# SM-1402 Basic Statistics

## Chapter 5: Regression analysis *[handout version]*

Dr. Haziq Jamil
Mathematical Sciences, Faculty of Science, UBD
https://haziqj.ml

Semester II 2022/23

# Learning outcomes

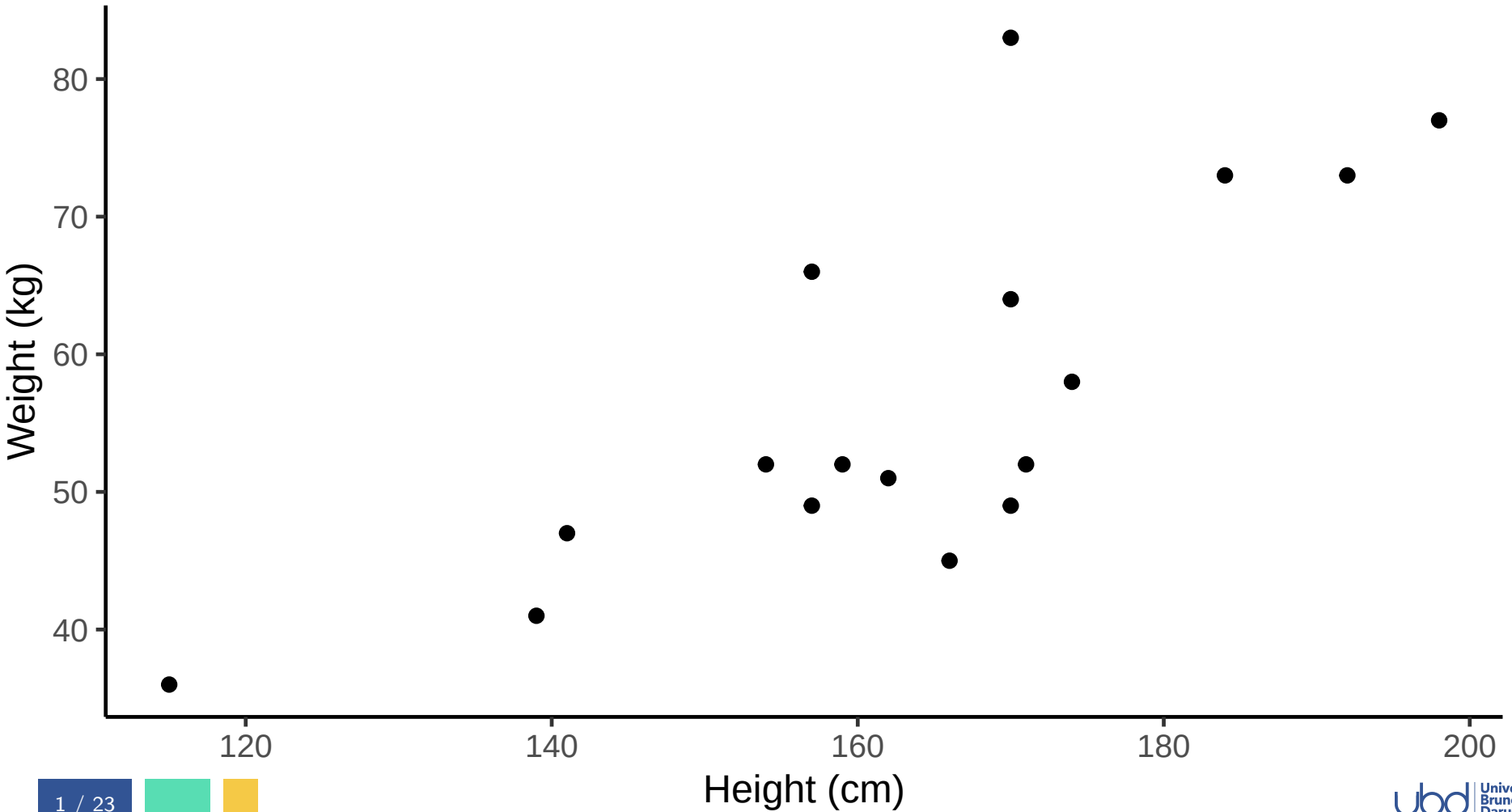
- Understand and quantify the linear association between two variables by calculating the correlation coefficient between them.
- Be familiar with the concept of least squares regression and know how to compute the least squares estimates for the intercept and slope for a simple linear regression model.
- Diagnose a linear model by analysing the residuals.
- Conduct a test of significance for the slope parameter of a linear regression model.
- Use the fitted model for statistical inference (interpreting the coefficients) and forecasting.

## Required reading

- Madsen (2016) Chapter 7.

# Linear associations

Relationship between height and weight of 17 boys in the Fitness Club survey. Is there a relationship between height ($X$) and weight ($Y$)?



# Quantifying the relationship

The correlation coefficient quantifies the *strength* of the linear association between two variables.

- It takes values between $-1$ (perfect negative) and $1$ (perfect positive).
- A value of 0 indicates no relationship whatsoever.
- Usually denoted by symbol $r$.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

