University of Ferrara

DIPARTIMENTO DI ECONOMIA E MANAGEMENT

Stefano Bonnini & Valentina Mini

Simple Linear Regression Analysis: interpretation

Lecture 4
2019, Feb 20th

Contents

1) How to perform SLR by hand (the model)
2) How to interpret the results
3) How to

   - perform a Simple Linear Regression analysis using R (*please, see "cakes" dataset*) and

   - interpret the output

1 – How to perform SLR by hand

The simple linear regression **model**

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- Dependent Variable: Y_i
- Population Y intercept: β_0
- Population Slope Coefficient: β_1
- Independent Variable: X_i
- Random Error term: ε_i

Linear component: $\beta_0 + \beta_1 X_i$

Random Error component: ε_i

The equation of **estimated Y value**

The simple linear regression equation provides an estimate of the population regression line

$$\hat{Y}_i = b_0 + b_1 X_i$$

- $\hat{Y}_i$: Estimated (or predicted) Y value for observation i
- b_0: Estimate of the regression intercept
- b_1: Estimate of the regression slope
- X_i: Value of X for observation i

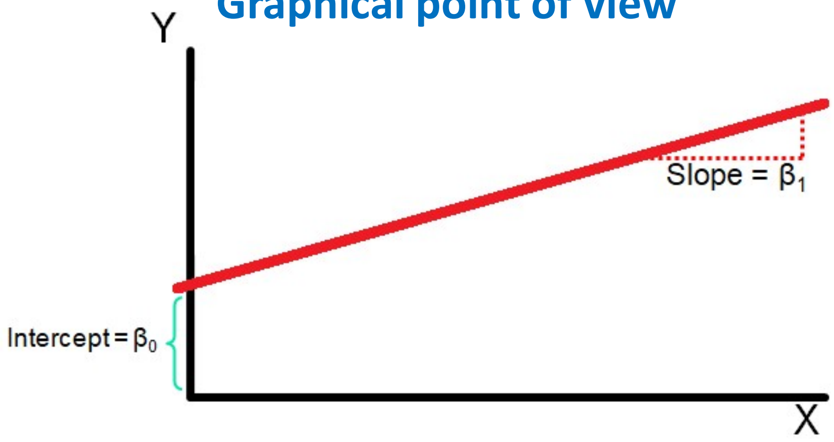
Starting from a database of observed variables (Yi and Xi)
we aim to identify the equation:

$$\hat{Y}_i = b_0 + b_1 X_i$$

To identify the actual equation we must find out the values of:

- b_0 (called intercept)
- B_1 (the slope)

Graphical point of view



Algebras point of view:
the least squares method (OLS)

b_0 and b_1 are obtained by finding the values that minimize the sum of the squared differences between Y and $\hat{Y}$:

$$\min \sum (Y_i - \hat{Y}_i)^2 = \min \sum (Y_i - (b_0 + b_1 X_i))^2$$

The values of our coefficients

$$\hat{Y}_i = b_0 + b_1 X_i$$

The **least squares estimates** of the intercept and slope in the simple linear regression model are

$$\mathbf{b_0} = \bar{y} - \mathbf{b_1}\bar{x}$$

$$\mathbf{b_1} = \frac{\sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}$$

where $\bar{y} = (1/n)\sum_{i=1}^{n} y_i$ and $\bar{x} = (1/n)\sum_{i=1}^{n} x_i$.

The meaning of the coefficients

$$\hat{Y}_i = b_0 + b_1 X_i$$

- b_0 ($\hat{\beta}_0$)is the estimated mean value of Y when the value of X is zero

- b_1 ($\hat{\beta}_1$)is the estimated change in the mean value of Y as a result of a one-unit change in X

week	sold_cakes (units)	unit_price $
1	280	4
2	290	4,2
3	300	5
4	300	5
5	300	5,1
6	310	5,2
7	320	5,5
8	330	5,7
9	340	5,7
10	350	5,8
11	350	5,8
12	350	5,9
13	360	4
14	370	4,2
15	380	4,3
16	380	4,3
17	410	5
18	410	5
19	420	5,5
20	430	5,7
21	430	5,8
22	440	6
23	450	7
24	450	5
25	450	5,5
26	460	5,6
27	460	5,6
28	470	5,8
29	470	6
30	490	6
31	500	7
32	500	7,5
33	505	8
34	510	8

