University of Ferrara

Stefano Bonnini & Valentina Mini

Simple Linear Regression Analysis: What about the inference?

Lecture 5

2019, Feb 22nd

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

The state of the art:

-We are studying **the linear relationship** between unit-price (x) and number of total cakes sold per week (y).

-The **goodness of fit (R^2)** is 0.4588: thus the 45.88% of total variance in y is explained by our model (on the other side, the 54.12% of that variance is still unexplained!)

- We **graphically tested the 4 main linear regression conditions** (plot: error terms and its relationship with the explanatory variable)

Now, our final aim is to understand whether this relationship between x and y does exist within the population as a whole

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

We start from sample and we're trying to estimate true population parameters

Because we analyze a sample (and not the entire population)
we need to make **inference** based on our sample.

We know that b_2 is unlikely to be exactly equals $ß_2$,

But, HOW CONFIDENT CAN WE BE THAT THERE IS AT LEAST A POSITIVE LINEAR RELATIONSHIP within the population BETWEEN UNIT-PRICE (x) AND TOTAL SOLD-CAKES (y)?

HOW CONFIDENT CAN WE BE THAT THERE IS AT LEAST
A NONZERO LINEAR RELATIONSHIP WITHIN THE POPULATION?

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

We may proceed in two different ways:

1) Testing the null/alternative hypothesis using t-statistic and T-student

2) Comparing the significance level (α) and the p-value of the coefficient

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

a) Testing the null/alternative hypothesis : the t-test for a population slope

Central question:
Is there a linear relationship between unit_price (X) and the number of cakes sold in a week (Y) in the general population?

Null and alternative hypotheses:

H_0: $\beta_1 = 0$(no linear relationship)

H_1: $\beta_1 \neq 0$(linear relationship does exist)

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

a) Testing the null/alternative hypothesis : the t-test for a population slope

Decision rule:

We should compare t-stat with a critical value ($t_{a/2}$)

$$t_{STAT} = \frac{b_1 - \beta_1}{S_{b_1}}$$

where:

b_1 = regression slope coefficient

β_1 = hypothesized slope

S_{b1} = standard error of the slope

We need to know:

-The significance level (α)

- The degree of freedom

If T-stat > $t_{\alpha/2}$ → we reject H_0

a) Testing the null/alternative hypothesis : the t-test for a population slope

```
> summary(reg_lin)

Call:
lm(formula = y ~ x)

Residuals:
   Min      1Q Median      3Q     Max
-76.16 -59.47  20.78  41.07  78.60

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  133.521     51.788   2.578   0.0147 *
x             47.577      9.134   5.208 1.08e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 53.62 on 32 degrees of freedom
Multiple R-squared:  0.4588,    Adjusted R-squared:  0.4419
F-statistic: 27.13 on 1 and 32 DF,  p-value: 1.084e-05
```

b_1 S_{b_1}

T-stat = 47.577 / 9.134 = 5.2088

a) Testing the null/alternative hypothesis : the t-test for a population slope

NOW WE NEED TO COMPUTE THE $t_{\alpha/2}$ value:

i) The significance level (α) is defined a priori (considering the value of our research) : let's imagine we want a **confidence level of 95%,** so our **significance level is (1-95** = 0.05) → α = 0.05 →

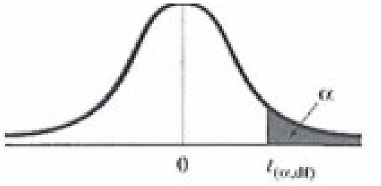
$\alpha/2 = 0.05/2 = 0.025$

ii) The degree of freedom (d.f.) for linear regression is n-2: thus in our case d.f. = 34-2 =32

iii) Using those information, let's check on a **T-student** table to discover the $t_{a/2}$ value: 2.037

