Math 243

Least Squares Regression Line – Inv. 5.8 and 5.9

Prediction

So far, we've

- *described* a dataset through graphs and numerical summaries,
- *tested* whether a parameter is a value (H_0 vs. H_a), and
- *estimated* a parameter (95% confidence interval)

Today, we'll

- *predict* the value of a quantitative variable based on the value of a second quantitative variable.

Inv. 5.8: Footlength vs. Height

Given the length of a person’s **foot**, *predict* their **height**.

Collect data:

Inv. 5.8: Footlength vs. Height

Given the length of a person's **foot**, *predict* their **height**.

Collect data:

Observational units: 20 statistics students

Explanatory variable: foot length in cm

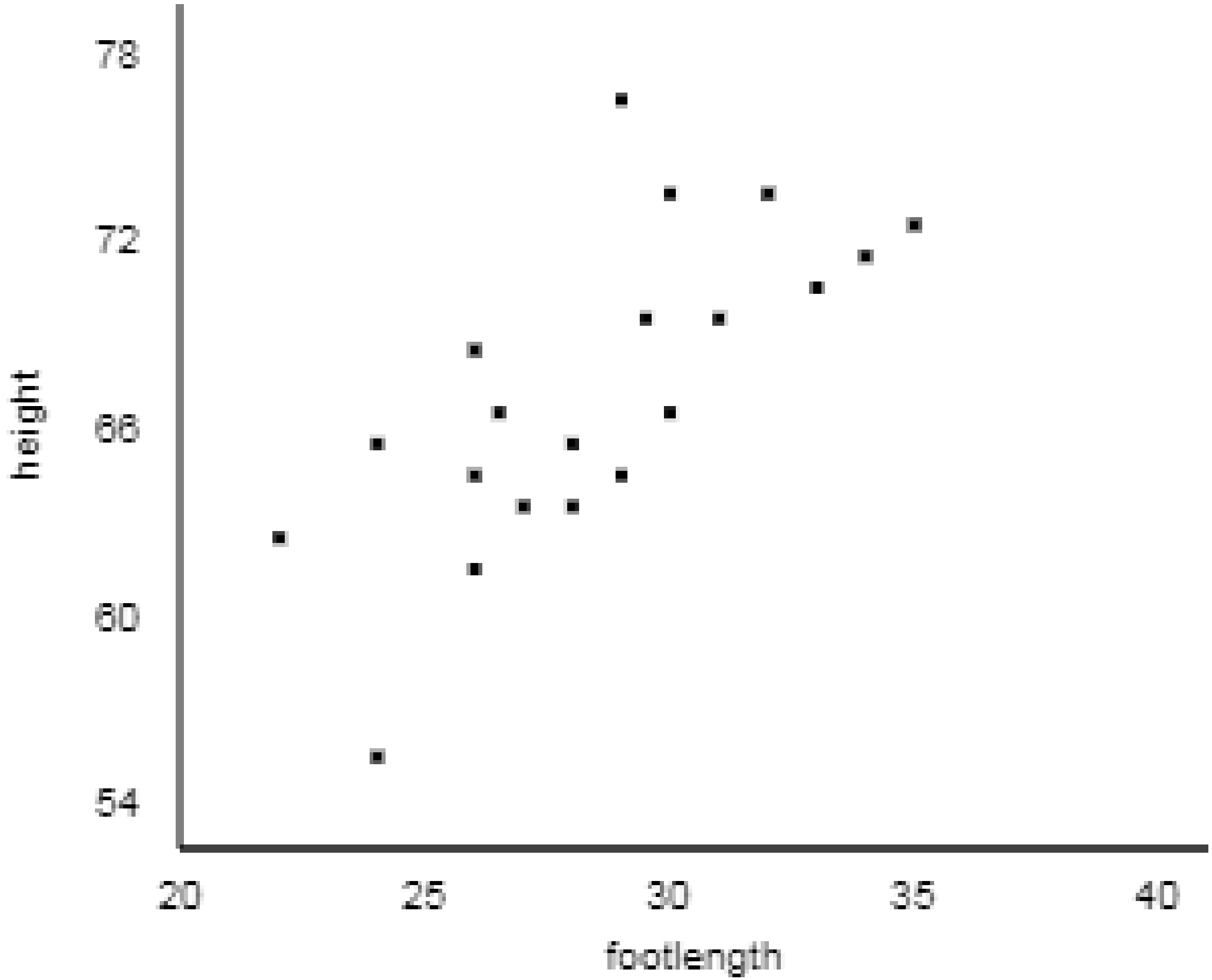
Response variable: height in inches

Inv. 5.8: Footlength vs. Height

Direction?