How to perform SLR

1) STARTING POINT: THE DATA

We have 34 observation (rows) and 2 variables (columns) collected in 34 weeks about:

- Units of cakes sold by week

  = measurement in “units of cake sold”

- Price per cake (unit) applied in that week

  = measurement in “$”

week	sold_cakes (units)	unit_price $
1	280	4
2	290	4,2
3	300	5
4	300	5
5	300	5,1
6	310	5,2
7	320	5,5
8	330	5,7
9	340	5,7
10	350	5,8
11	350	5,8
12	350	5,9
13	360	4
14	370	4,2
15	380	4,3
16	380	4,3
17	410	5
18	410	5
19	420	5,5
20	430	5,7
21	430	5,8
22	440	6
23	450	7
24	450	5
25	450	5,5
26	460	5,6
27	460	5,6
28	470	5,8
29	470	6
30	490	6
31	500	7
32	500	7,5
33	505	8
34	510	8

2) SECOND STEP: IMMAGINE THE RELATIONSHIP OF DEPENDENCE

Which variable is the explanatory one?

Which variable is the dependent variables?

Try to identify the model!

Y= b0+b1*x1

CENTRAL RESEARCH QUESTION:

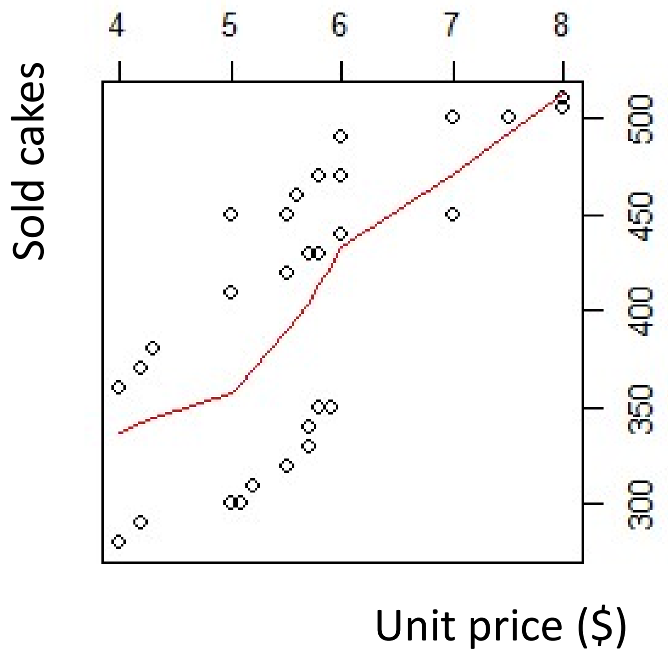
Is the number of cakes sold per week affected by the unit's price?

To investigate this question we define our model:

*Units of sold cakes = b0 + b1 * price per unit*

week	sold_cakes (units)	unit_price $
1	280	4
2	290	4,2
3	300	5
4	300	5
5	300	5,1
6	310	5,2
7	320	5,5
8	330	5,7
9	340	5,7
10	350	5,8
11	350	5,8
12	350	5,9
13	360	4
14	370	4,2
15	380	4,3
16	380	4,3
17	410	5
18	410	5
19	420	5,5
20	430	5,7
21	430	5,8
22	440	6
23	450	7
24	450	5
25	450	5,5
26	460	5,6
27	460	5,6
28	470	5,8
29	470	6
30	490	6
31	500	7
32	500	7,5
33	505	8
34	510	8

3) THIRD STEP: OBSERVE THE PLOT AND MAKE COMMENTS ABOUT THE POSSIBLE RELATIONSHIP BETWEEN VARIABLES



#comments: Do you think we can expect a linear causal relationship between Price and Sold_cakes?

4^{th} STEP: CALCULATE THE COEFFICIENTS

5 basic elements to estimate the coefficients

- n
- $\sum_{i=1}^{n} x$
- $\sum_{i=1}^{n} y$
- $\sum_{i=1}^{n} xx$
- $\sum_{i=1}^{n} xy$

$$b_1 = ssxy/ssx$$

SSXY=

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

SSX=

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

$$b_0 = \bar{Y} - b_1 \bar{x}$$

$$\bar{Y} = \frac{\sum_{1=i}^{n} y_i}{n}$$

$$\bar{x} = \frac{\sum_{1=i}^{n} x_i}{n}$$

week	sold_cakes (units)	unit_price $
1	280	4
2	290	4,2
3	300	5
4	300	5
5	300	5,1
6	310	5,2
7	320	5,5
8	330	5,7
9	340	5,7
10	350	5,8
11	350	5,8
12	350	5,9
13	360	4
14	370	4,2
15	380	4,3
16	380	4,3
17	410	5
18	410	5
19	420	5,5
20	430	5,7
21	430	5,8
22	440	6
23	450	7
24	450	5
25	450	5,5
26	460	5,6
27	460	5,6
28	470	5,8
29	470	6
30	490	6
31	500	7
32	500	7,5
33	505	8
34	510	8

