Linear Regression Models
Simple models

Konstantinos I. Bougioukas, MSc, PhD

School of Psychology
Faculty of Philosophy
Aristotle University of Thessaloniki

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)

Basic Properties

- They should be **valid**: provide explanations or 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**.

To Explain or to Predict?

Modeling for explanation

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

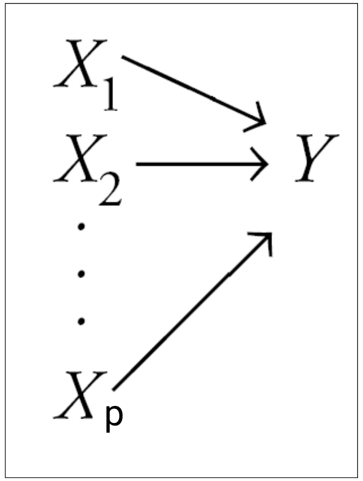
- Identification of 'important' explanatory variables
- Understanding the effects of explanatory variables
- Adjustment for variables uncontrollable by experimental design

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.

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.

Basic assumptions

Linearity: linear combination of variables

- (Relaxation: splines, fractional polynomials, etc.)

$$\widehat{y}_i = \widehat{\beta}_o + \widehat{\beta}_1 * x_{1i} + \widehat{\beta}_2 * x_{2i} + \widehat{\beta}_3 * x_{3i} + \ldots + \widehat{\beta}_p * x_{pi}$$

Additivity: sum of main effects

- (Relaxation: include interactions etc.)

Data

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

- Body weight of the infant in g (**weight**)
- 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

Dependent variable

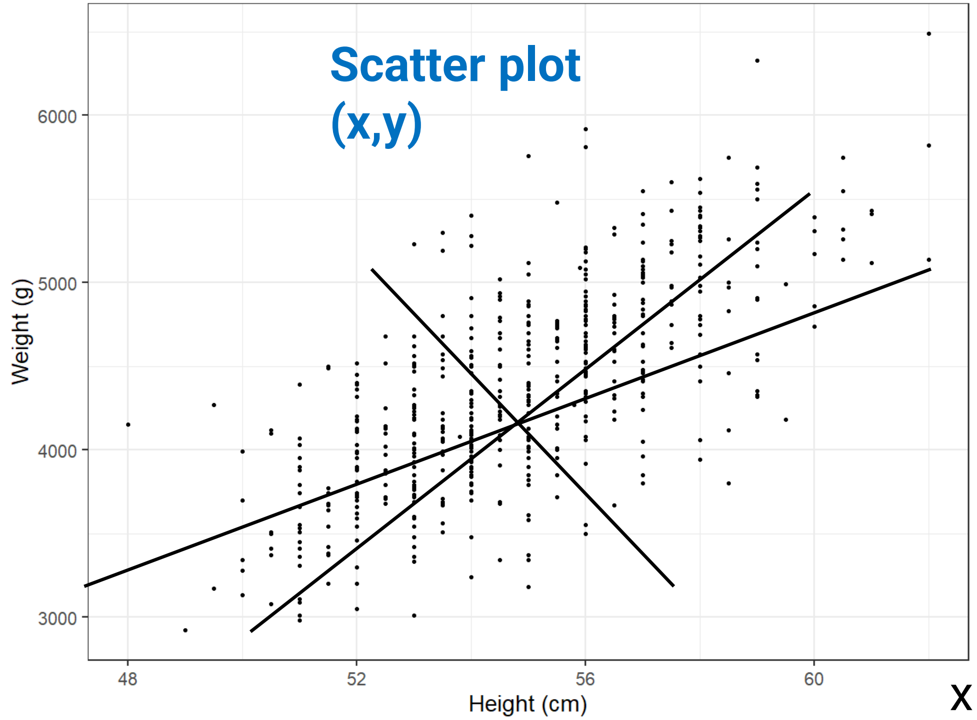
$$y_i = \boxed{\beta_o + \beta_1 * x_i} + \varepsilon_i$$

