Regression I:

Hypothesis testing & predictions
+ Graphic Methods

Session 10

Programación Estadística con Python

Alberto Sanz, Ph.D

asanz@edem.es

MASTER EN DATA ANALYTICS PARA LA EMPRESA

Goals

EDEM
Centro Universitario

- Hypothesis testing over the relatioship of two quantitative variables by the means of regression.
  - Numeric approach (coefficients & p.values)
  - Graphic approach (Slope line)

- Measurements of Model fit
  - Numeric methods (residuals and R2)
  - Graphic methods (Scatterplot + trend line)

- Linear modeling & prediction:
  - Numeric methods (The regression function)
  - Graphic methods (Slope line + confidence interval+ rugs)

Regression (I)

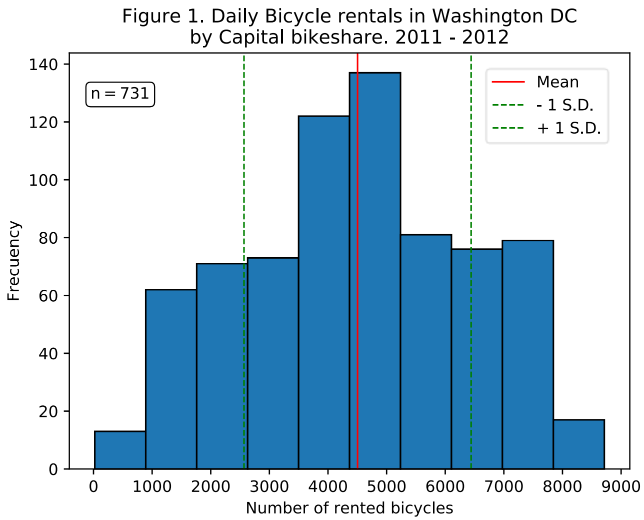
EDEM Centro Universitario

1. **Always DESCRIBE** the variables involved in the regression model separately. Check and validate the integrity of the data prior to any analysis.
2. **EXPLORE** of bivariate relation: **Scatterplot / Pearson's r**
3. **Fit your linear regression model carefully.** Pay attention to:
   a) **Slope & intercept**
   b) **P. value**
   c) **Model fit**
   d) **Sample size**
   e) **Model Diagnostics**

Research Question

Why some days are rent *more* bikes?

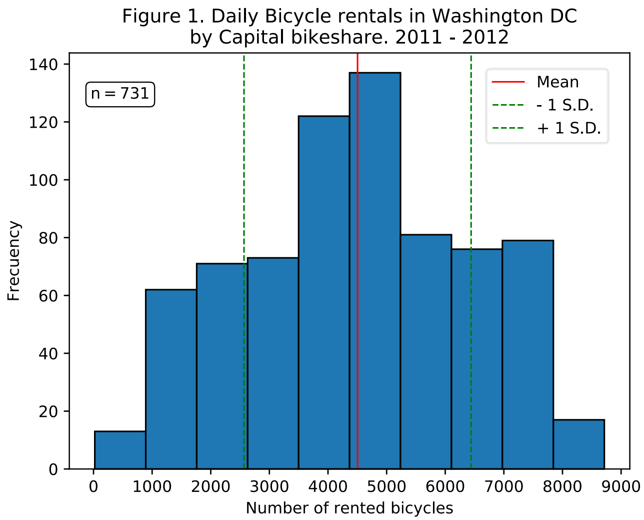
- Temperature ?



- H0.: There is no linear association (r=0) between the *number of rentals* and the *temperature*.
- H1.: There is a linear association (r≠0) between the *number of rentals* and the *temperature*.

