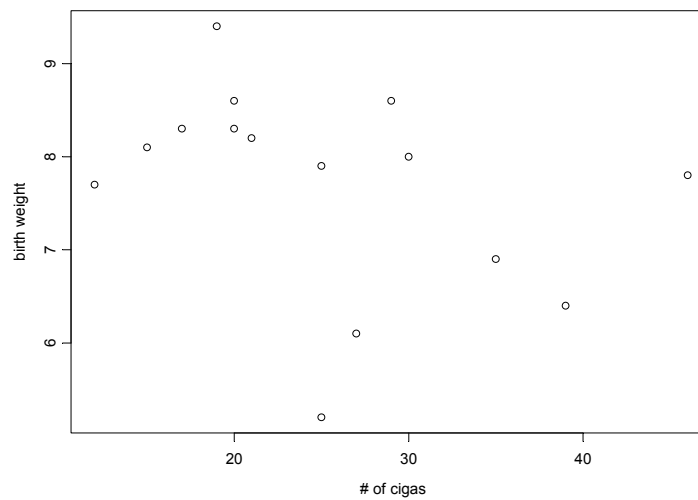


## Simple Linear Regression: A Numerical Example

A study is conducted to investigate the relationship between cigarette smoking during pregnancy and the weights of newborn infants. A sample of 15 woman smokers kept accurate records of the number of cigarettes smoked during their pregnancies, and weights of their children were recorded at birth. The data are given in the following table.

id	Cigarettes Per Day (X)	Birth Weight (Y)
1	12	7.7
2	15	8.1
3	35	6.9
4	21	8.2
5	20	8.6
6	17	8.3
7	19	9.4
8	46	7.8
9	20	8.3
10	25	5.2
11	39	6.4
12	25	7.9
13	30	8
14	27	6.1
15	29	8.6

### 1. Scatterplot of the Data. (the Fitted LS Line has been Added)



## 2. Preliminary Calculation, including the Pearson's Correlation Coefficient

id	Cigarettes Per Day (X)	Birth Weight (Y)	$X_i^2$	$Y_i^2$	$X_i \cdot Y_i$
1	12	7.7	144	59.29	92.4
2	15	8.1	225	65.61	121.5
3	35	6.9	1225	47.61	241.5
4	21	8.2	441	67.24	172.2
5	20	8.6	400	73.96	172.0
6	17	8.3	289	68.89	141.1
7	19	9.4	361	88.36	178.6
8	46	7.8	2116	60.84	358.8
9	20	8.3	400	68.89	166.0
10	25	5.2	625	27.04	130.0
11	39	6.4	1521	40.96	249.6
12	25	7.9	625	62.41	197.5
13	30	8	900	64.00	240.0
14	27	6.1	729	37.21	164.7
15	29	8.6	841	73.96	249.4
<b>Sum</b>	<b>380</b>	<b>115.5</b>	<b>10,842</b>	<b>906.27</b>	<b>2875.3</b>

First it can be found that  $\sum x_i = 380$ ,  $\sum y_i = 115.5$ ,  $\sum x_i^2 = 10842$ ,  $\sum y_i^2 = 906.3$ , and  $\sum x_i y_i = 2875.3$ . Hence,

$$\bar{x} = 380/15 = 25.33, \quad \bar{y} = 115.5/15 = 7.7$$

$$\begin{aligned} SS_{xx} &= \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n \cdot \bar{x}^2 \\ &= 10,842 - 15 \times 25.33^2 \\ &= 1215.33 \end{aligned}$$

$$\begin{aligned} SS_{yy} &= \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - n \cdot \bar{y}^2 \\ &= 906.27 - 15 \times 7.7^2 \\ &= 16.92 \end{aligned}$$

$$\begin{aligned} SS_{xy} &= \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - n \cdot \bar{x} \cdot \bar{y} \\ &= 2875.3 - 15 \times 25.33 \times 7.7 \\ &= -50.7 \end{aligned}$$

Therefore,  $r = \frac{SS_{xy}}{\sqrt{SS_{xx} \cdot SS_{yy}}} = \frac{-50.7}{\sqrt{1215.33 \times 16.92}} = -0.3536$ , which shows a somehow weak negative linear association.

3. Compute the Spearman correlation coefficient,  $r_s(X, Y)$ .

$r_s(X, Y)$  is defined as the Pearson correlation coefficient based on ranks of  $X$  and  $Y$ .

**A Shortcut Formula for Computing Spearman's Rank Correlation Coefficient**

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)},$$

where  $d_i$  is the difference in the ranks of the  $i$ -th observations for  $X$  and  $Y$ .

id	$X_i$	Rank	$Y_i$	Rank	$d_i$	$d_i^2$
1	12	1	7.7	5	-4	16
2	15	2	8.1	9	-7	49
3	35	13	6.9	4	9	81
4	21	7	8.2	10	-3	9
5	20	5.5	8.6	13.5	-8	64
6	17	3	8.3	11.5	-8.5	72.25
7	19	4	9.4	15	-11	121
8	46	15	7.8	6	9	81
9	20	5.5	8.3	11.5	-6	36
10	25	8.5	5.2	1	7.5	56.25
11	39	14	6.4	3	11	121
12	25	8.5	7.9	7	1.5	2.25
13	30	12	8	8	4	16
14	27	10	6.1	2	8	64
15	29	11	8.6	13.5	-2.5	6.25
					<b>Sum =</b>	<b>795</b>