i=1,2,…,n

Systematic component

Random error

Y

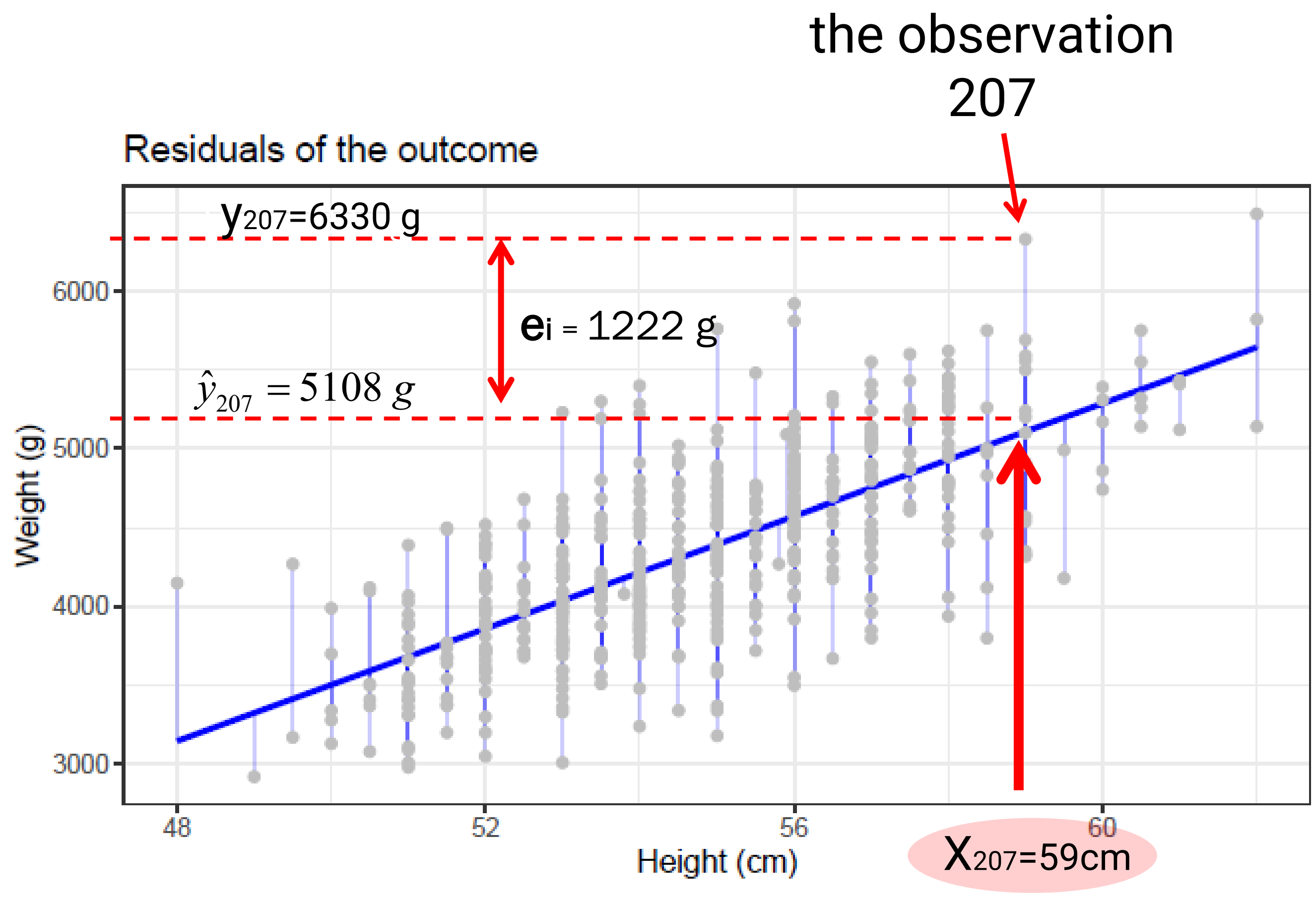


X

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

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

Continuous explanatory variable

Question:

What is the association between weight and height ?