Describing quantitative variables

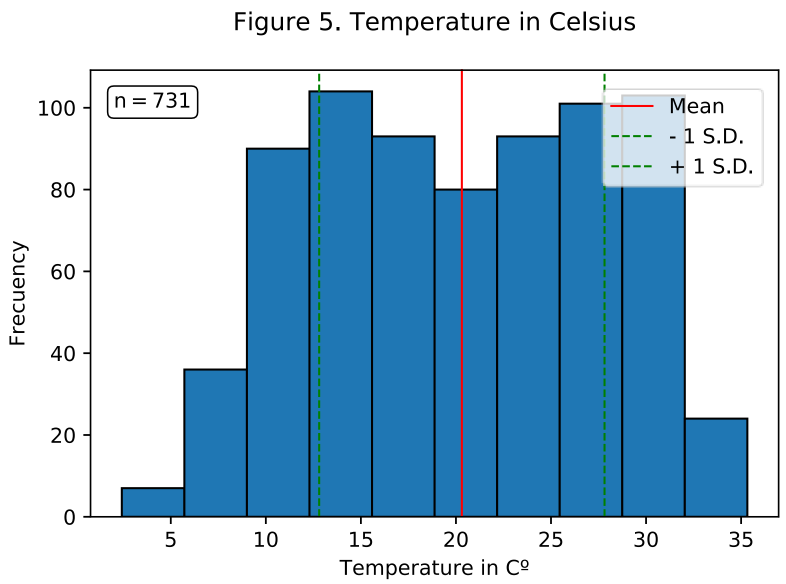

```
x=wbr['cnt']
plt.hist(x, bins=10,
edgecolor='black')
plt.xticks(np.arange(0, 10000,
step=1000))
plt.title('Figure 4. Daily Bicycle
rentals in Washington DC'
          '\n'
          'by Capital bikeshare.
2011 - 2012')
plt.ylabel('Frecuency')
plt.xlabel('Number of rented
bicycles')


props = dict(boxstyle='round',
facecolor='white', lw=0.5)
textstr = '$\mathrm{n}=%.0f$'%(n)
plt.text (-50,128, textstr ,
bbox=props)
```



Alberto Sanz, Ph.D. asanz@edem.es

Describing quantitative variables


```
##histogram ver4
x=wbr['temp_celsius']
plt.hist(x, bins=10,
edgecolor='black')
#plt.xticks(np.arange(0, 10000,
step=1000))
plt.title('Figure 5. Temperature in
Celsius'
          '\n')
plt.ylabel('Frecuency')
plt.xlabel('Temperature in C°')
props = dict(boxstyle='round',
facecolor='white', lw=0.5)
textstr = '$\mathrm{n}=%.0f$'%(n)
plt.text (2,100, textstr ,
bbox=props)
```