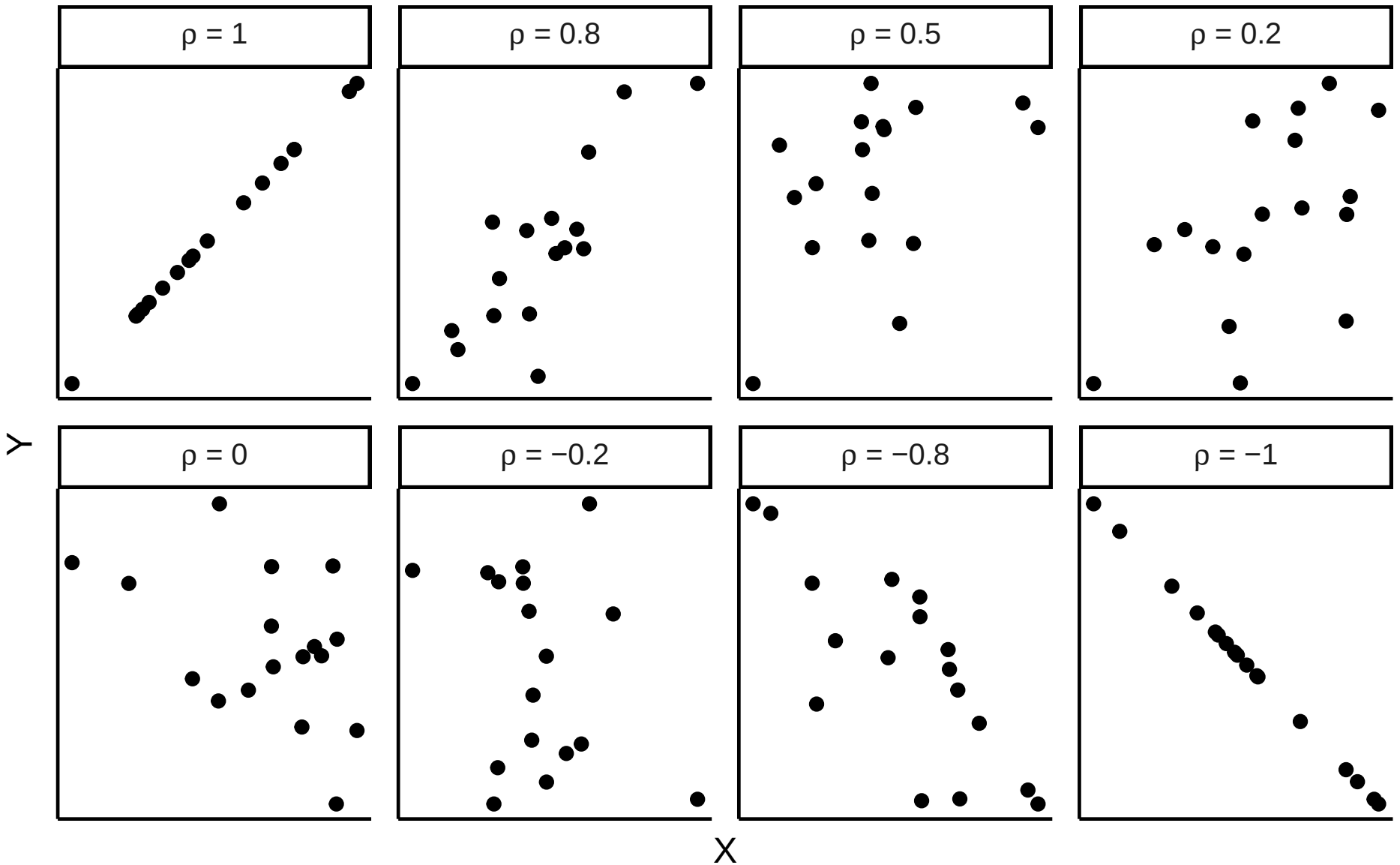
For the current data, the correlation is $r = 0.765$ (in Excel, use `CORREL`). This seems to suggest that a linear relationship is present.

# Example

| $i$ | $X$ | $Y$ | $X-\bar{X}$ | $(X-\bar{X})^2$ | $Y-\bar{Y}$ | $(Y-\bar{Y})^2$ | $(X-\bar{X})(Y-\bar{Y})$ |
|---|---|---|---|---|---|---|---|
| 1 | 157 | 66 | -6.47 | 41.9 | 9.06 | 82.06 | -58.62 |
| 2 | 115 | 36 | -48.47 | 2349.4 | -20.94 | 438.53 | 1015.03 |
| 3 | 174 | 58 | 10.53 | 110.9 | 1.06 | 1.12 | 11.15 |
| 4 | 171 | 52 | 7.53 | 56.7 | -4.94 | 24.41 | -37.20 |
| 5 | 141 | 47 | -22.47 | 504.9 | -9.94 | 98.83 | 223.38 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| | | SUM | | | | | |

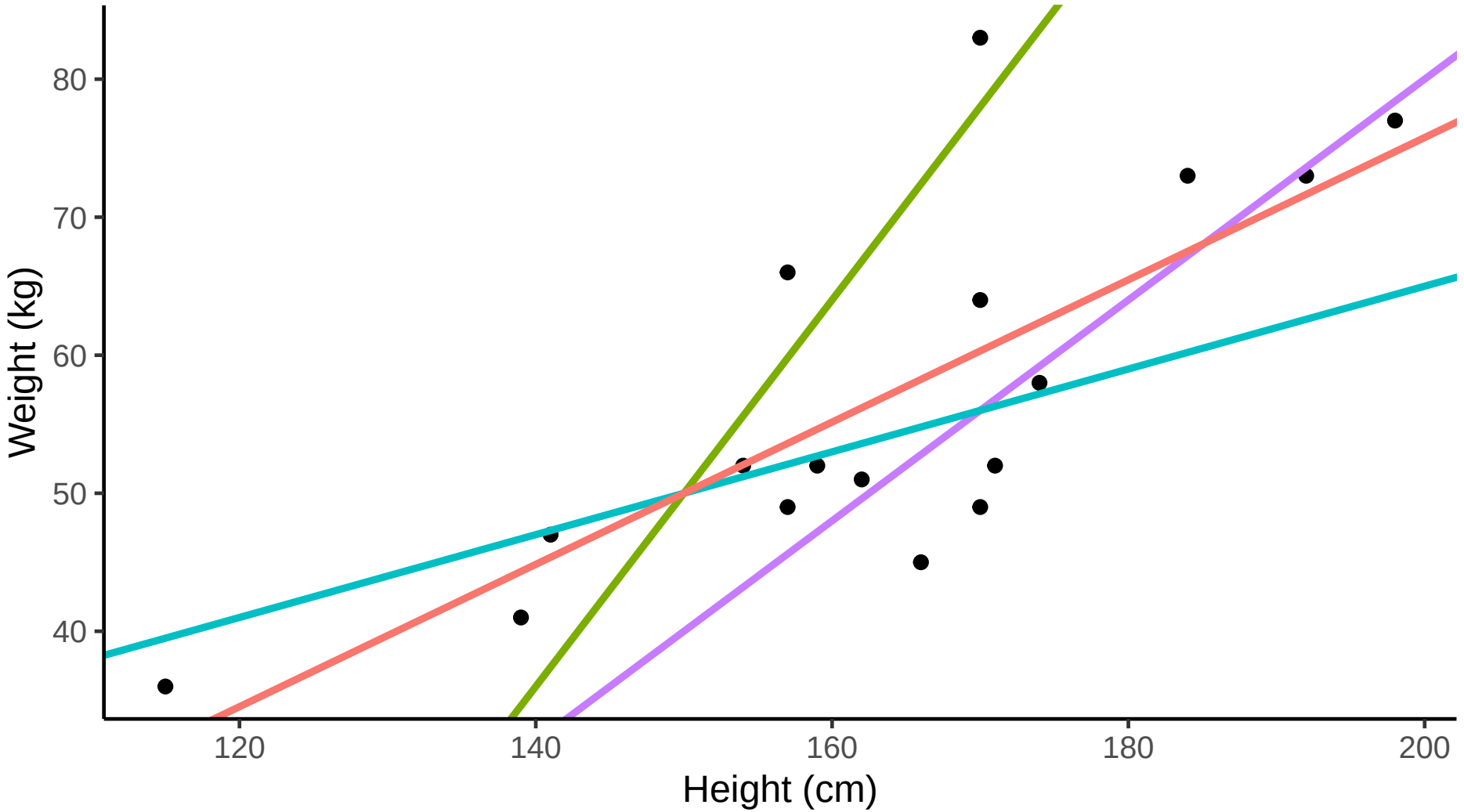
- $\sum X = 2779$, $\sum Y = 968$, $\bar{X} = 163.47$, $\bar{Y} = 56.94$
- $\sum (X-\bar{X})^2 = 6378.235$, $\sum (Y-\bar{Y})^2 = 2898.941$
- $\sum (X-\bar{X})(Y-\bar{Y}) = 3287.471$

# Correlations



# Eyeballing the line

Evidently, a linear trend exists. But there are many lines that can be drawn–which line passes "the best" through all the data points?