a) Testing the null/alternative hypothesis : the t-test for a population slope

Tavola della distribuzione T di Student



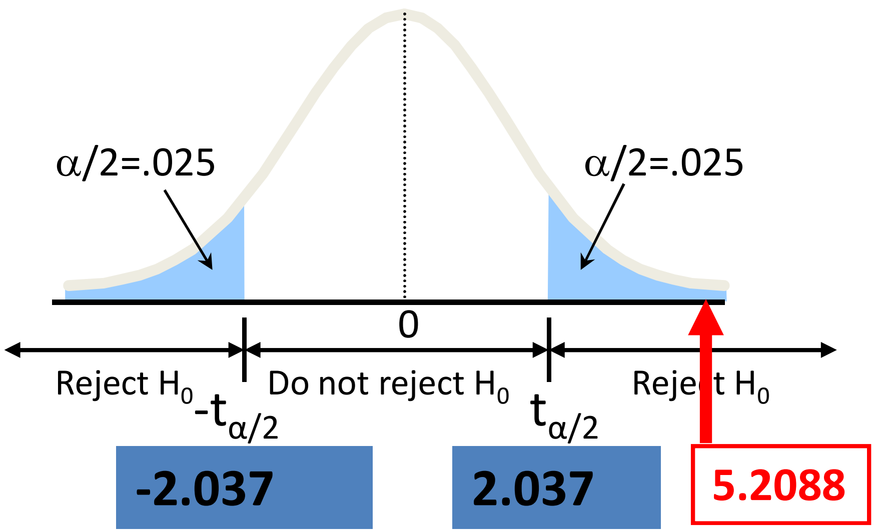
Gradi di libertà	Area nella coda di destra 0.1	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	3.078	6.314	12.706	15.894	31.821	63.656	127.321	318.289	636.578
2	1.886	2.920	4.303	4.849	6.965	9.925	14.089	22.328	31.600
3	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.214	12.924
4	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.894	6.869
6	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.610	3.922
19	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.689
28	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.660
30	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
31	1.309	1.696	2.040	2.144	2.453	2.744	3.022	3.375	3.633
32	1.309	1.694	2.037	2.141	2.449	2.738	3.015	3.365	3.622
33	1.308	1.692	2.035	2.138	2.445	2.733	3.008	3.356	3.611

a) Testing the null/alternative hypothesis : the t-test for a population slope

Last step: the comparison of computed values

T-stat = 5.2088
$t_{\alpha/2}$ = 2.037

T-stat > $t_{\alpha/2}$ → the statistic falls within the rejection area!



We reject the null hypothesis (H_0) , thus:

there is sufficient evidence that (at a 95% of confidence level) within the population the unit-price affects the number of cakes sold in a week

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

We may proceed in two different ways:

1) Testing the null/alternative hypothesis

2) Comparing the significance level (α) and the p-value of the coefficient

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

2)Comparing the significance level (α) and the p-value of the coefficient

Central question:
Is there a linear relationship between unit_price (X) and the number of cakes sold in a week (Y) in the general population?

Compare:
-The level of significance (α)
-The p(value) of the slope coefficient

Decision rule:
If the p-value < α → we reject H0

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

2)Comparing the significance level (α) and the p-value of the coefficient

let's imagine we want a **confidence level of 95%,** so our **level of significance is (1-95** = 0.05) → **α = 0.05**

The **p-value** associated to the b1 coefficient ≅ 0

```
> summary(reg_lin)

Call:
lm(formula = y ~ x)

Residuals:
   Min      1Q Median      3Q     Max
-76.16 -59.47  20.78  41.07  78.60

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  133.521     51.788   2.578   0.0147 *
x             47.577      9.134   5.208 1.08e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '

Residual standard error: 53.62 on 32 degrees of freedom
Multiple R-squared:  0.4588,    Adjusted R-squared:  0.441
F-statistic: 27.13 on 1 and 32 DF,  p-value: 1.084e-05
```

The p-value is $P(t \neq 5.208) \simeq 0$

We compare the p-value to our level of significance:

0.05 (α) > 0 (p-value) →

there is sufficient evidence that (at a 95% of confidence level) within the population the unit-price affects the number of cakes sold in a week