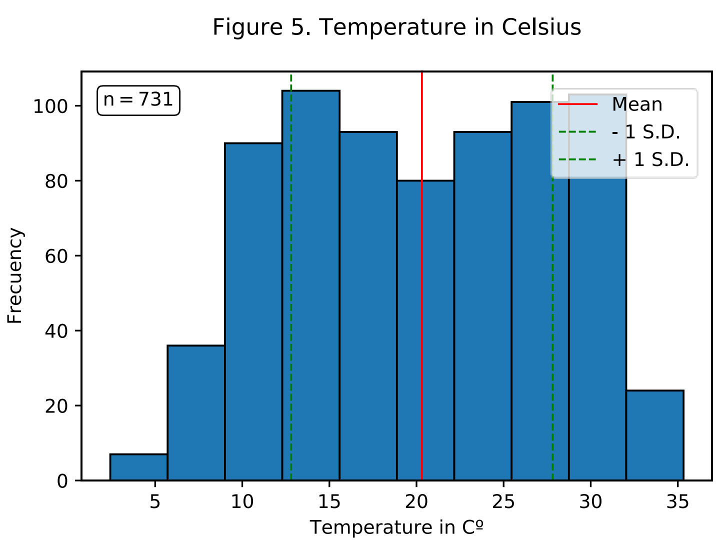


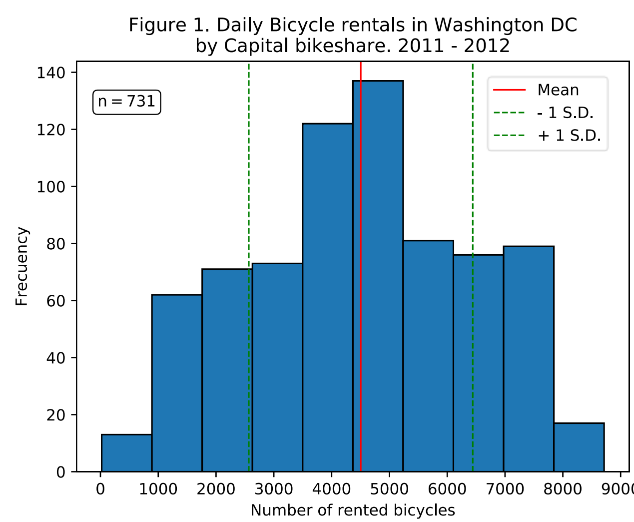
Regression

1. Describe the two variables involved in hypothesis

Temperature

Rentals





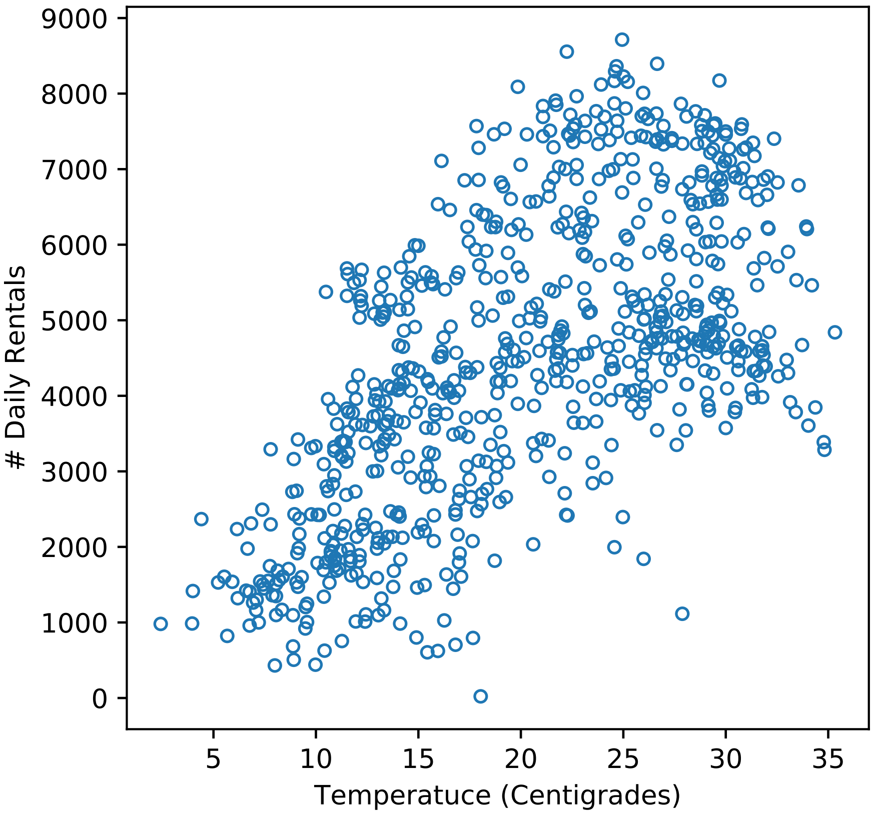
Regression

EDEM
Centro Universitario

2. Scatterplot

```
x=wbr.temp_Celsius
y=wbr.cnt
plt.scatter (x,y)
```

Figure 9. Daily bicycle rentals, by temperature.



Regression

3. Pearson's r

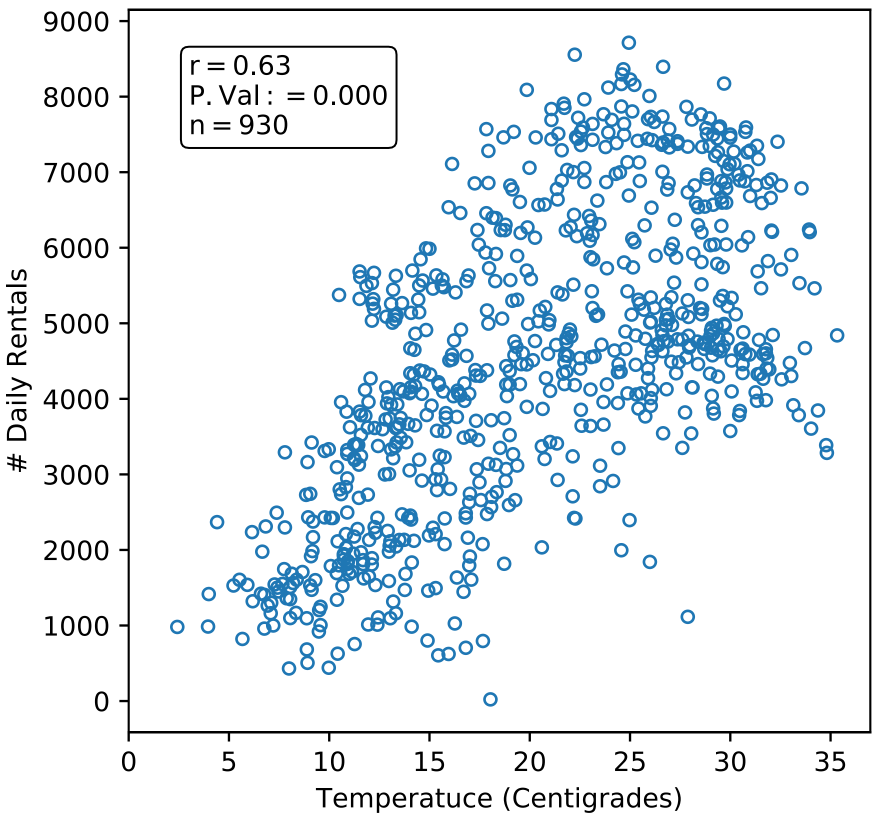
```
from scipy.stats.stats import pearsonr
res = pearsonr(x, y)
print (res)

[1] (0.627494009033349195, 2.810622397590l415e-81)
```

This is
Perason's r

This is
The P.Value

Scatterplot + Pearson's r + test

Figure 9. Daily bicycle rentals, by temperature.