# Equation of a line

The equation of a line in two dimensions is

$$Y = \alpha + \beta X$$

Any straight line is "parameterised" by

- an intercept $\alpha$; and
- a slope $\beta$.

In the previous example, the $Y$ variable (response) is the weight of boys (in cm), whereas the $X$ variable (explanatory) is the height of boys (in kg).

Obviously, we don't expect the line to pass through all the data points due to some randomness in the data points.

## Definition 1 (Errors)

The *(random) errors* $\epsilon$ of the linear regression model is the amount by which an observation differs from its expected value

$$\epsilon = Y - (\alpha + \beta X)$$

# Least squares regression

Our model then becomes

$$Y = \alpha + \beta X + \epsilon$$

(Typically, we also assume that $\epsilon \sim \mathrm{N}(0, \sigma^2)$). So perhaps the "best line" to fit is the line which gives as small an error as possible.

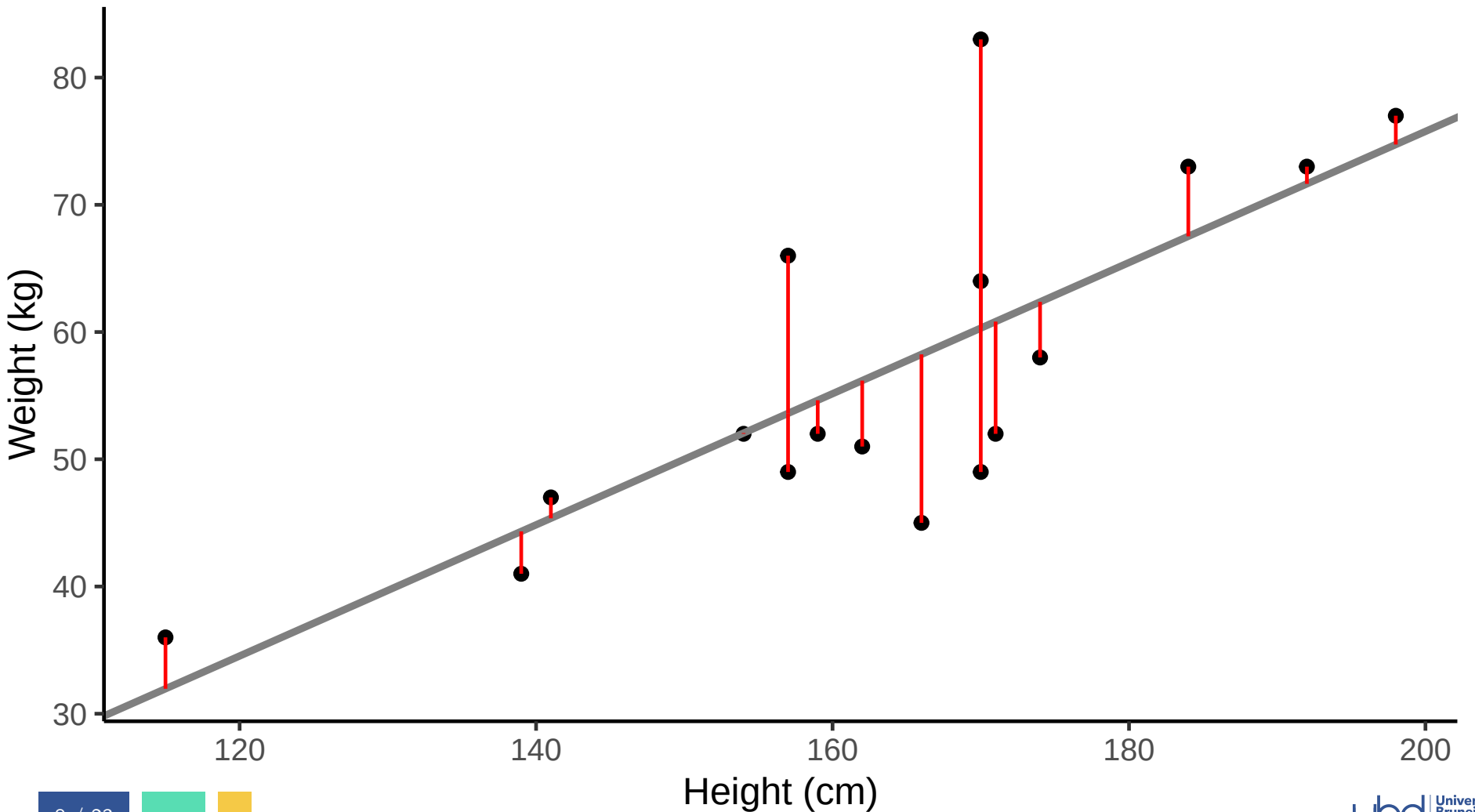
The method of least squares aims to find values $\alpha$ and $\beta$ which minimises the sum of squared errors

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

As a remark, errors are *random variables*, whose quantity is unknown beforehand. Once we have a straight line and observe the difference, these are known as *residuals* (often use hats $\hat{\epsilon}$ to distinguish).

# Least squares regression (cont.)

"Adjust" the regression line (intercept $\alpha$ and slope $\beta$) such that the sum of all the vertical red lines (errors) are as small as possible.



# Formula for least squares regression

For this simple linear model, the solution to the least squares minimisation problem is pretty straightforward.

## Proposition 2 (Least squares solutions)

For the linear regression model $Y = \alpha + \beta X + \epsilon$, where $\epsilon$ are random errors, the least squares solution $\hat{\alpha}$ and $\hat{\beta}$ are given by

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \qquad \text{and} \qquad \hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Remarks

- No need to memorise these formula.
- In Excel, use `INTERCEPT` and `SLOPE` to calculate $\hat{\alpha}$ and $\hat{\beta}$ respectively. (See the book for more details)

# $R^2$ value

Another useful determinant of how well the model captures the variability of the data is the coefficient of determination $R^2$.