Linear?

Strength?

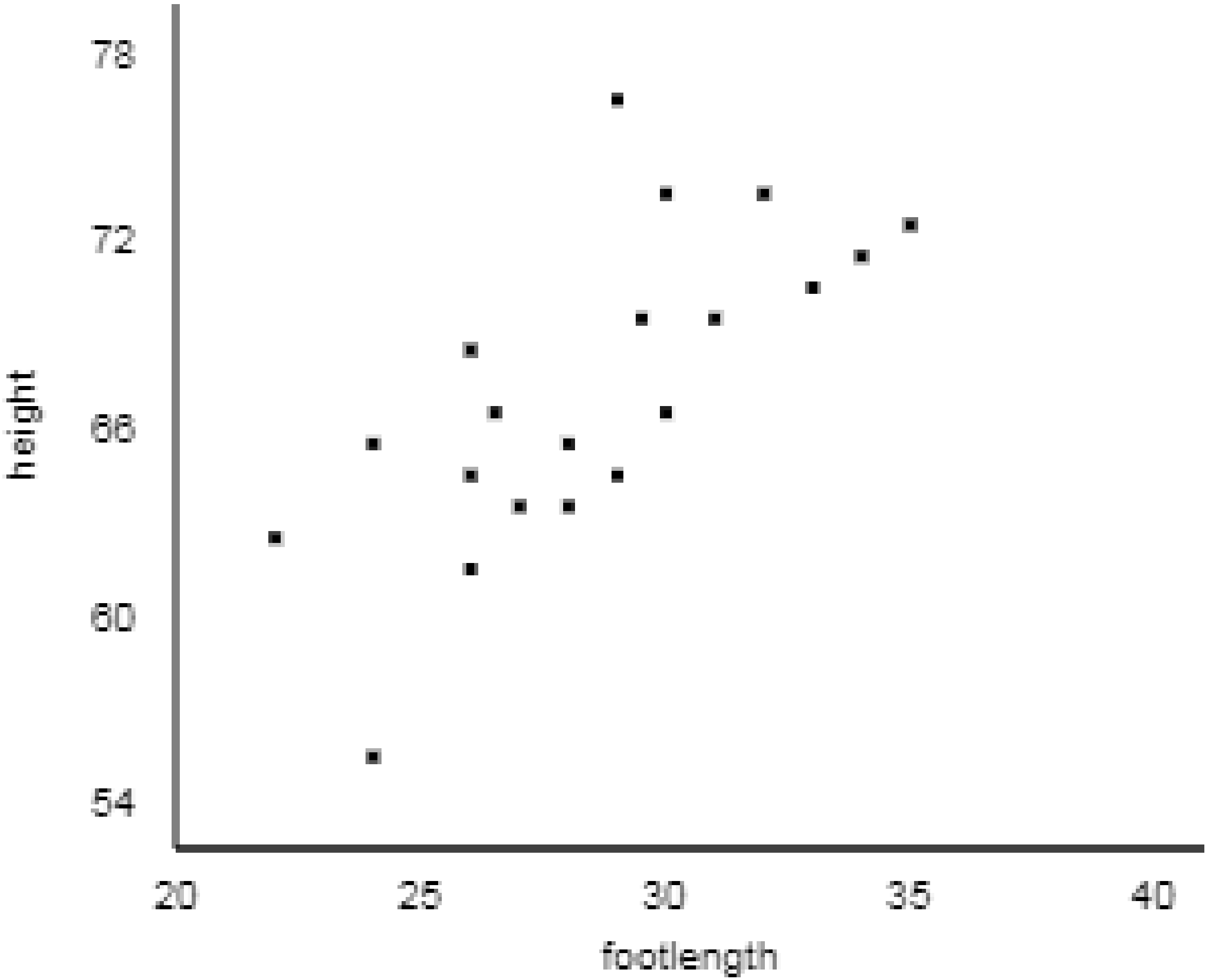


Inv. 5.8: Footlength vs. Height