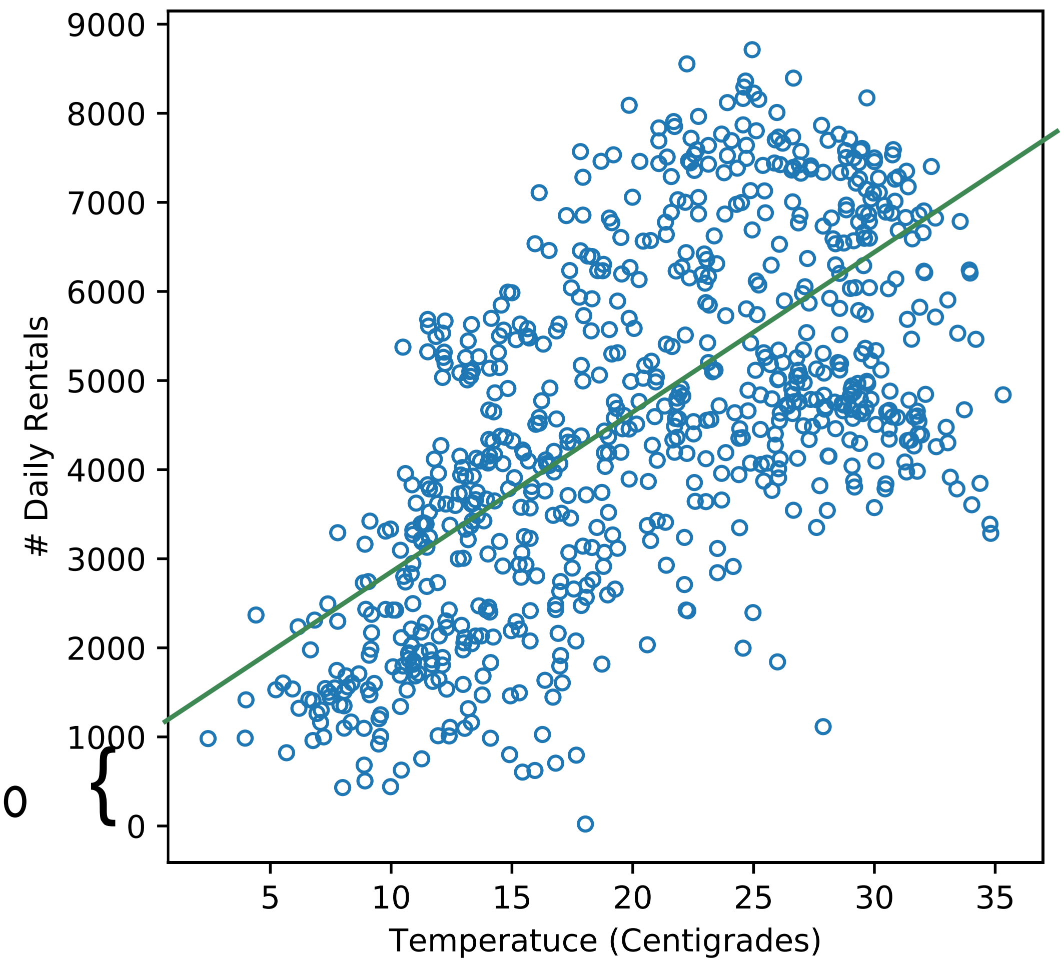


The Regression Model

$Y= a + bx$

EDEM Centro Universitario

Figure 9. Daily bicycle rentals, by temperature.



$a \approx 1200$

$b \approx 800/5 \approx 160$

$a \approx 1200$

The Regression Model


```python
# Regression
from statsmodels.formula.api import ols

model1 = ols('cnt ~ temp_celsius', data=wbr).fit()
model1.summary2()
```

```
                 Results: Ordinary least squares
==================================================================
Model:              OLS              Adj. R-squared:     0.393
Dependent Variable: cnt              AIC:                12777.5357
Date:               2019-12-11 12:23 BIC:                12786.7245
No. Observations:   731              Log-Likelihood:     -6386.8
Df Model:           1                F-statistic:        473.5
Df Residuals:       729              Prob (F-statistic): 2.81e-81
R-squared:          0.394            Scale:              2.2783e+06
------------------------------------------------------------------
                 Coef.   Std.Err.    t    P>|t|   [0.025    0.975]
------------------------------------------------------------------
Intercept      1214.6421 161.1635  7.5367 0.0000 898.2421 1531.0421
temp_celsius    161.9685   7.4436 21.7594 0.0000 147.3551  176.5820
------------------------------------------------------------------
Omnibus:               20.477        Durbin-Watson:          0.468
Prob(Omnibus):         0.000         Jarque-Bera (JB):       12.566
Skew:                  0.167         Prob(JB):               0.002
Kurtosis:              2.452         Condition No.:          63
==================================================================
```

