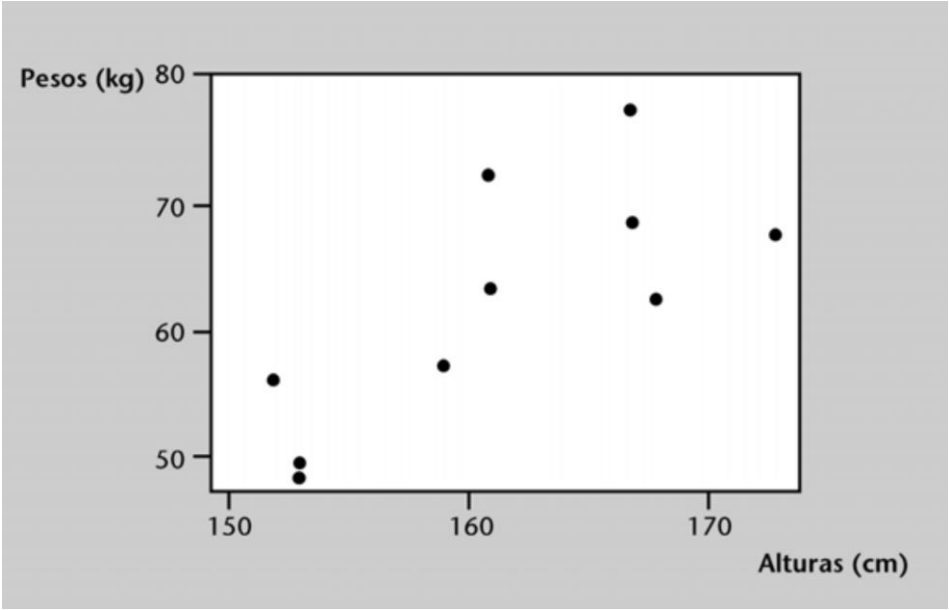
Linear Regression - Example

Department of Computer Languages and Systems

Data

Individuo	1	2	3	4	5	6	7	8	9	10
X altura (cm)	161	152	167	153	161	168	167	153	159	173
Y peso (kg)	63	56	77	49	72	62	68	48	57	67



Calculation of β_0 and β_1

i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	161	63	–0,4	1,1	0,16	–0,44
2	152	56	–9,4	–5,9	88,36	55,46
3	167	77	5,6	15,1	31,36	84,56
4	153	49	–8,4	–12,9	70,56	108,36
5	161	72	–0,4	10,1	0,16	–4,04
6	168	62	6,6	0,1	43,56	0,66
7	167	68	5,6	6,1	31,36	34,16
8	153	48	–8,4	–13,9	70,56	116,76
9	159	57	–2,4	–4,9	5,76	11,76
10	173	67	11,6	5,1	134,56	59,16
Σ	**1.614**	**619**			**476,40**	**466,40**

Calculation of β_0 and β_1 (cont.)

Sample means: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 161.4$ $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 61.9$

Sample variance: $S_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{476.40}{10-1} = 52.933$

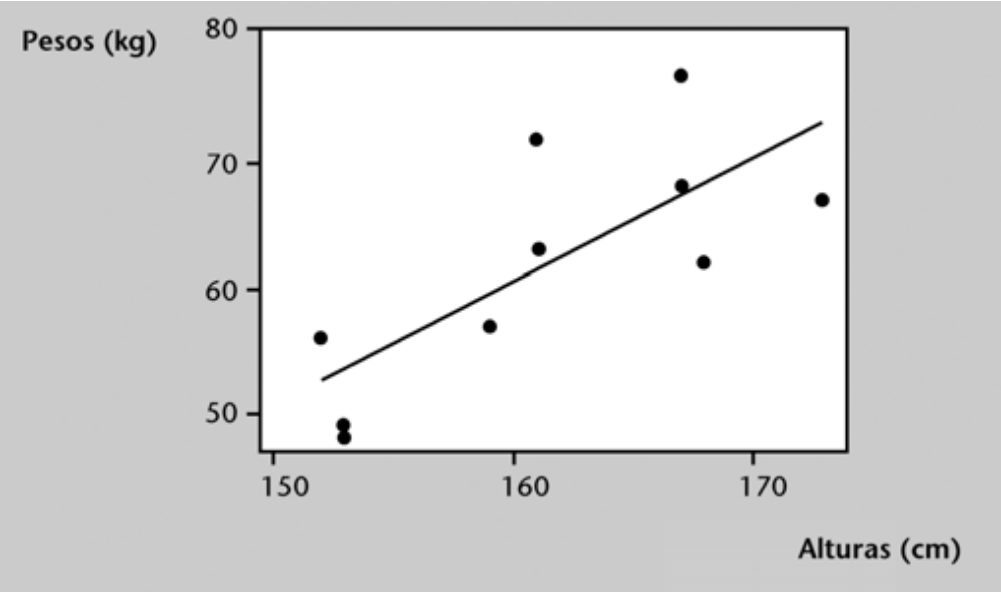
Sample covariance: $S_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{466.40}{10-1} = 51.822$

$$\widehat{\beta_1} = \frac{S_{xy}}{S_x^2} = \frac{51.822}{52.933} = 0.979$$

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x} = 61.9 - 0.979009 \cdot 161.4 = -96.112$$