Therefore,

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} = 1 - \frac{6 \times 795}{15 \times (15^2 - 1)} = 1 - 1.42 = -0.42$$

**4, Compute the LSE's of  $(\beta_0, \beta_1)$  and then add the LS straight line to the scatter plot.**

Model:  $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ , where  $\varepsilon_i \sim N(0, \sigma^2)$

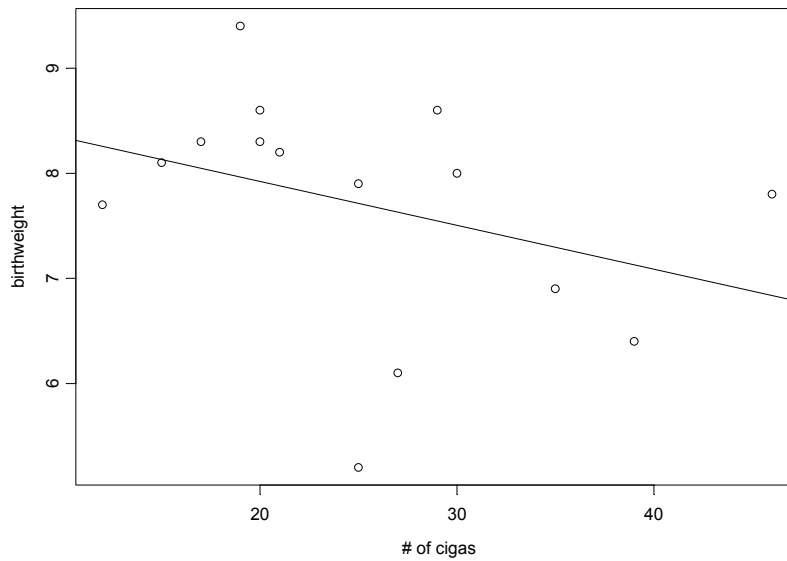
Compute the least square estimates (LSE) of  $(\beta_0, \beta_1)$ .

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{-50.7}{1215.33} = -0.0417$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x} = 7.7 - (-0.0417 \times 25.33) = 8.7568$$

Hence the fitted straight line is  $y = 8.7568 - 0.0417x$ .

To add the line, pick any two points, e.g.,  $(0, \hat{\beta}_0) = (0, 8.7568)$  and  $(\bar{x}, \bar{y}) = (25.33, 7.7)$



**5, Complete the following ANOVA table**

Source	df	SS	MS	F Value	P-value
Model	1	2.115	2.115	1.857	0.196
Error	13	14.805	1.139		
Total	14	16.915			

**Note:**

- ❖  $SSTotal = SS_{yy} = 16.92$
- ❖  $SSR(\text{regression}) = \hat{\beta}_1^2 \cdot SS_{xx}$  or  $\hat{\beta}_1 \cdot SS_{xy} = (-0.0417) \times (-50.7) = 2.115$
- ❖  $F_{.95}^{(1,13)} = 4.667$

- ❖ The coefficient of Determination  $R^2 = 14.28\%$
- ❖ A natural estimate of the constant variance  $\hat{\sigma}_2 = MSE = 1.139$

## 6, Statistical Inference on $\beta_1$

95% Confidence Interval for  $\beta_1$ :

$$\begin{aligned} \hat{\beta}_1 \pm t_{.975}^{(13)} \cdot \sqrt{\frac{\hat{\sigma}^2}{SS_{xx}}} &\Rightarrow -0.0417 \pm 2.160 \times \sqrt{\frac{1.139}{1215.22}} \\ &\Rightarrow -0.0417 \pm 2.160 \times 0.0306 \Rightarrow (-0.1078, 0.0244) \end{aligned}$$

**Question:** Test to see if the following statement is true at  $\alpha = 0.05$ : every 10 cigarettes increase per day for the mother would result in (more than) half-pound reduction in the infant birth weight.

**Hint:** This is equivalent to test  $H_0 : 10 \cdot \beta_1 = -0.5 \Leftrightarrow \beta_1 = -0.05$   
 $H_a : \beta_1 < -0.05$

## 7, Prediction

Estimation: Construct a 95% confidence interval for the mean infant birth weight of all mothers who smoke 10 cigarettes per day.

$$\begin{aligned} (\hat{\beta}_0 + \hat{\beta}_1 x_p) \pm t_{.975}^{(13)} \sqrt{\hat{\sigma}^2 \cdot \left( \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}} \right)} \\ \{8.757 + (-0.0417) \times 10\} \pm 2.160 \times \sqrt{1.138 \times \left( \frac{1}{15} + \frac{(10-25.33)^2}{1215.33} \right)} \Rightarrow (7.164, 9.516) \end{aligned}$$

Prediction: Given a mother who smokes 10 cigarettes per day, construct a 95% prediction interval for the birth weight of her baby.

$$\begin{aligned} (\hat{\beta}_0 + \hat{\beta}_1 x_p) \pm t_{.975}^{(13)} \sqrt{\hat{\sigma}^2 \cdot \left( 1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}} \right)} \\ \{8.757 + (-0.0417) \times 10\} \pm 2.160 \times \sqrt{1.138 \times \left( 1 + \frac{1}{15} + \frac{(10-25.33)^2}{1215.33} \right)} \Rightarrow (5.752, 10.928) \end{aligned}$$

### 8, Confidence and Prediction Bands

**Linear Fit with Confidence/Prediction Bands**