The Regression Model


```python
# Regression
from statsmodels.formula.api import ols

model1 = ols('cnt ~ temp_celsius', data=wbr).fit()
model1.summary2()
```

```
                 Results: Ordinary least squares
==================================================================
Model:              OLS              Adj. R-squared:     0.393
Dependent Variable: cnt              AIC:                12777.5357
Date:               2019-12-11 12:23 BIC:                12786.7245
No. Observations:   731              Log-Likelihood:     -6386.8
Df Model:           1                F-statistic:        473.5
Df Residuals:       729              Prob (F-statistic): 2.81e-81
R-squared:          0.394            Scale:              2.2783e+06
------------------------------------------------------------------
                 Coef.   Std.Err.    t     P>|t|   [0.025   0.975]
------------------------------------------------------------------
Intercept      1214.6421 161.1635  7.5367 0.0000 898.2421 1531.0421
temp_celsius    161.9685   7.4436 21.7594 0.0000 147.3551  176.5820
------------------------------------------------------------------
Omnibus:               20.477        Durbin-Watson:           0.468
Prob(Omnibus):         0.000         Jarque-Bera (JB):        12.566
Skew:                  0.167         Prob(JB):                0.002
Kurtosis:              2.452         Condition No.:           63
==================================================================
```

The Regression Model


```
# Regression
from statsmodels.formula.api import ols

model1 = ols('cnt ~ temp_celsius', data=wbr).fit()
model1.summary2()
```

```
                 Results: Ordinary least squares
=====================================================================
Model:              OLS              Adj. R-squared:     0.393
Dependent Variable: cnt              AIC:                12777.5357
Date:               2019-12-11 12:23 BIC:                12786.7245
No. Observations:   731              Log-Likelihood:     -6386.8
Df Model:           1                F-statistic:        473.5
Df Residuals:       729              Prob (F-statistic): 2.81e-81
R-squared:          0.394            Scale:              2.2783e+06
---------------------------------------------------------------------
                 Coef.   Std.Err.    t     P>|t|    [0.025    0.975]
---------------------------------------------------------------------
Intercept      1214.6421 161.1635  7.5367 0.0000 898.2421 1531.0421
temp_celsius    161.9685   7.4436 21.7594 0.0000 147.3551  176.5820
---------------------------------------------------------------------
Omnibus:              20.477        Durbin-Watson:           0.468
Prob(Omnibus):        0.000         Jarque-Bera (JB):        12.566
Skew:                 0.167         Prob(JB):                0.002
Kurtosis:             2.452         Condition No.:           63
=====================================================================
```

The Regression Model

EDEM Centro Universitario

```python
# Regression
from statsmodels.formula.api import ols

model1 = ols('cnt ~ temp_celsius', data=wbr).fit()
model1.summary2()
```

```
                 Results: Ordinary least squares
==================================================================
No. Observations:   731              Log-Likelihood:     -6386.8
R-squared:          0.394            Scale:              2.2783e+06
------------------------------------------------------------------
                 Coef.    Std.Err.     t     P>|t|   [0.025    0.975]
------------------------------------------------------------------
Intercept      1214.6421 161.1635  7.5367 0.0000 898.2421 1531.0421
temp_celsius    161.9685   7.4436 21.7594 0.0000 147.3551  176.5820
------------------------------------------------------------------
==================================================================
```

Alberto Sanz, Ph.D. asanz@edem.es

The Regression Model


```python
# Regression
from statsmodels.formula.api import ols

model1 = ols('cnt ~ temp_celsius', data=wbr).fit()
model1.summary2()
```

```
                     Results: Ordinary least squares
====================================================================
No. Observations:    731              Log-Likelihood:      -6386.8
R-squared:           0.394            Scale:               2.2783e+06
--------------------------------------------------------------------
                 Coef.    Std.Err.     t     P>|t|   [0.025    0.975]
--------------------------------------------------------------------
Intercept      1214.6421  161.1635  7.5367  0.0000  898.2421 1531.0421
temp_celsius    161.9685    7.4436 21.7594  0.0000  147.3551  176.5820
--------------------------------------------------------------------
====================================================================
```