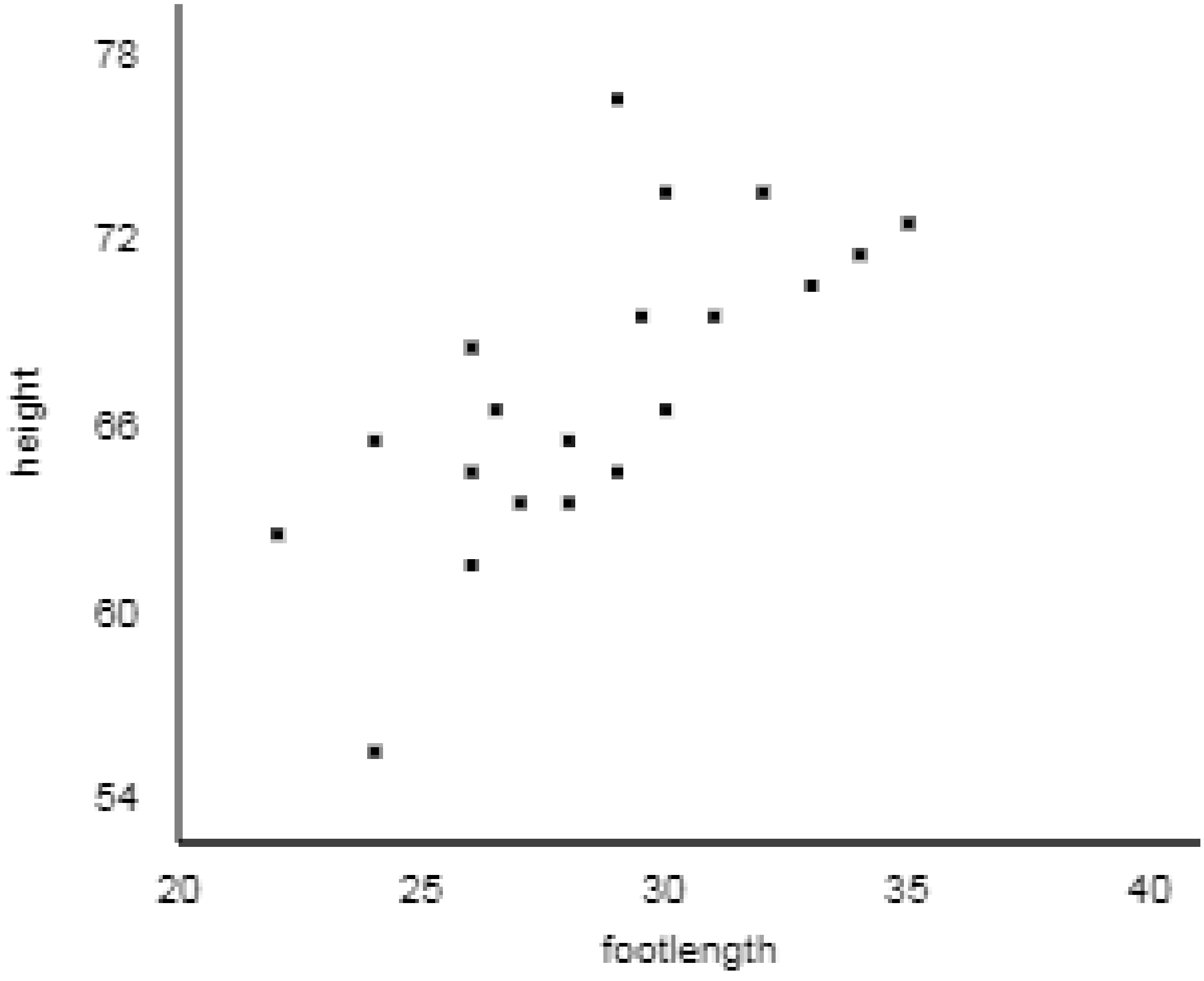
Direction?
positive

Linear?
Yes

Strength?
r =0.711

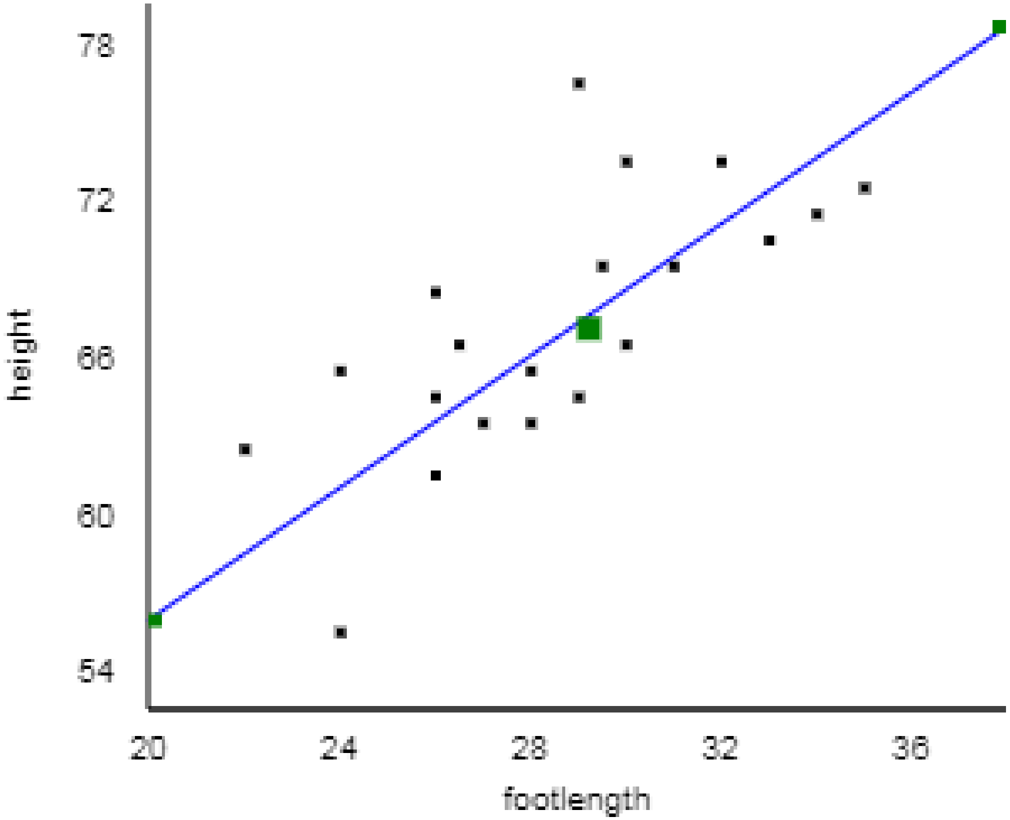


Suppose you come across a footprint that is 28 cm long. How tall do you predict the maker was?