## Definition 3 (Coefficient of determination)

For the simple linear regression model, let

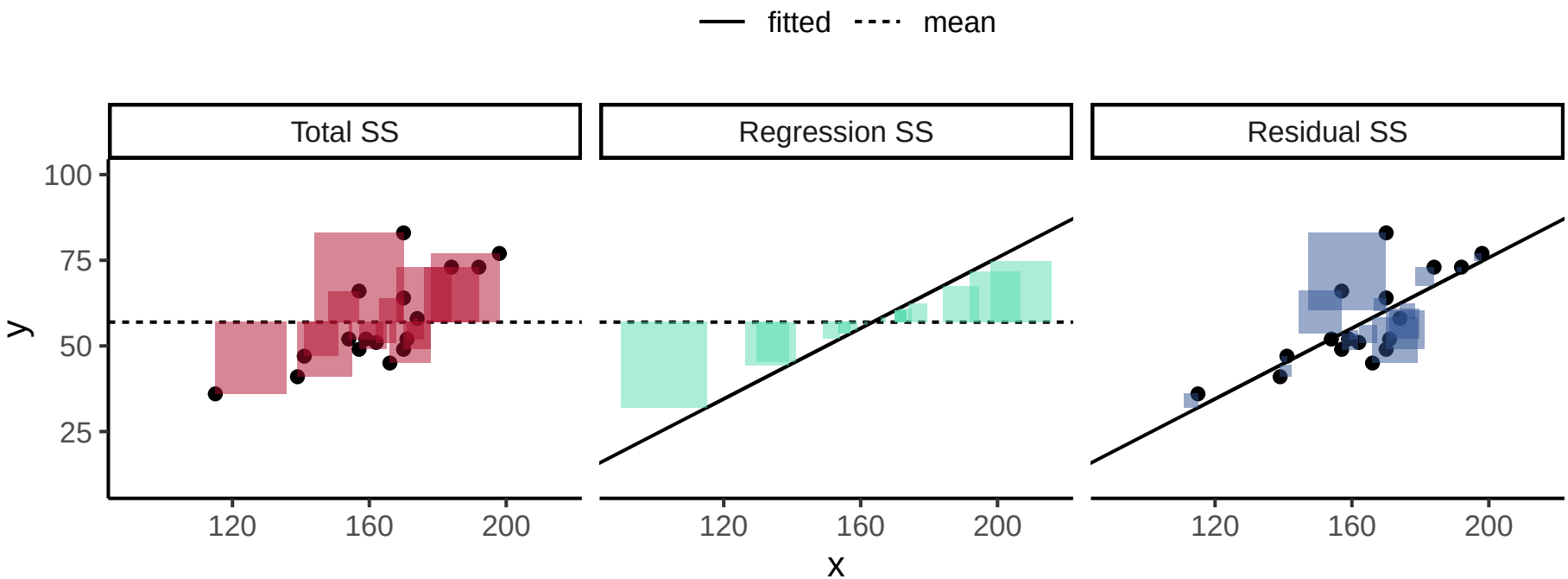
- $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$ represent the *fitted* values
- $\mathrm{SS}_{\mathrm{tot}} = \sum_{i=1}^{n} (y_i - \bar{y})^2$ represent the *total sum of squares*
- $\mathrm{SS}_{\mathrm{reg}} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ represent the *regression sum of squares*
- $\mathrm{SS}_{\mathrm{resid}} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ represent the *residual sum of squares*

Then, the coefficient of determination $R^2 \in [0, 1]$ is

$$R^2 = \frac{\mathrm{SS}_{\mathrm{reg}}}{\mathrm{SS}_{\mathrm{tot}}} = 1 - \frac{\mathrm{SS}_{\mathrm{resid}}}{\mathrm{SS}_{\mathrm{tot}}}$$

The $R^2$ measures how much the variation in the data is explained by the model. Values closer to 1 are desirable.

# $R^2$ value (cont.)




- $\text{SS}_{\text{tot}} = \sum_{i=1}^{n} (y_i - \bar{y})^2$
- $\text{SS}_{\text{reg}} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$
- $\text{SS}_{\text{resid}} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

$$R^2 = \frac{\text{SS}_{\text{reg}}}{\text{SS}_{\text{tot}}} = 1 - \frac{\text{SS}_{\text{resid}}}{\text{SS}_{\text{tot}}}$$

# Residuals

## Definition 4 (Residuals)

The residuals are the difference between the the observed value and the fitted values (values on the straight line), i.e.

$$\hat{\epsilon}_i = y_i - \hat{y}_i = y_i - (\hat{\alpha} + \hat{\beta} x_i)$$

We expect that error values

1. Follow a bell-shaped normal distribution;
2. Are not too extreme-valued (no outliers); and
3. Are truly random and show no discernible pattern with any other variable.

Analysing the residuals can help diagnose inconsistencies of linear model assumptions in the data set.

# Calculating residuals

For the Fitness Club example, we may calculate the least squares estimators

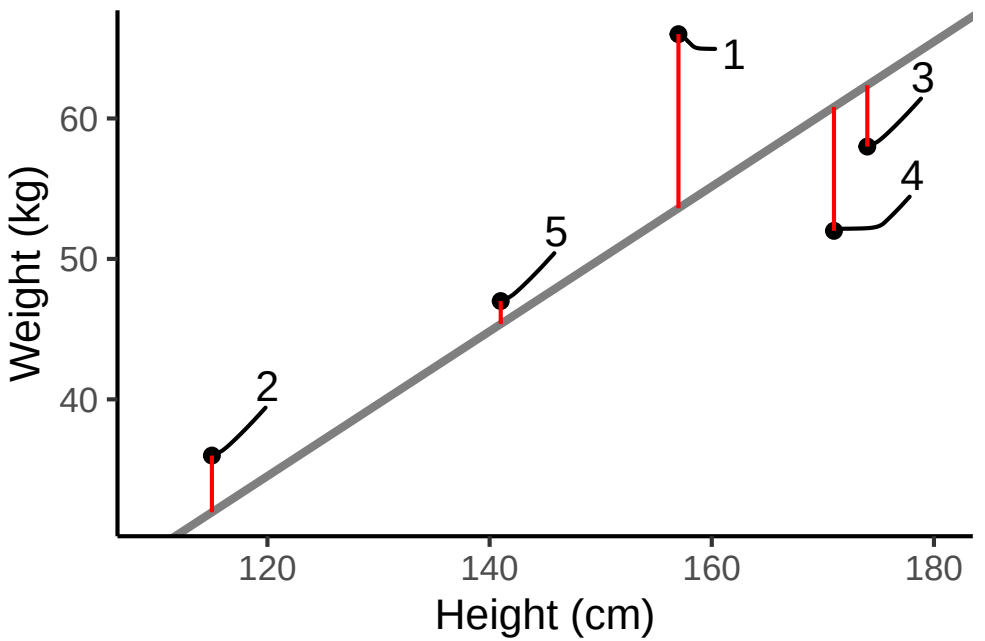
- $\hat{\alpha} = -27.31$
- $\hat{\beta} = 0.52$