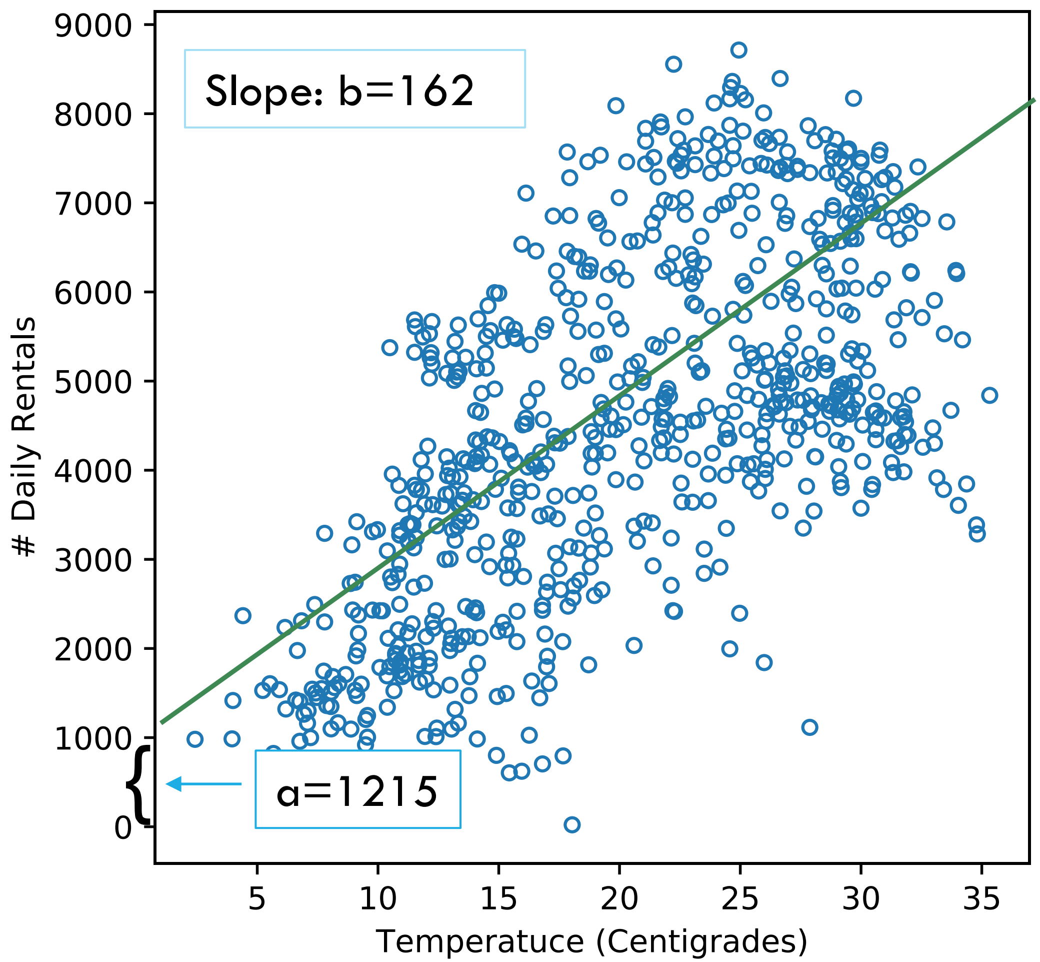
The Regression Model

```
# Regression
from statsmodels.formula.api import ols

model1 = ols('cnt ~ temp_celsius', data=wbr).fit()
model1.summary2()
```

```
                    Results: Ordinary least squares
==================================================================
No. Observations:    731              Log-Likelihood:     -6386.8
R-squared:           0.394            Scale:              2.2783e+06
------------------------------------------------------------------
                 Coef.    Std.Err.    t      P>|t|   [0.025    0.975]
------------------------------------------------------------------
Intercept      1214.6421 161.1635  7.5367  0.0000 898.2421 1531.0421
temp_celsius    161.9685   7.4436 21.7594  0.0000 147.3551  176.5820
------------------------------------------------------------------
==================================================================
```

The Regression Model

Figure 9. Daily bicycle rentals, by temperature.

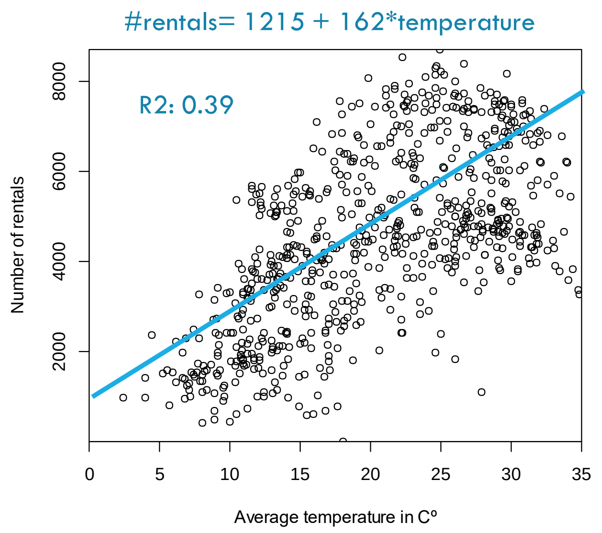



```
Results: Ordinary least squares
===================================
No. Observations:    731
R-squared:               0.394
-----------------------------------
                  Coef.       P>|t|
-----------------------------------
Intercept      1214.6421    0.0000
temp_celsius    161.9685    0.0000
-----------------------------------
```

