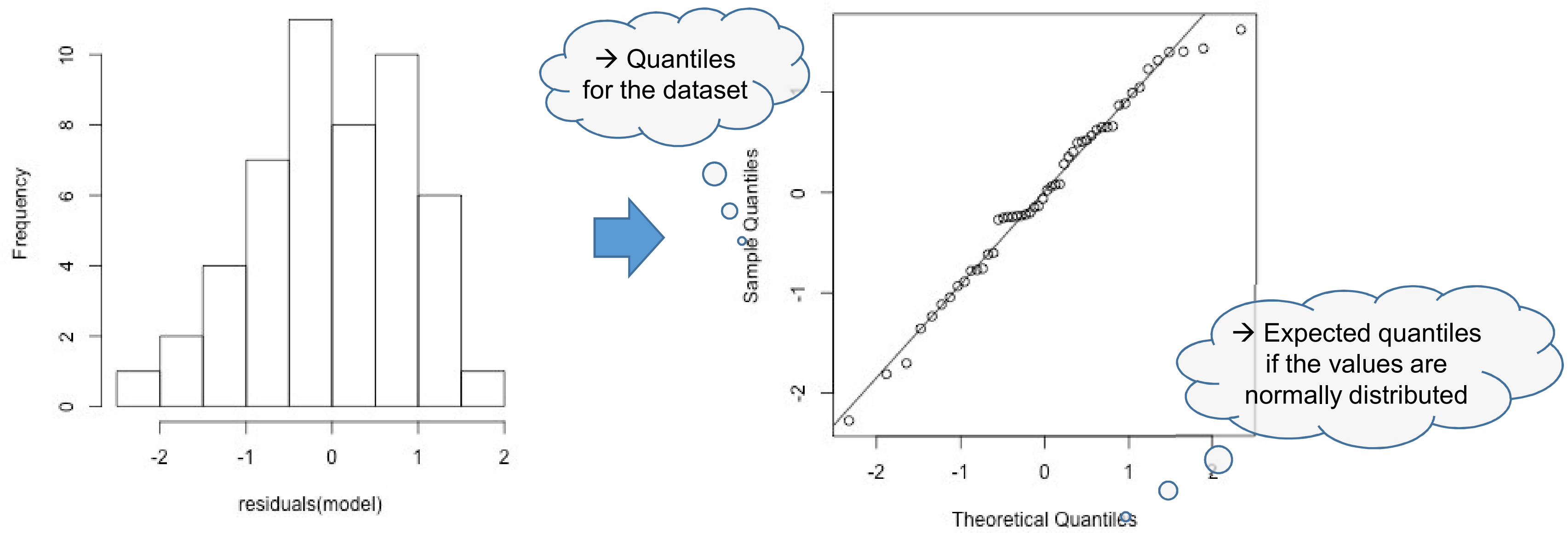
When there is heteroscedasticity
- Standard errors might not be reliable
- Compute *robust* standard errors (e.g., with R)

Checking for normally distributed residuals

→ Q-Q (quantile-quantile) plot

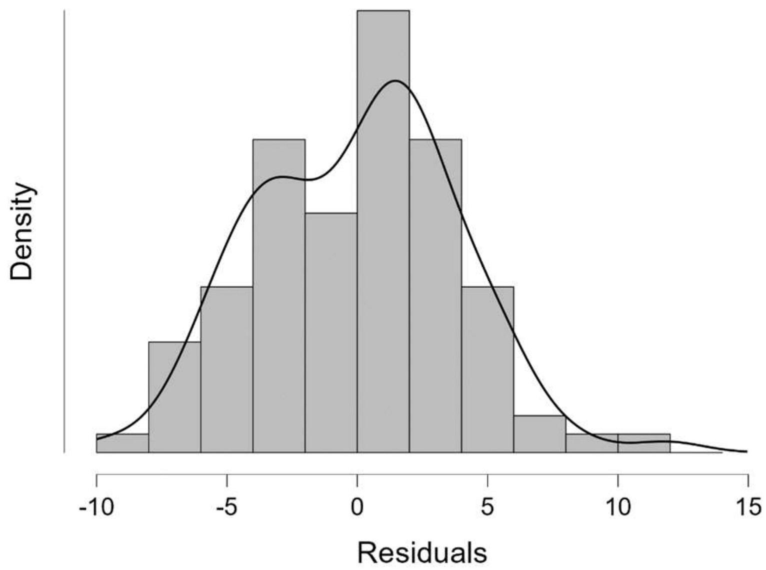
(*Quantile*: Expresses how many values in the distribution are below a certain value; e.g., the 50%-quantile (which is the median) is the value that is larger than 50% of the other data points in the distribution