Thus, for each explanatory variable $x_i$, we may calculate the *fitted value* as

$$\hat{y}_i = -27.31 + 0.52 \times x_i,$$

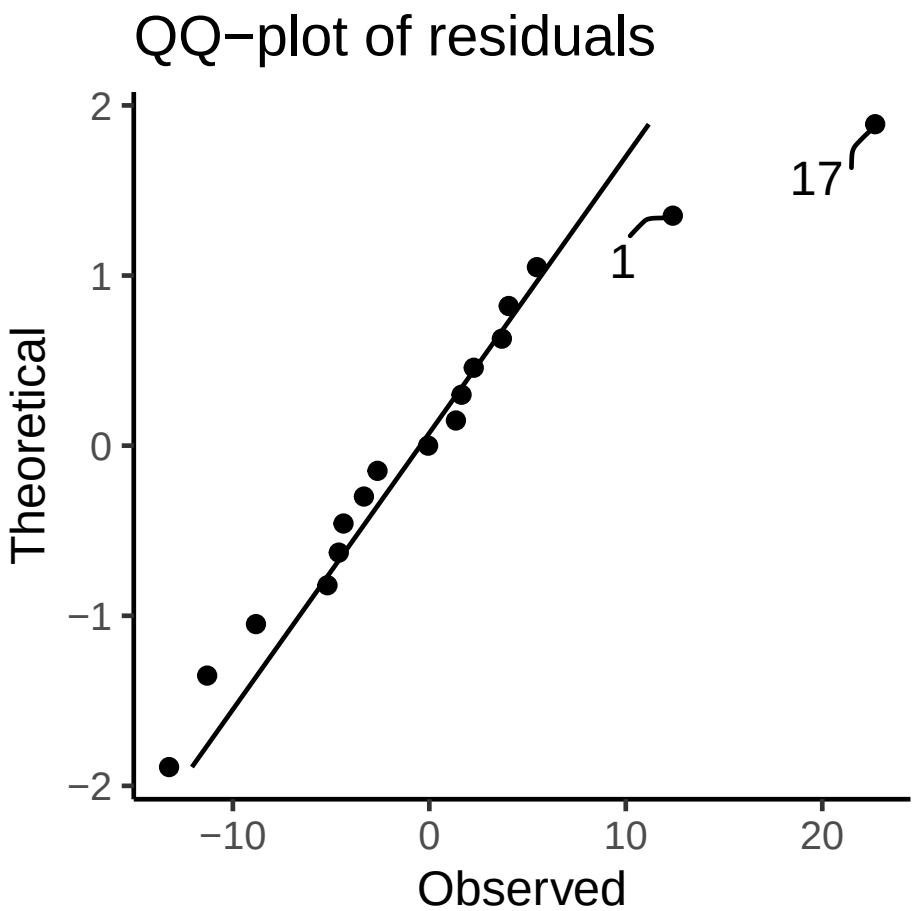
and the residual as $\hat{\epsilon}_i = y_i - \hat{y}_i$.

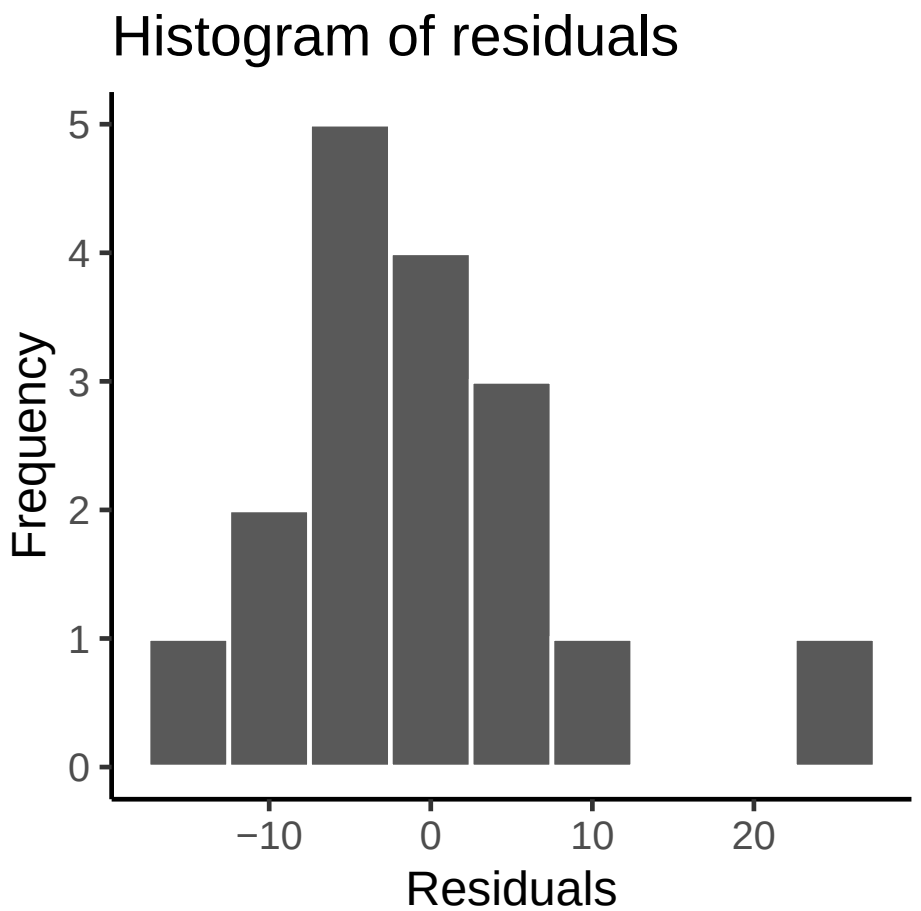
| $i$ | $X$ | $Y$ | $\hat{Y}$ | $\hat{\epsilon}$ |
|---|---|---|---|---|
| 1 | 157 | 66 | 53.6 | 12.39 |
| 2 | 115 | 36 | 32.0 | 4.04 |
| 3 | 174 | 58 | 62.4 | -4.37 |
| 4 | 171 | 52 | 60.8 | -8.82 |
| 5 | 141 | 47 | 45.4 | 1.64 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |



# Checking residuals

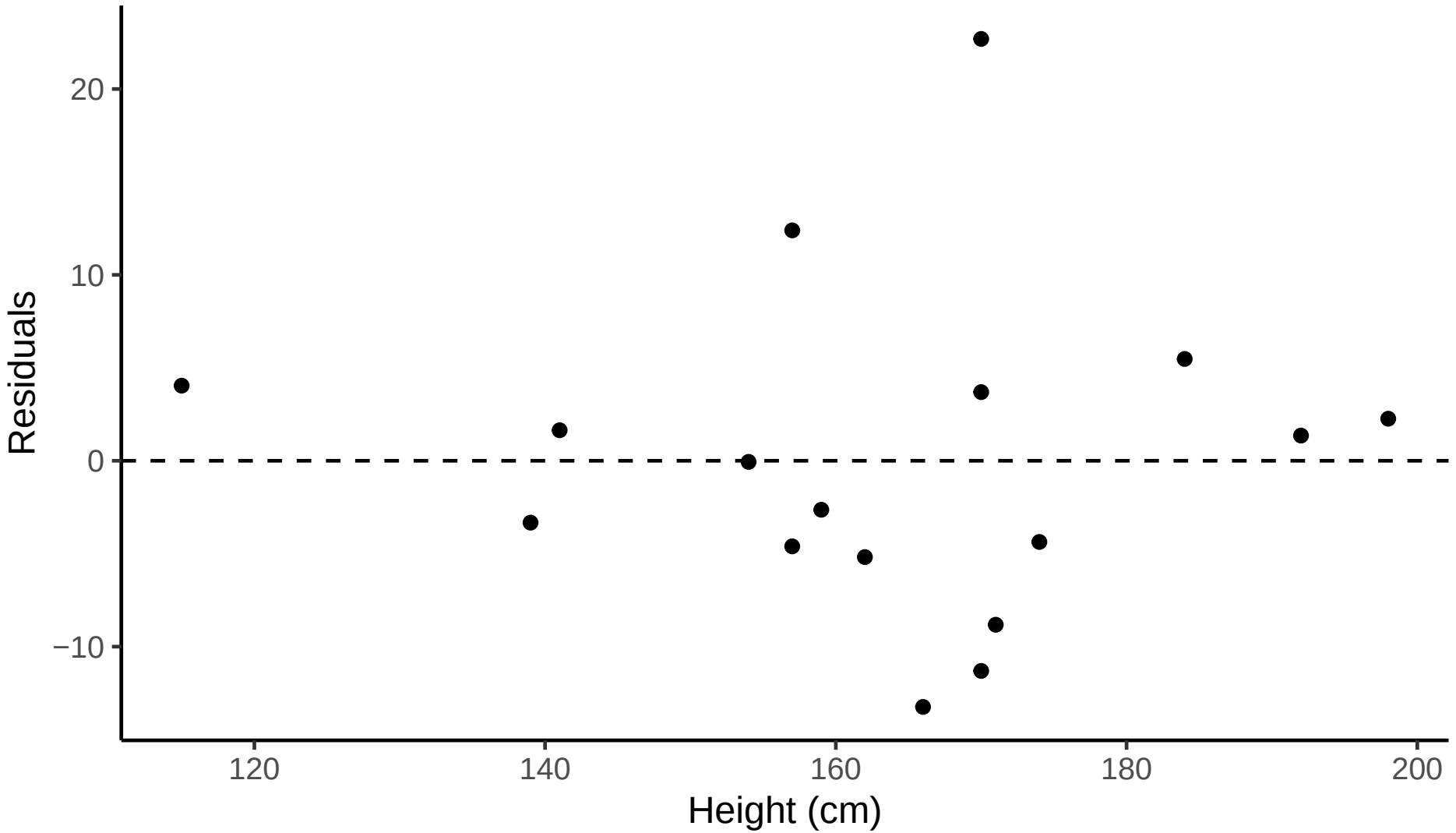
| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 12.39 | 4.04 | -4.37 | -8.82 | 1.64 | -13.24 | -5.18 | -4.61 | -3.33 |
| -2.64 | -11.31 | 2.26 | 1.35 | -0.06 | 3.69 | 5.48 | 22.69 | |