Y = a + b*x

#rentals= 1215 + 162 * temperature

Conclussion



Conclussion:

As P. Value < 0.000

We can reject H0 with a confidence higer tan 99.9

Using (**only**) temperature as predictors, we can anticipate as much as **40 % of the variability in bike rentals!!!!**

❌ H0.: There is no linear association between the *number of rentals* and the *temperature*. (the slope of the regression line=0).

✓ H1.:There is a linear association between the *number of rentals* and the *temperature*. (the slope of the regression line ≠ 0).

The Regression Model


```
model1  = ols('cnt ~ temp_celsius', data=wbr).fit()
model1b = ols('cnt ~ windspeed_kh', data=wbr).fit()
print(model1b.summary2())
```

```
                  Results: Ordinary least squares
==================================================================
Model:              OLS              Adj. R-squared:     0.054
Dependent Variable: cnt              AIC:                13102.0108
Date:               2019-12-11 15:56 BIC:                13111.1996
No. Observations:   731              Log-Likelihood:     -6549.0
Df Model:           1                F-statistic:        42.44
Df Residuals:       729              Prob (F-statistic): 1.36e-10
R-squared:          0.055            Scale:              3.5512e+06
------------------------------------------------------------------
                Coef.   Std.Err.    t    P>|t|    [0.025    0.975]
------------------------------------------------------------------
Intercept     5621.1529 185.0624 30.3744 0.0000 5257.8341 5984.4717
windspeed_kh   -87.5062  13.4327 -6.5144 0.0000 -113.8775  -61.1348
------------------------------------------------------------------
Omnibus:               45.655        Durbin-Watson:           0.350
Prob(Omnibus):         0.000         Jarque-Bera (JB):        17.090
Skew:                  -0.026        Prob(JB):                0.000
Kurtosis:              2.253         Condition No.:           37
==================================================================
```

The Multiple Regression Model


```
model1 = ols('cnt ~ temp_celsius', data=wbr).fit()
model2 = ols('cnt ~  temp_celsius  + windspeed_kh', data=wbr).fit()
print(mode2.summary2())
```

```
 Results: Ordinary least squares
====================================================================
Model:              OLS              Adj. R-squared:     0.411
Dependent Variable: cnt              AIC:                12756.4931
Date:               2019-12-11 16:03 BIC:                12770.2763
No. Observations:   731              Log-Likelihood:     -6375.2
Df Model:           2                F-statistic:        255.6
Df Residuals:       728              Prob (F-statistic): 7.99e-85
R-squared:          0.413            Scale:              2.2106e+06
--------------------------------------------------------------------
                Coef.    Std.Err.    t     P>|t|    [0.025    0.975]
--------------------------------------------------------------------
Intercept     1991.0459 225.9615  8.8114 0.0000 1547.4319 2434.6599
temp_celsius   156.3058   7.4254 21.0500 0.0000  141.7279  170.8836
windspeed_kh   -51.8225  10.7328 -4.8284 0.0000  -72.8934  -30.7515
--------------------------------------------------------------------
Omnibus:              25.144       Durbin-Watson:          0.467
Prob(Omnibus):        0.000        Jarque-Bera (JB):       15.379
Skew:                 0.206        Prob(JB):               0.000
Kurtosis:             2.422        Condition No.:          102
====================================================================
```

Models of increasing complexity


```
model1 = ols('cnt ~ temp_celsius', data=wbr).fit()

model2 = ols('cnt ~  temp_celsius  + windspeed_kh ,data=wbr).fit()

model3 = ols('cnt ~ temp_celsius +  windspeed_kh + hum'
            , data=wbr).fit()

model4 = ols('cnt ~ temp_celsius +  windspeed_kh + hum + workingday'
            , data=wbr).fit()
```