4th STEP: CALCULATE THE COEFFICIENTS

we need to calculate the following elements:

n= 34 = total number of observations (rows)

$\sum_{i=1}^{n} x$ = total sum of unit_price (4+4.2+5+5+5.1+…..+8+8) = 189.7 $

Average x = 189.7/34 = 5.58 $

$\sum_{i=1}^{n} y$ = total sum of sold_cakes (280+290+300+300+…505+510)=13565 cakes

Average y = 13565/34 = 398.97 cakes

$\sum_{i=1}^{n} xx$ = 4*4+4.2*4.2+5*5+…+8*8+8*8=1092.87

$\sum_{i=1}^{n} xy$ = 4*280+4.2*290+…+8*505+8*210 =77324

4th STEP: CALCULATE THE COEFFICIENTS

$$b_1 = ssxy/ssx = \frac{\sum_{1=i}^{n} y_i x_i - \frac{(\sum_{1=i}^{n} x_i)(\sum_{1=i}^{n} y_i)}{n}}{\sum_{1=i}^{n} x_i^2 - \frac{(\sum_{1=i}^{n} x_i)^2}{n}} = \frac{77324 - (189.7*13565)/34}{1092-(189.7^2)/34} = 48.809$$

$$b_0 = \bar{Y} - b_1 \bar{x} = 398.97 - (48.809*5.58) = 125.616$$

week	sold_cakes (units)	unit_price $
1	280	4
2	290	4,2
3	300	5
4	300	5
5	300	5,1
6	310	5,2
7	320	5,5
8	330	5,7
9	340	5,7
10	350	5,8
11	350	5,8
12	350	5,9
13	360	4
14	370	4,2
15	380	4,3
16	380	4,3
17	410	5
18	410	5
19	420	5,5
20	430	5,7
21	430	5,8
22	440	6
23	450	7
24	450	5
25	450	5,5
26	460	5,6
27	460	5,6
28	470	5,8
29	470	6
30	490	6
31	500	7
32	500	7,5
33	505	8
34	510	8

5th STEP: TRANSCRIPT THE MODEL

$$\widehat{SOLD\_CAKES} = 125.616+UNIT\_PRICES*48.809$$

NOW WE CAN INDIVIDUATE THE ESTIMATED Y VALUES:

WEEK1:

-Estimated Y value: 125.616+4*48.809 = 320.852 sold_cakes

-Real (observed) Y value : 280

The difference between 280 and 320.852 is the error made by our model.

#exercise: please calculate the estimated Y value for the 2nd week.

2 -How to interpret the results

$$\widehat{SOLD\_CAKES} = 125.616 + UNIT\_PRICES * 48.809$$

6th step: interpreting the result

- b_1 = when the price of one cake increases by 1 $, we expect that the number of sold_cakes increases by 48.809 units

- b_0 = when the price of one cake is 0$ → for that week the estimated sold_cakes will be 125.616

be careful about the real meaning of your interpretation!!!

NB: the problem of the X_1 (unit' price) range

$$\widehat{SOLD\_CAKES} = 125.616 + UNIT\_PRICES * 48.809$$

7th step: making predictions

1) Control the X range

In our case X(4$; 8$)

1) Make the prediction for values within the range

I.e. : How many cakes we expect to sell in a week in which the applied price is 5.3$ per cake?