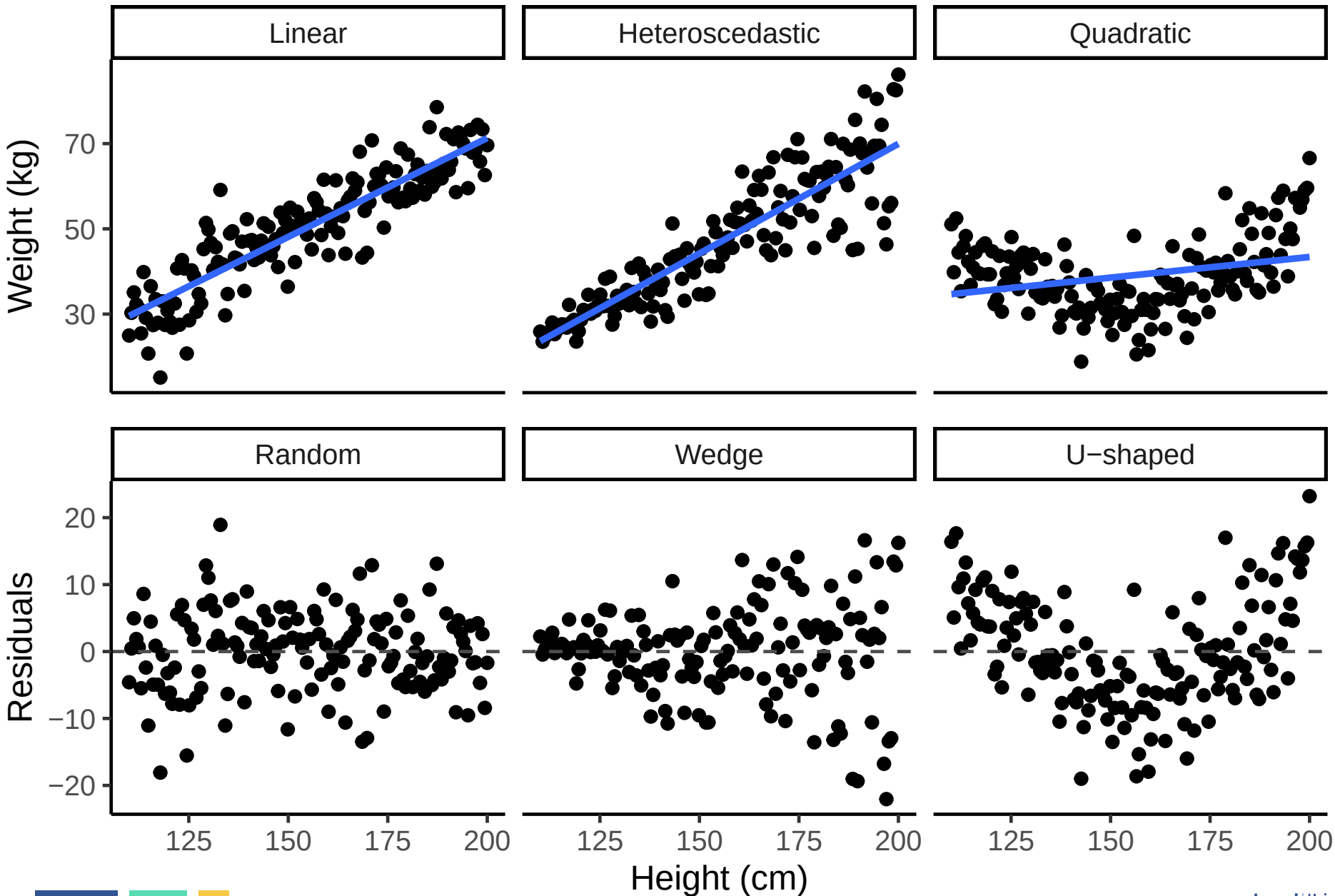



# Checking residuals (cont.)

Plot $\hat{\epsilon}$ against $x$; ensure that only a random pattern is seen.



# Homoscedasticity



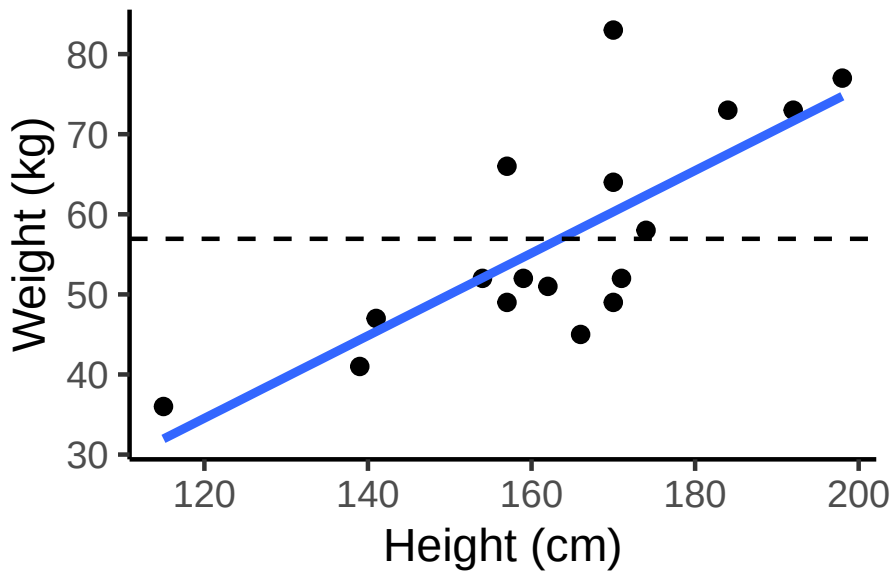
# Is there a relationship?

Can we be sure that the linear relationship is "real"? Is it any different, say, from just a horizontal line at the average?

$$H_0 : Y = \alpha \qquad \text{v.s.} \qquad H_1 : Y = \alpha + \beta X$$

Using the general approach we learnt last chapter, we can test the hypothesis.

1. Assume that the hypothesis is true (i.e. $\beta = 0$).
2. Calculate the $p$-value, i.e. the probability of observing the data or "more extreme cases" of it.
3. If $p < 0.05$, reject the (null) hypothesis; if not, we accept it.



# Test statistic

## Definition 5 (Test statistic for slope)

To test the hypothesis that $H_0 : \beta = 0$ in the simple linear regression, calculate

$$T = r\sqrt{\frac{n-2}{1-r^2}},$$

where $r$ is the correlation coefficient between $X$ and $Y$. This value is then compared against critical values of the $t_{n-2}$ distribution.

## Example 6

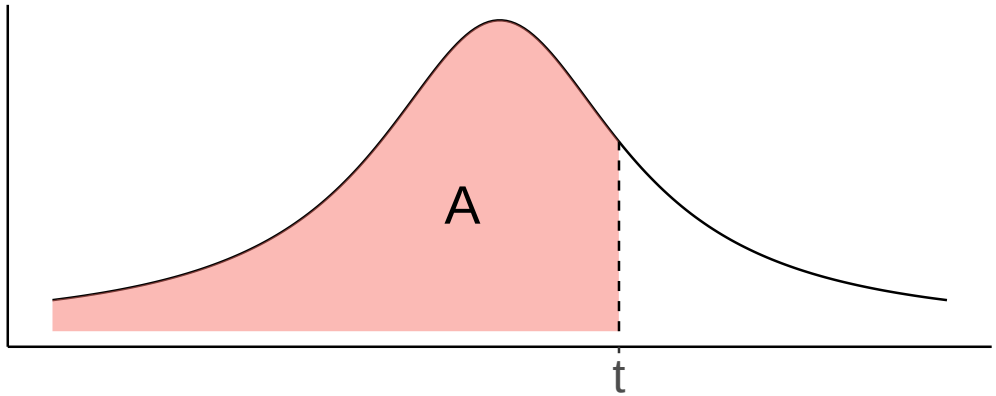
Earlier we computed $r = 0.765$ for our data set ($n = 17$). Thus,

$$T = 0.765 \times \sqrt{\frac{17-2}{1-0.765^2}} = 4.594.$$

From the tables: $\Pr(t_{15} > 1.753) = 0.05$, so $\Pr(t_{15} > 4.594) \ll 0.05$.