Hypothesis Testing

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

- **H_0: β_1=0** (no association)
- **H_1: $\beta_1 \neq 0$** (there is association)

Results and interpretation

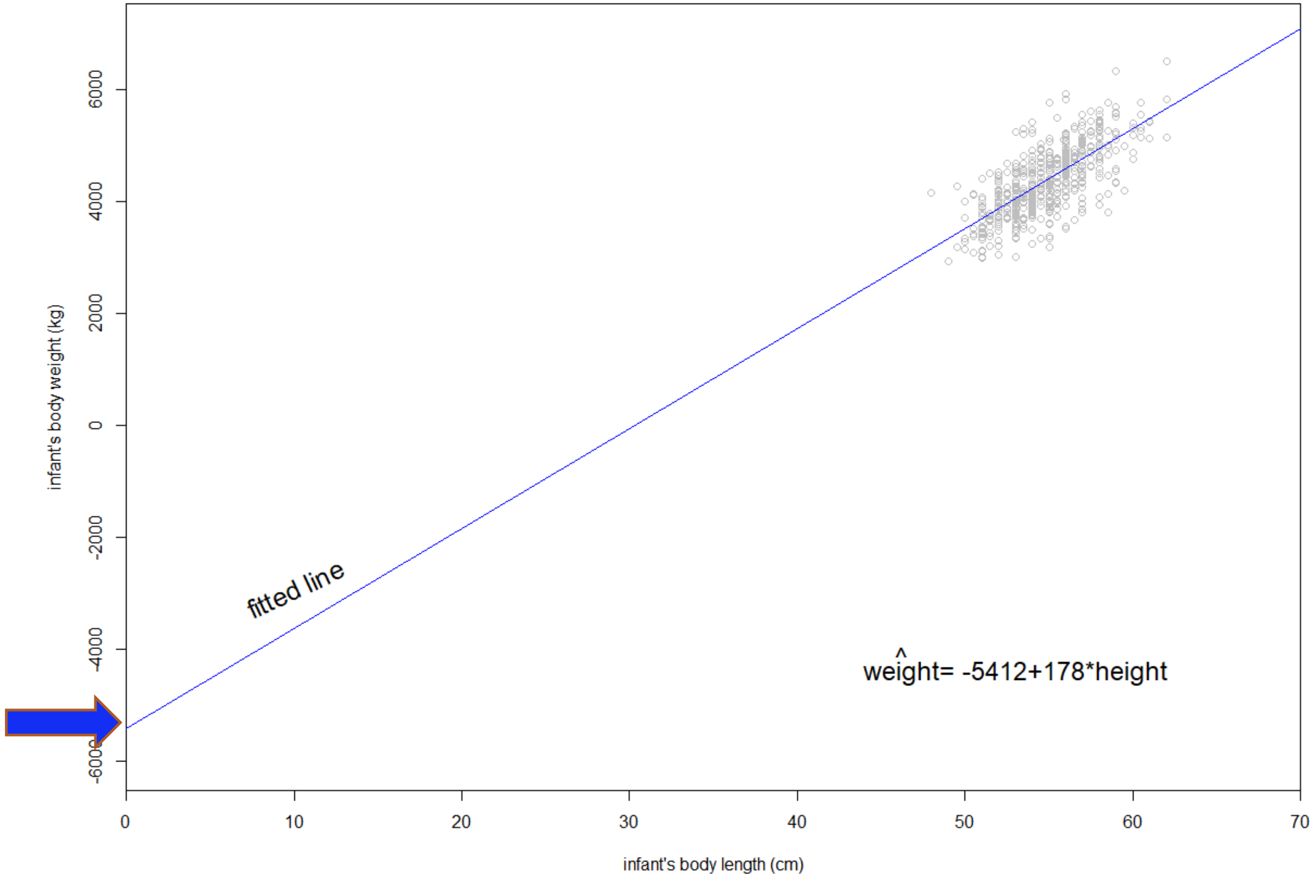
$$\hat{weight} = -5412 + 178 * height$$

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

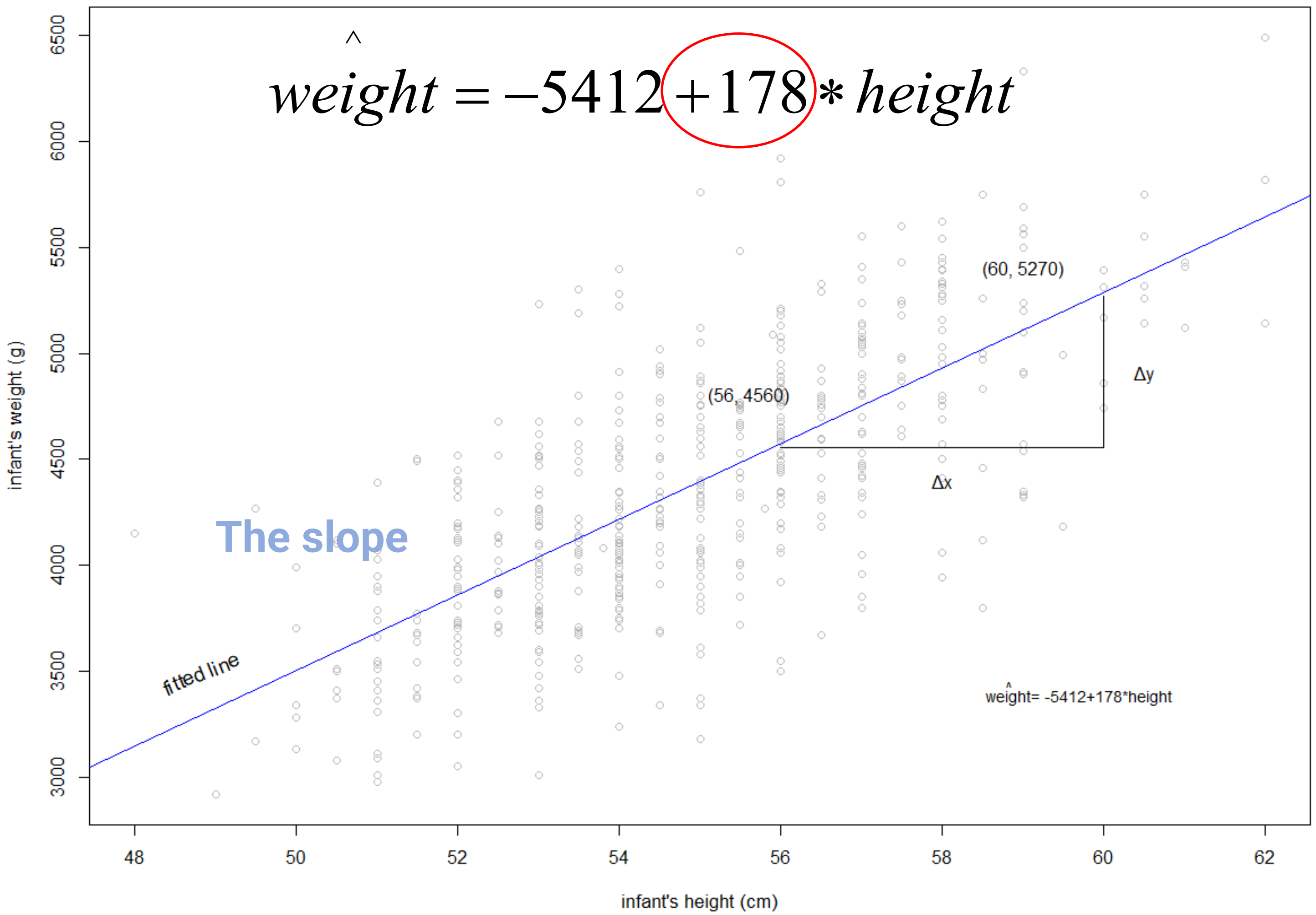
$$\hat{weight} = -5412 + 178 * height$$

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:

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

Binary explanatory variable

Question:

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

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

$$\widehat{\text{weight}} = b_o + b_1 \cdot \text{gender}$$

Results and interpretation

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

$$\widehat{\text{weight}} = 4140 + 452 \cdot \text{gender}$$

- For **females**:

  Weight = **4140 + 452*0 = 4140 g**

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

- For **males**:

  Weight = **4140 + 452*1 = 4140 + 452 = 4592 g**

  The coefficient value 452 is **the difference** (4592 – 4140) 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 of 4141 g (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?

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

Dummy variables

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

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

Dummy variables

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

$$\text{parity1} = \begin{cases} 1 & \text{if infant has one sibling} \\ 0 & \text{otherwise (ref.)} \end{cases}$$

$$\text{parity2} = \begin{cases} 1 & \text{if infant has 2 or more siblings} \\ 0 & \text{otherwise (ref.)} \end{cases}$$

parity	One sibling	2 or more siblings
Singleton (ref.)	0	0
One sibling (parity1)	1	0
2 or more siblings (parity2)	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

$$\widehat{\text{weight}} = 4259 + 130 \cdot \text{parity1} + 192 \cdot \text{parity2}$$

- 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.**

- 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**.