The Multiple Regression Model


```
model4 = ols('cnt ~ temp_celsius +  windspeed_kh + hum + workingday'
          , data=wbr).fit()

print(model4.summary2())
```

```
====================================================================
Model:              OLS              Adj. R-squared:     0.459
Dependent Variable: cnt              AIC:                12696.4930
Date:               2019-12-11 16:20 BIC:                12719.4650
No. Observations:   731              Log-Likelihood:     -6343.2
Df Model:           4                F-statistic:        155.7
Df Residuals:       726              Prob (F-statistic): 3.61e-96
R-squared:          0.462            Scale:              2.0309e+06
--------------------------------------------------------------------
                Coef.   Std.Err.    t    P>|t|    [0.025    0.975]
--------------------------------------------------------------------
Intercept     4009.3688 344.5244 11.6374 0.0000 3332.9858 4685.7517
temp_celsius   161.2124   7.1558 22.5289 0.0000  147.1639  175.2609
windspeed_kh   -71.6672  10.5792 -6.7743 0.0000  -92.4367  -50.8976
hum            -31.0683   3.8398 -8.0911 0.0000  -38.6067  -23.5299
workingday     125.8049 113.5505  1.1079 0.2683  -97.1217  348.7315
--------------------------------------------------------------------
Omnibus:              10.037       Durbin-Watson:          0.404
Prob(Omnibus):        0.007        Jarque-Bera (JB):       7.868
Skew:                 0.160        Prob(JB):               0.020
Kurtosis:             2.604        Condition No.:          449
====================================================================
```

Models of increasing complexity

EDEM Centro Universitario


```
model1 = ols('cnt ~ temp_celsius', data=wbr).fit()

model2 = ols('cnt ~  temp_celsius  + windspeed_kh ,data=wbr).fit()

model3 = ols('cnt ~ temp_celsius +  windspeed_kh + hum'
              , data=wbr).fit()

model4 = ols('cnt ~ temp_celsius +  windspeed_kh + hum + workingday'
            , data=wbr).fit()

#Stargazer
#!pip install stargazer
from stargazer.stargazer import Stargazer

Stargazer([model1,model2,model3,model4]).render_html()
```

Tip: Visit https://pypi.org/project/stargazer/ for stargazer functionalities
Visit https://github.com/mwburke/stargazer/blob/master/examples.ipynb for use examples

Tip: Stargezer will output HTML code. 1) You can render it into a nice (and editable) table in:
https://htmledit.squarefree.com/
Or 2) You can save the code in a plain text document with .html extensión and read it in word

Model reporting with Stargazer

EDEM
Centro Universitario

Table 1. Models of number of daily bicycle rentals in Washington D.C.

	Model 1	Model 2	Model 3	Model 4
Temperature Cº	162.0***	156.3***	161.6***	161.2***
	(7.4)	(7.4)	(7.1)	(7.2)
Windspeed_k/h		-51.8***	-71.7***	-71.7***
		(10.7)	(10.6)	(10.6)
Humidity			-31.0***	-31.1***
			(3.8)	(3.8)
Workingday (0/1)				125.8
				(113.6)
Intercept	1214.6***	1991.0***	4084.4***	4009.4***
	(161.2)	(226.0)	(337.9)	(344.5)
Observations	731	731	731	731
R^2	0.4	0.4	0.5	0.5
Note:				*p<0.1; **p<0.05; ***p<0.01

Models of increasing complexity

EDEM Centro Universitario

	Dependent Variable: Number of bicycle rentals in Washington				
	Model 1	Model 2	Model 3	Model 4	Model 6
Temperature in Cº	161.969***	156.306***	161.598***	161.212***	646.078***
	(7.444)	(7.425)	(7.148)	(7.156)	(38.263)
Temperature in Cº squared					-12.022***
					(0.935)
Windspeed (Km/h)		-51.822***	-71.745***	-71.667***	-85.550***
		(10.733)	(10.581)	(10.579)	(9.614)
Humidity (in %)			-31.001***	-31.068***	-42.666***
			(3.840)	(3.840)	(3.583)
Working day (0:No, 1:Yes)				125.805	85.370
				(113.551)	(102.588)
Constant	1,214.642***	1,991.046***	4,084.363***	4,009.369***	730.179*
	(161.164)	(225.962)	(337.862)	(344.524)	(402.305)
Observations	731	731	731	731	731
R^2	0.394	0.413	0.461	0.462	0.562
Note:					*p<0.1 **p<0.05 ***p<0.01

Regression I. Summing UP

EDEM Centro Universitario

1. **Always DESCRIBE** the variables involved in the regression model separately. Check and validate the integrity of the data prior to any analysis.
2. **EXPLORE** of bivariate relation: **Scatterplot / Pearson's r**
3. **Fit your linear regression model carefully.** Pay attention to:
   a) **Slope & intercept**
   b) **P. value**
   c) **Model fit**
   d) **Sample size**

Questions?

Thank you !

Alberto Sanz
asanz@edem.es