# The $t$-distribution

Each table entry is the value $t$, where $\int_{-\infty}^{t} f(x)\,dx = A$ with $X \sim t_k$.




| | $A$ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $k$ | 0.80 | 0.85 | 0.90 | 0.95 | 0.975 | 0.990 | 0.9925 | 0.9950 | 0.9975 |
| 11 | 0.876 | 1.088 | 1.363 | 1.796 | 2.201 | 2.718 | 2.879 | 3.106 | 3.497 |
| 12 | 0.873 | 1.083 | 1.356 | 1.782 | 2.179 | 2.681 | 2.836 | 3.055 | 3.428 |
| 13 | 0.870 | 1.079 | 1.350 | 1.771 | 2.160 | 2.650 | 2.801 | 3.012 | 3.372 |
| 14 | 0.868 | 1.076 | 1.345 | 1.761 | 2.145 | 2.624 | 2.771 | 2.977 | 3.326 |
| 15 | 0.866 | 1.074 | 1.341 | 1.753 | 2.131 | 2.602 | 2.746 | 2.947 | 3.286 |

# Interpreting the model

For our simple linear regression model,

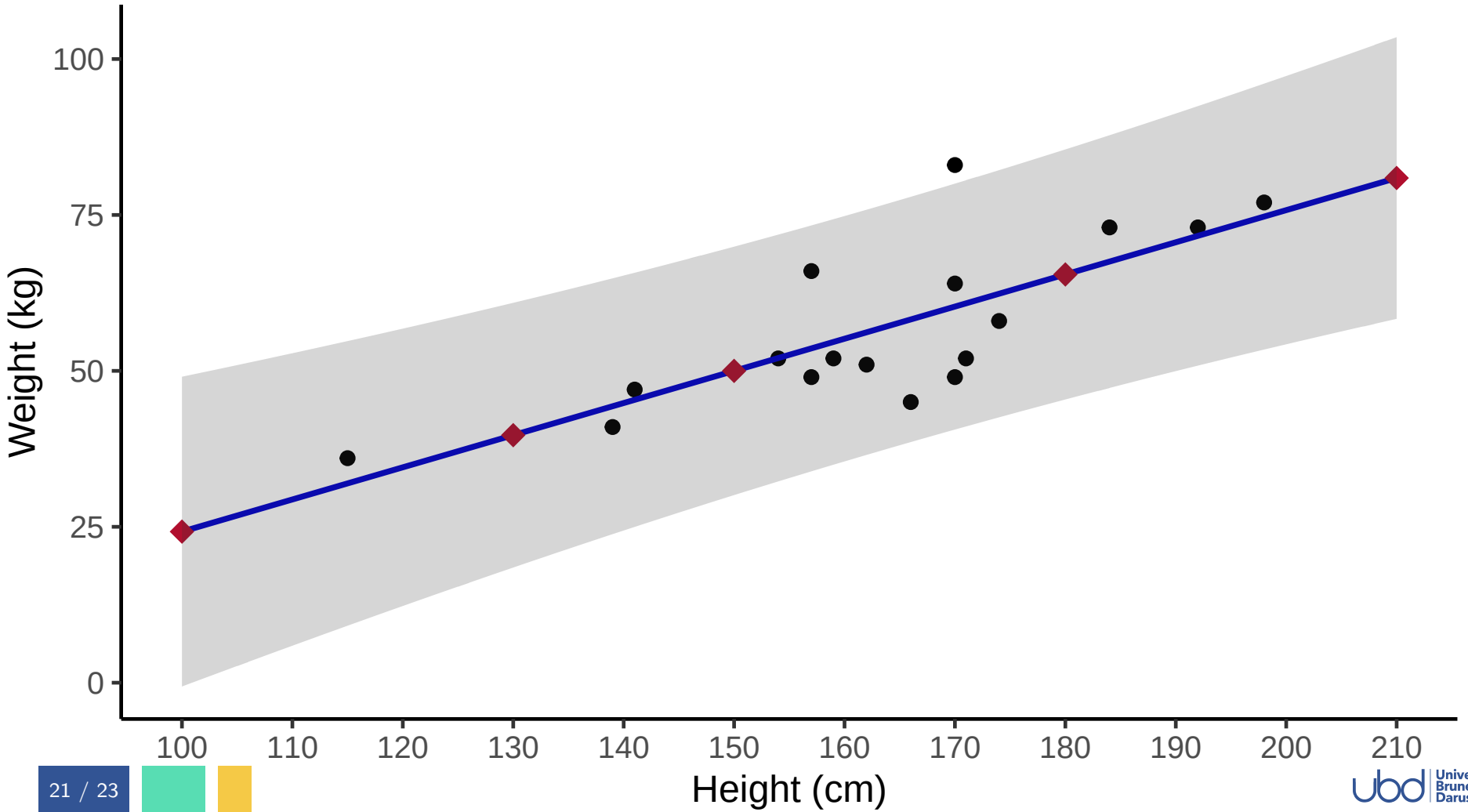
- Suppose that the value of $X = 0$. Then $Y = \alpha + \beta X = \alpha$. Hence $\alpha$ is the expected value of $Y$ when the explanatory variable is zero.
- Consider two values

$$Y = \alpha + \beta X$$
$$Y' = \alpha + \beta(X + 1)$$

If we subtract $Y' - Y$ we get the value $\beta$. Hence, the slopes give the (average) increase in the response variable given a unit change in the explanatory variable.

# Forecasting

Since we have the estimated regression line $Y = -27.31 + 0.52X$, we can plug in any value for $X$, even new data points for predictive purposes.



# Forecasting (cont.)

When forecasting a new value $y_*$ given input $x_*$, bear in mind the following.

- Forecasting works when $x_*$ is close to the observed data $\{x_1, \ldots, x_n\}$. We cannot expect it to work well for data points that are very far to the left or right.
- Every prediction comes with some uncertainty. While this is beyond the scope of this module, if interested, the uncertainty is captured by the $100(1-\alpha)\%$ interval

$$\hat{y}_* \pm t_{n-2}(\alpha/2) \times \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}},$$

where $\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^n (y - \hat{y}_i)^2}{n-2}}$ is called the *residual standard error*, and $t_{n-2}(\alpha/2)$ is the top $\alpha/2$th point[1] of the $t_{n-2}$ distribution.

[1] E.g. for a 95% interval with $n = 17$, then $t_{n-2}(\alpha/2) = t_{15}(0.025) = 2.131$.

# Further topics

- Our explanatory variable need not be continuous. In fact, they can be categorical. For example,

$$\texttt{Weight} = \alpha + \beta\ \texttt{WtTrain}$$

  I.e., modelling the weight of individuals by whether or not they do weight training ("Yes"/"No").
- We can build a model using more than one explanatory variables, e.g.

$$\texttt{Weight} = \alpha + \beta_1\texttt{Height} + \beta_2\texttt{WtTrain} + \beta_3\texttt{Sex} + \beta_4\texttt{Age}$$

  This can easily be done in most modern statistical softwares including MS Excel. Note that the formulae presented in this lecture is exclusively for simple models ($Y$ and $X$ variables only).
- Binary explanatory variables (aka logistic regression).