Prediction from a line

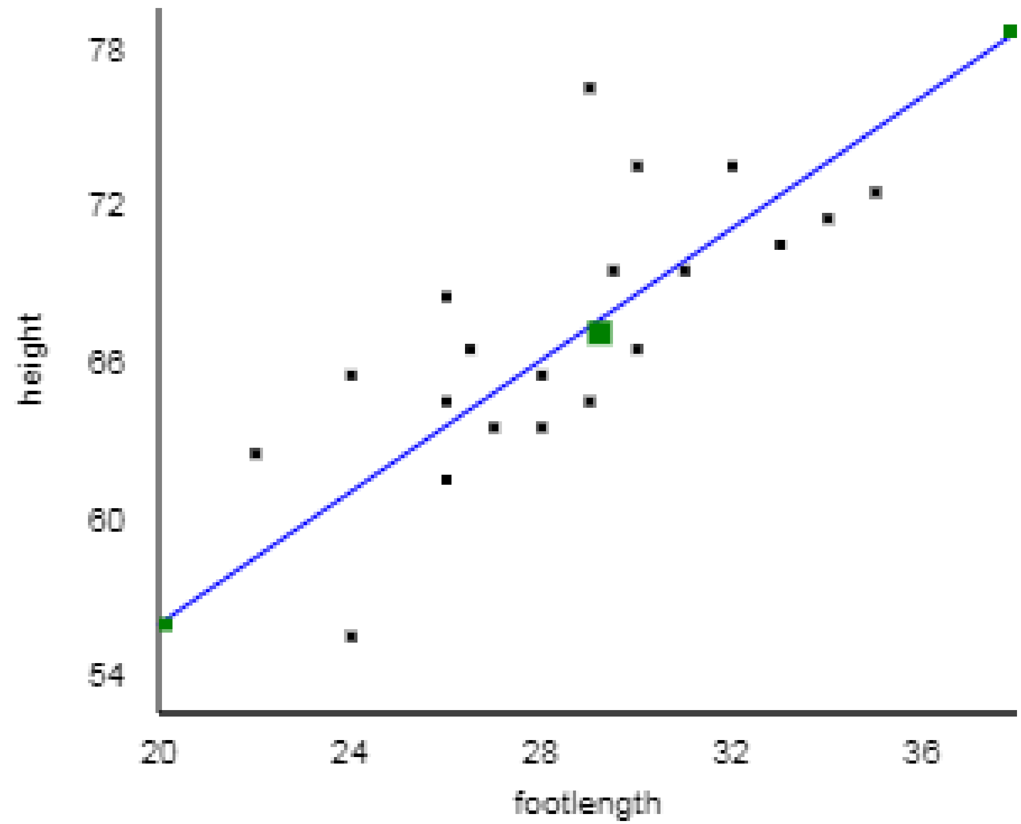
By drawing a **line** through the points, I can consistently predict the value of the response variable for a given value of the explanatory variable.



Prediction from a line

By agreeing on ***how to draw*** a **line** through the points, **we** can consistently predict the value of the response variable for a given value of the explanatory variable.

<u>Idea:</u> Choose the line that minimizes the distances from the points to the prediction line



In HW 9, you'll be asked to find the equation of the line you would use.

Analyzing Two Quantitative Variables

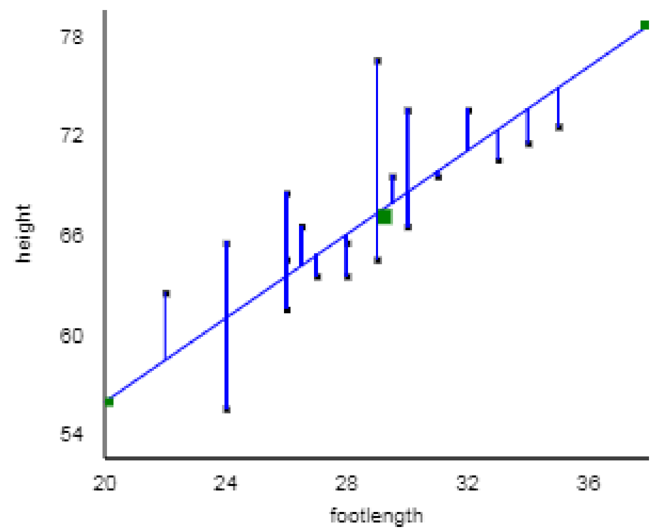
Sample data:

(Explanatory, Response)

26.0	62
26.5	67
28.0	66
28.0	64
26.0	69
35.0	73
30.0	74
31.0	70
29.0	65
34.0	72

Use Data | Revert | Clear

n = 20



Show Movable Line: ☑

height^ = 31.18 + 1.26 x footlength

Show Regression Line: ☐

Show Residuals: ☑

SAE=58.24

Show Squared Residuals: ☐

Some terminology for choosing the "best" line

A **residual** is the difference between the *predicted* value and the *observed* value

The *sum of the absolute residuals* is denoted **SAE**

The sum of squared residuals is denoted **SSE**

Choosing the “best” line

- Could choose the line that minimizes either SAE or SSE.