→ 125.616+5.3*48.809 = 384.304 cakes → 384 cakes c.a.

$$\widehat{SOLD\_CAKES} = 125.616 + UNIT\_PRICES * 48.809$$

8th step: assessing the goodness of fit

using the coefficient of determination

It provides a measure of how well observed outcomes are replicated by the model

$$R^2 = SSR/SST$$

$$SSR = SUM\ (\hat{Y}-\bar{Y})^2$$

$$SST << SSR+SSE = SUM(y_i-\bar{Y})^2$$

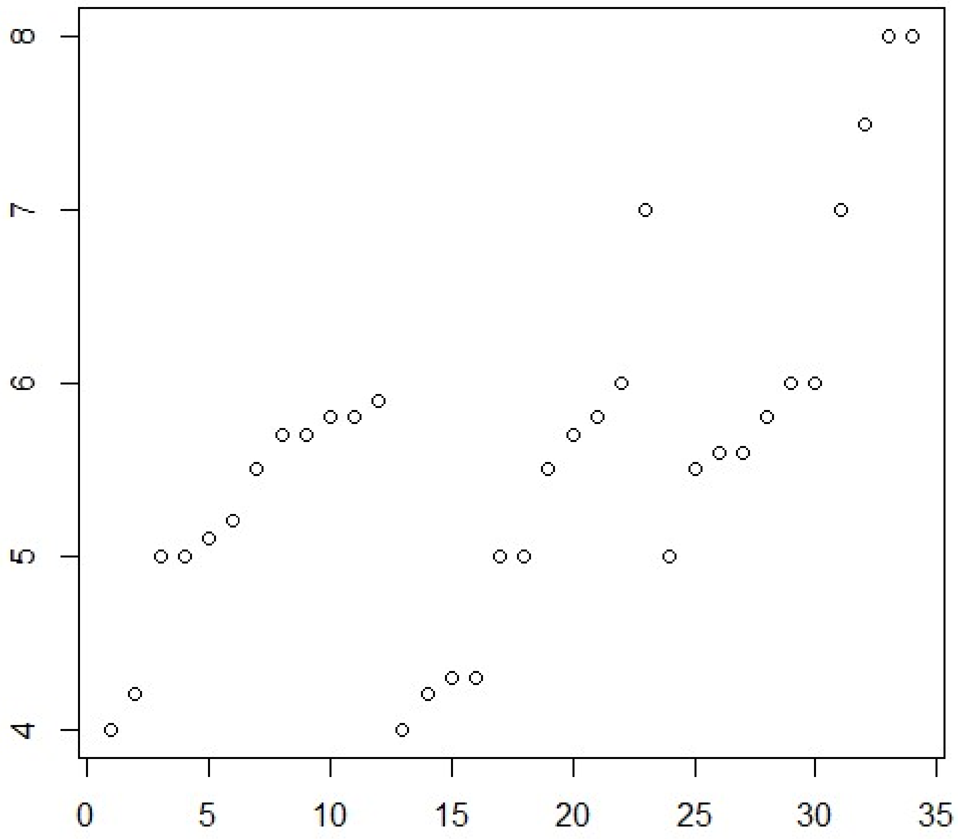
In our case: $R^2 = 0.4588$ → using our model the 45.88% of total variance is explained

The unexplained variance (1-0.4588) may be due to additional variables or different relationship between the observed variables.

$$\widehat{SOLD\_CAKES} = 125.616 + UNIT\_PRICES * 48.809$$

9th step: interpreting the residuals of our model (to confirm the 4 basic assumptions)