Regression line

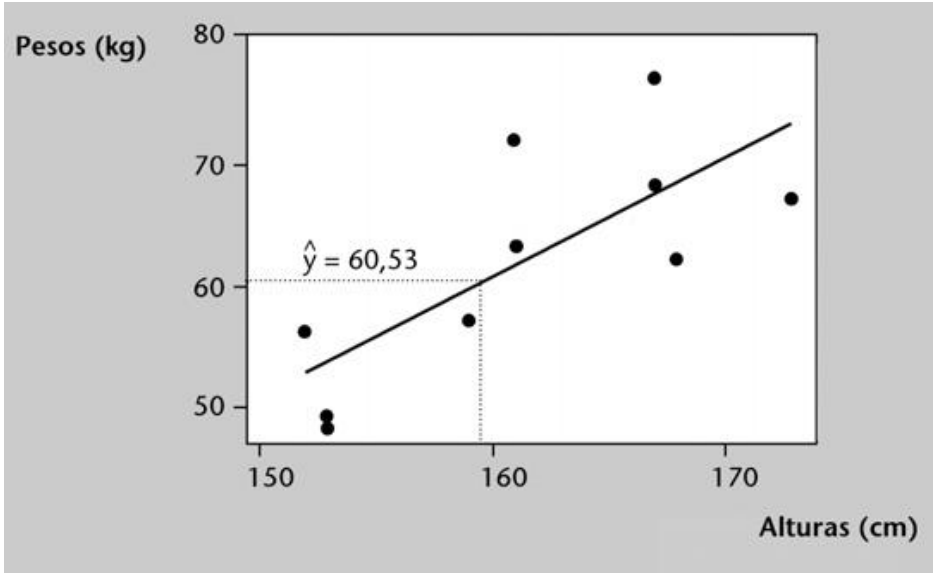
$$\hat{y} = -96.112 + 0.979x$$



Making a prediction

Estimate the weight value for a person 160 cm tall:

$$\hat{y} = -96.112 + 0.979x = -96.112 + 0.979 \cdot 160 = 60.53$$



Calculation of R^2

i	x_i	y_i	$\hat{y}_i$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$\hat{y}_i - \bar{y}$	$(\hat{y}_i - \bar{y})^2$	ε_i	ε_i^2
1	161	63	61,51	1,10	1,21	–0,39	0,15	1,49	2,23
2	152	56	52,70	–5,90	34,81	–9,20	84,69	3,30	10,91
3	167	77	67,38	15,10	228,01	5,48	30,06	9,62	92,50
4	153	49	53,68	–12,90	166,41	–8,22	67,63	–4,68	21,87
5	161	72	61,51	10,10	102,01	–0,39	0,15	10,49	110,07
6	168	62	68,36	0,10	0,01	6,46	41,75	–6,36	40,47
7	167	68	67,38	6,10	37,21	5,48	30,06	0,62	0,38
8	153	48	53,68	–13,90	193,21	–8,22	67,63	–5,68	32,22
9	159	57	59,55	–4,90	24,01	–2,35	5,52	–2,55	6,50
10	173	67	73,26	5,10	26,01	11,36	128,97	–6,26	39,14
Σ		**619**			**812,90**		**456,61**		**356,29**

Calculation of R^2 (cont.)

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = 456.61 \qquad SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 = 812.90$$

$$R^2 = \frac{SSR}{SST} = \frac{456.61}{812.90} = 0.561$$

$R^2 = 0.5617$ indicates that the linear regression model only explains 56.17% of the variance of the observations

Calculation of r

i	x_i	y_i	$\bar{x} - x_i$	$\bar{y} - y_i$	$(\bar{x} - x_i)^2$	$(\bar{y} - y_i)^2$	$(\bar{x} - x_i)(\bar{y} - y_i)$
1	161	63	0,4	–1,1	0,16	1,21	–0,44
2	152	56	9,4	5,9	88,36	34,81	55,46
3	167	77	–5,6	–15,1	31,36	228,01	84,56
4	153	49	8,4	12,9	70,56	166,41	108,36
5	161	72	0,4	–10,1	0,16	102,01	–4,04
6	168	62	–6,6	–0,1	43,56	0,01	0,66
7	167	68	–5,6	–6,1	31,36	37,21	34,16
8	153	48	8,4	13,9	70,56	193,21	116,76
9	159	57	2,4	4,9	5,76	24,01	11,76
10	173	67	–11,6	–5,1	134,56	26,01	59,16
Σ	**1.614**	**619**			**476,40**	**812,90**	**466,40**

Calculation of r (cont.)