- Historically, people have chosen the line that minimizes SSE because it is possible to compute without computers: this line is called the “**least squares line**” or “**regression line**”

The Least Squares Line

Analyzing Two Quantitative Variables

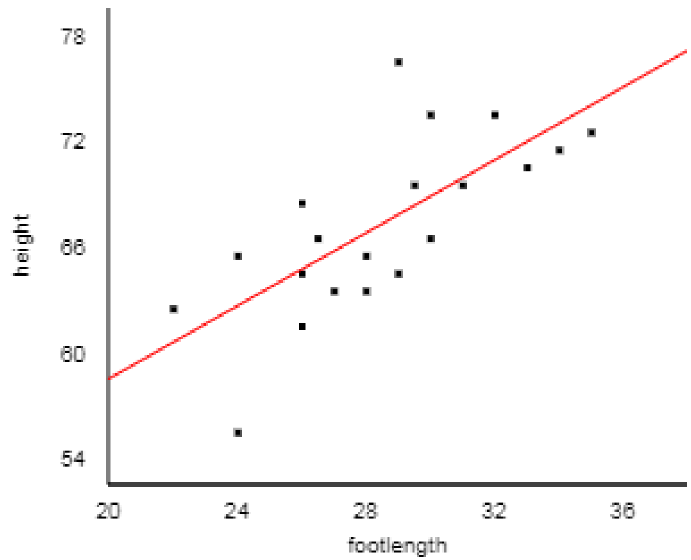
Sample data:

(Explanatory, Response)

26.0	62
26.5	67
28.0	66
28.0	64
26.0	69
35.0	73
30.0	74
31.0	70
29.0	65
34.0	72

Use Data Revert Clear

n = 20



Show Movable Line: ☐

Show Regression Line: ☑

height^ = 38.30 + 1.03 x *footlength*

Equation of a line

Recall the equation of a line from algebra:

$$y=mx+b$$

Example: $y=3x+1$

What does the 3 mean?

What does the 1 mean?

Equation of a line

Recall the equation of a line from algebra:

$$y=mx+b$$

Example: $y=3x+1$

What does the 3 mean?

If x increases by 1 unit then y increases by 3 units

What does the 1 mean?

When $x = 0$, $y = 1$.

Equation of a line

Recall the equation of a line from algebra:

$$y=mx+b$$

Example: $y=-3x+1$

What does the -3 mean?

If x increases by 1 unit then y decreases by 3 units

Equation of a "least squares line"

$$\hat{y} = b_0 + b_1 x$$

Here,

- $\hat{y}$ is the **predicted value** of the response variable when the value of the explanatory variable is x

- b_1 is the **regression slope**,

- b_0 is the **regression intercept**

The “least squares line” from Inv. 5.8

$$\widehat{height} = 38.1 + 1.03 footlength$$

Here,

- $\widehat{height}$ is the **predicted** height for a given footlength
- 1.03 is the **regression slope,**
- 38.1 is the **regression intercept**

Interpreting the “least squares line” from Inv. 5.8

$$\widehat{height} = 38.3 + 1.03 footlength$$

Slope:

If the footlength increases by 1 cm then the predicted height increases by 1.03 inches.

Intercept:

The predicted height is 38.3 inches when the footlength is 0.

Using the "least squares line" from Inv. 5.8

$$\widehat{height} = 38.3 + 1.03 footlength$$

Predict the height of someone whose footlength is 28 cm:

$$\widehat{height} = 38.3 + 1.03(28) = 67.14$$

Coefficient of determination, R^2

- Provides a measure of how useful the least squares line is

R^2 = percent error of the SSE of prediction line $\bar{y}$ and the SSE of the least squares line

where SSE is the sum of the squared residuals

Interpretation of R^2 in Inv. 5.8

R^2 = percent of variability of the response variable y that is ***explained*** by the least squares line with the explanatory variable x

R^2 =50.6% so…

The least square line with footlength ***explains*** 50.6% of the variability in height.

Interpretation of R^2

R^2 = percent of variability of the response variable y that is explained by the least squares line with the explanatory variable x

Notes:

- R^2 is always between 0% and 100%
- $R^2 = r^2$, where r is the correlation coefficient.