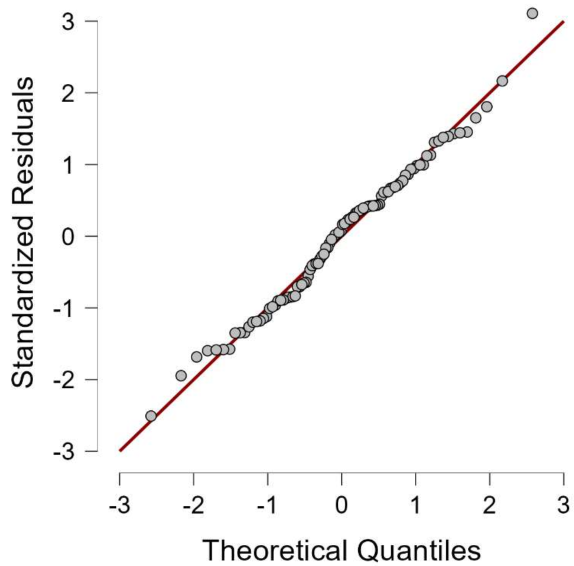
If the residuals are normally distributed, the Q-Q plot will show a straight line



Checking for normally distributed residuals

$$\widehat{Height} = b_0 + b \times Weight$$



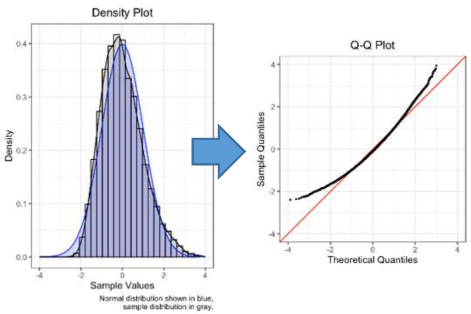


Checking for normally distributed residuals

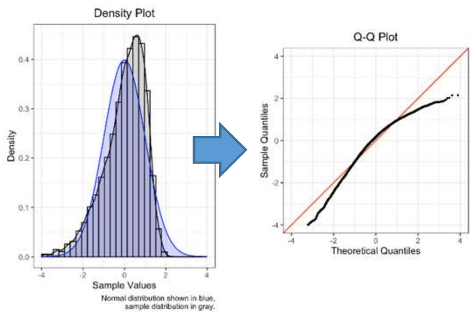
→ Q-Q (quantile-quantile) plots

Some types of deviations from a normal distribution

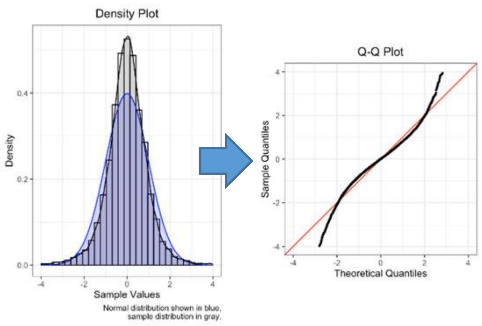
“Right skew“



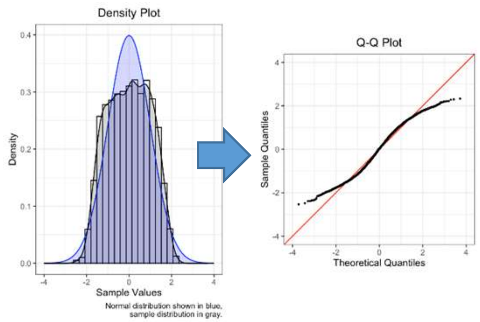
“Left skew“



“Fat tailed“



“Thin tailed“



Self-quiz questions

- What are the key purposes of estimating a regression model?
- What are the key parameters of a regression model?
- How are the parameters of a regression model estimated?
- Why can it be helpful to center a predictor?
- How is a regression model evaluated statistically—both in terms of the overall model and in terms of the regression coefficients?
- What are key assumptions in simple linear regression—and how can you check whether the assumptions are fulfilled?

Background readings for next session

Howell, D. C. (2017). Multiple regression. In: D. C. Howell, *Fundamental statistics for the behavioral sciences (9th ed.)* (p. 265–298). Wadsworth Cengage Learning, Belmont.