Sample covariance: $S_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{466.40}{10-1} = 51.822$

Sample standard deviation (X): $S_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{476.40}{10-1} = 52.933 \longrightarrow S_x = 7.276$

Sample standard deviation (Y): $S_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = \frac{812.90}{10-1} = 90.322 \longrightarrow S_y = 9.504$

$$r = \frac{S_{xy}}{S_x S_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{51.822}{7.276 \cdot 9.504} = 0.749$$

$r = 0.749$ indicates a moderate relationship between these two variables: as height grows, weight also grows

Relationship between R^2 and r

We have obtained the following values:

$$R^2 = 0.561 \qquad \text{and} \qquad r = 0.749$$

Now you can check the relationship between R^2 and r:

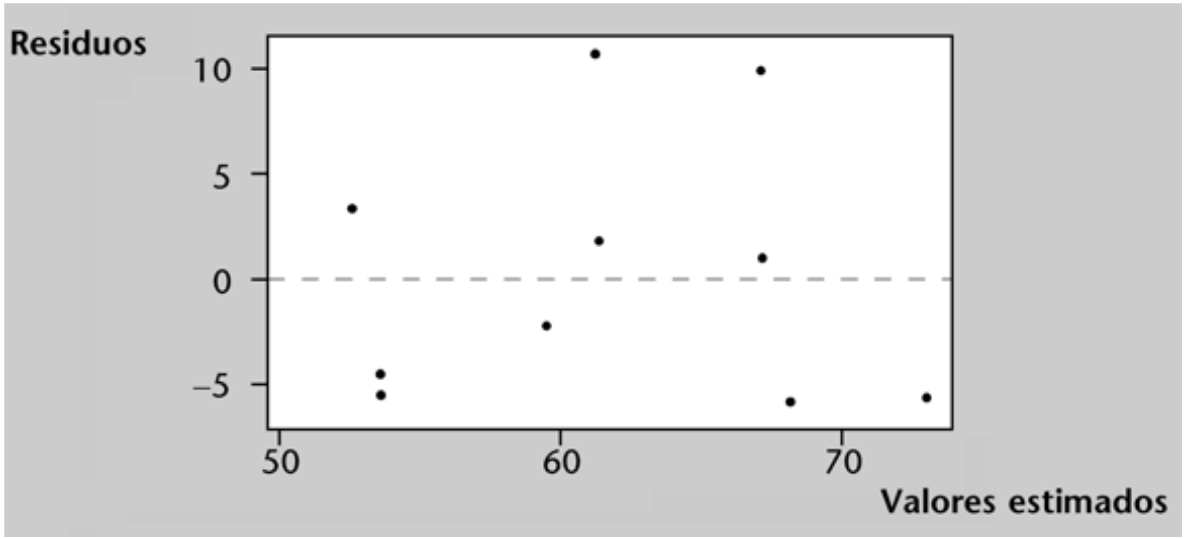
$$\boxed{r^2 = 0.749^2 = 0.561}$$

Regression diagnosis: residual analysis

i	x_i	y_i	$\hat{y}_i$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$\hat{y}_i - \bar{y}$	$(\hat{y}_i - \bar{y})^2$	e_i	e_i^2
1	161	63	61,51	1,10	1,21	–0,39	0,15	1,49	2,23
2	152	56	52,70	–5,90	34,81	–9,20	84,69	3,30	10,91
3	167	77	67,38	15,10	228,01	5,48	30,06	9,62	92,50
4	153	49	53,68	–12,90	166,41	–8,22	67,63	–4,68	21,87
5	161	72	61,51	10,10	102,01	–0,39	0,15	10,49	110,07
6	168	62	68,36	0,10	0,01	6,46	41,75	–6,36	40,47
7	167	68	67,38	6,10	37,21	5,48	30,06	0,62	0,38
8	153	48	53,68	–13,90	193,21	–8,22	67,63	–5,68	32,22
9	159	57	59,55	–4,90	24,01	–2,35	5,52	–2,55	6,50
10	173	67	73,26	5,10	26,01	11,36	128,97	–6,26	39,14
Σ		**619**			**812,90**		**456,61**		**356,29**

Regression diagnosis: residual analysis (cont.)

Draw the residual plot:



We cannot observe any type of structure in the representation. Therefore, we can conclude that the regression model obtained is a good model to explain the relationship between the two variables

Hypothesis testing

1. specify the null and alternative hypotheses:

   Null hypothesis: $H_0: \beta_1 = 0$ (the variable x is not explanatory)
   Alternative hypothesis: $H_a: \beta_1 \neq 0$ (the variable x is explanatory)

2. set a significance level:

   $\alpha = 0.05$

Hypothesis testing (cont.)

3. construct a statistic T to test the null hypothesis H_0:

$$SE(\beta_1) = \sqrt{\frac{MSE}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-2)}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \sqrt{\frac{356.29/(10-2)}{476.4}} = 0.306$$

$$T = \frac{\beta_1}{SE(\beta_1)} = \frac{0.979}{0.306} = 3.199$$

4. define a decision rule to reject, or not, the null hypothesis H_0:

$$P(|t_{n-2}| > T) = 2P(t_{n-2} > T) = 2P(t_8 > 3.199) = 2 \cdot 0.0063 = \mathbf{0.0126 < 0.05}$$

Since the *p*-value is less than α, we reject the null hypothesis