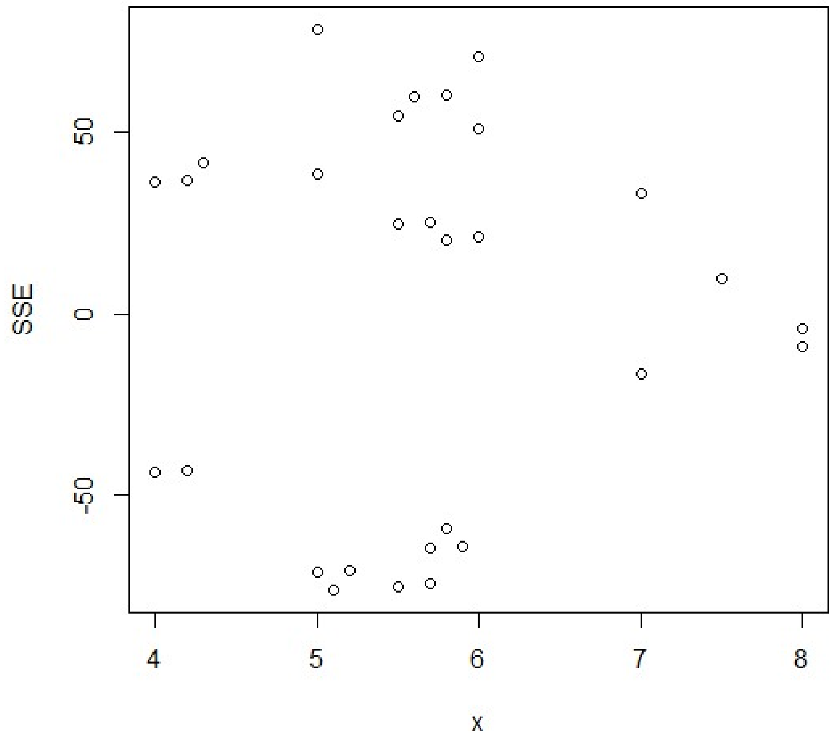
•Examine for linearity



$$\widehat{SOLD\_CAKES} = 125.616 + UNIT\_PRICES * 48.809$$

9th step: interpreting the residuals of our model (to confirm the 4 basic assumptions)

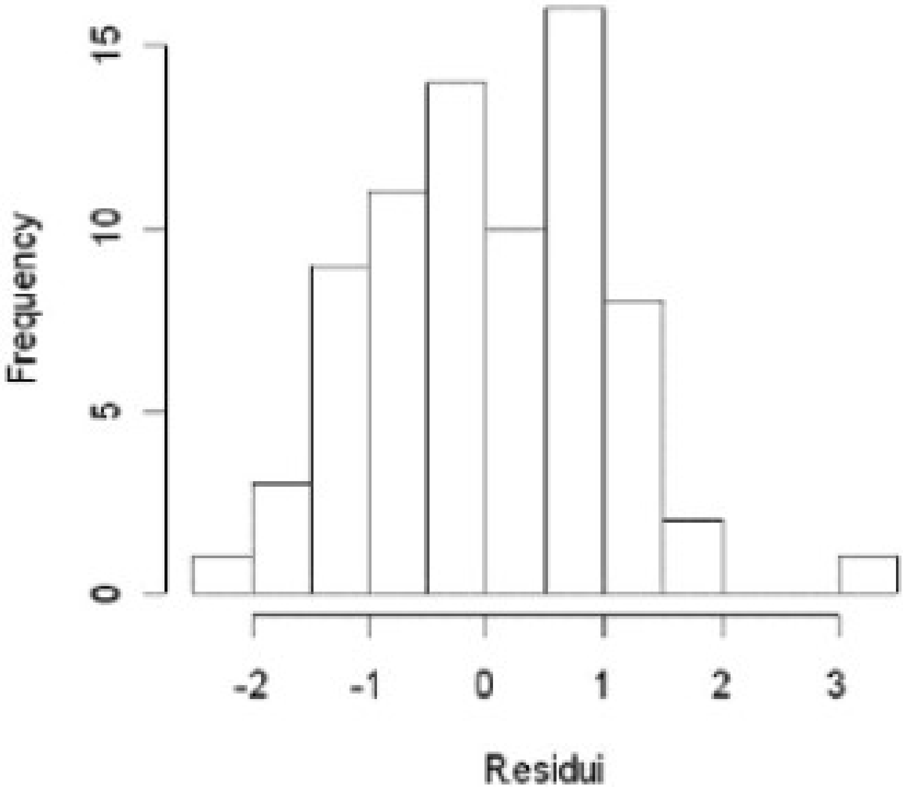
- Evaluate independence assumption
- Examine for constant variance for all levels of X (homoscedasticity)



$$\widehat{SOLD\_CAKES} = 125.616 + UNIT\_PRICES * 48.809$$

9th step: interpreting the residuals of our model (to confirm the 4 basic assumptions)

- Evaluate normal distribution of residuals (histogram of the residuals)



4 - How to perform LRM using R

Simple Linear Regression Model using R

UNIFE

Spring Semester

Mini V. 20-02-2019

RESEARCH QUESTION:

does exist a linear causal relationship between the number of cakes sold in a week (by a firm) and the unit's price (the price applied per cake)?

Let's observe a given dataset and perform a simple linear regression analysis

#Analysis: step by step

0. LET'S PREPARE THE DATASET

1. Visualize the relationship: the scatter plot

2. Identify the estimated model

3. The model on a graph

4. Prediction: the expected Y values given a X value

5. The model's goodness of fit

6. Graphical analysis of Linear Regression Model's assumptions

7. what about the inference? #