EDEM

Centro Universitario

Regression II:

Multiple linear regression

Sessions 11-14

Programación Estadística con Python

Alberto Sanz, Ph.D

asanz@edem.es

MASTER EN DATA ANALYTICS PARA LA EMPRESA

Goals (+ to be developed)

EDEM Centro Universitario

- Multiple linear regression.
  - Comprehensive models
  - Gaining control over alternative explanations: *Ceteris paribus*

- Dealing with the qualitative in regression models (I)
  - Interpretation of coefficients from dichotomies

- Dealing with non linearity
  - Curvilinear modeling under linear regression

- Graphic methods in multiple linear regression:
  - Coefficent plot
  - Confidence interval plot

Multiple Regression

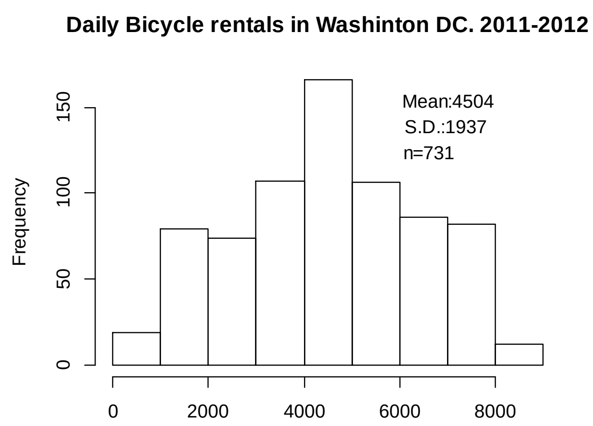
1. **Always DESCRIBE** the variables involved in the regression model separately. Check and validate the integrity of the data prior to any analysis.
2. **EXPLORE** bivariate relation: **Scatterplot / Pearson's r / Simple Regression**
3. **Fit your multiple linear regression model carefully. Attend to:**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

Research Question

EDEM
Centro Universitario

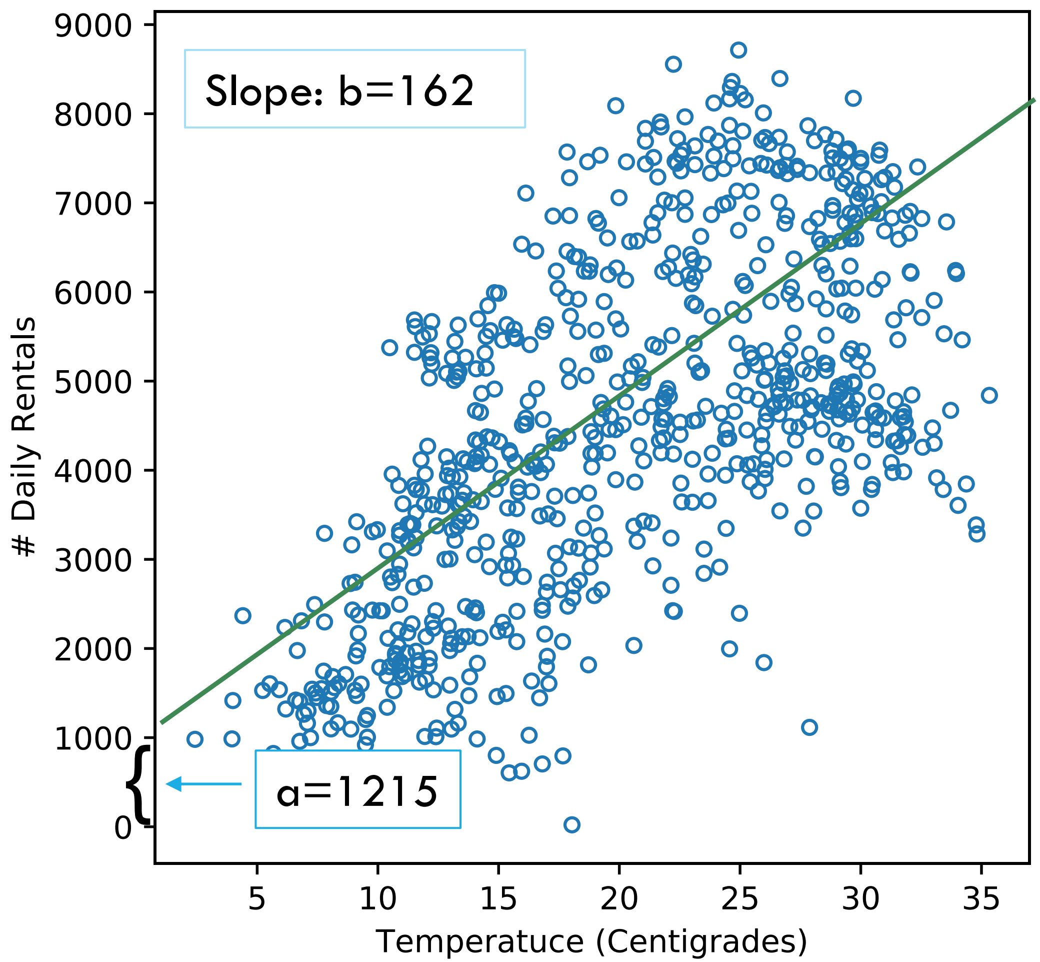
Why some days are rent *more* bikes?

- Temperature ?
- Windspeed ?
- Humidity ?
- Holiday ?
- ...



The (simple) 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

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
====================================================================
```

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

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()
```

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
====================================================================
```

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

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

	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

Basic diagnostics for simple LM

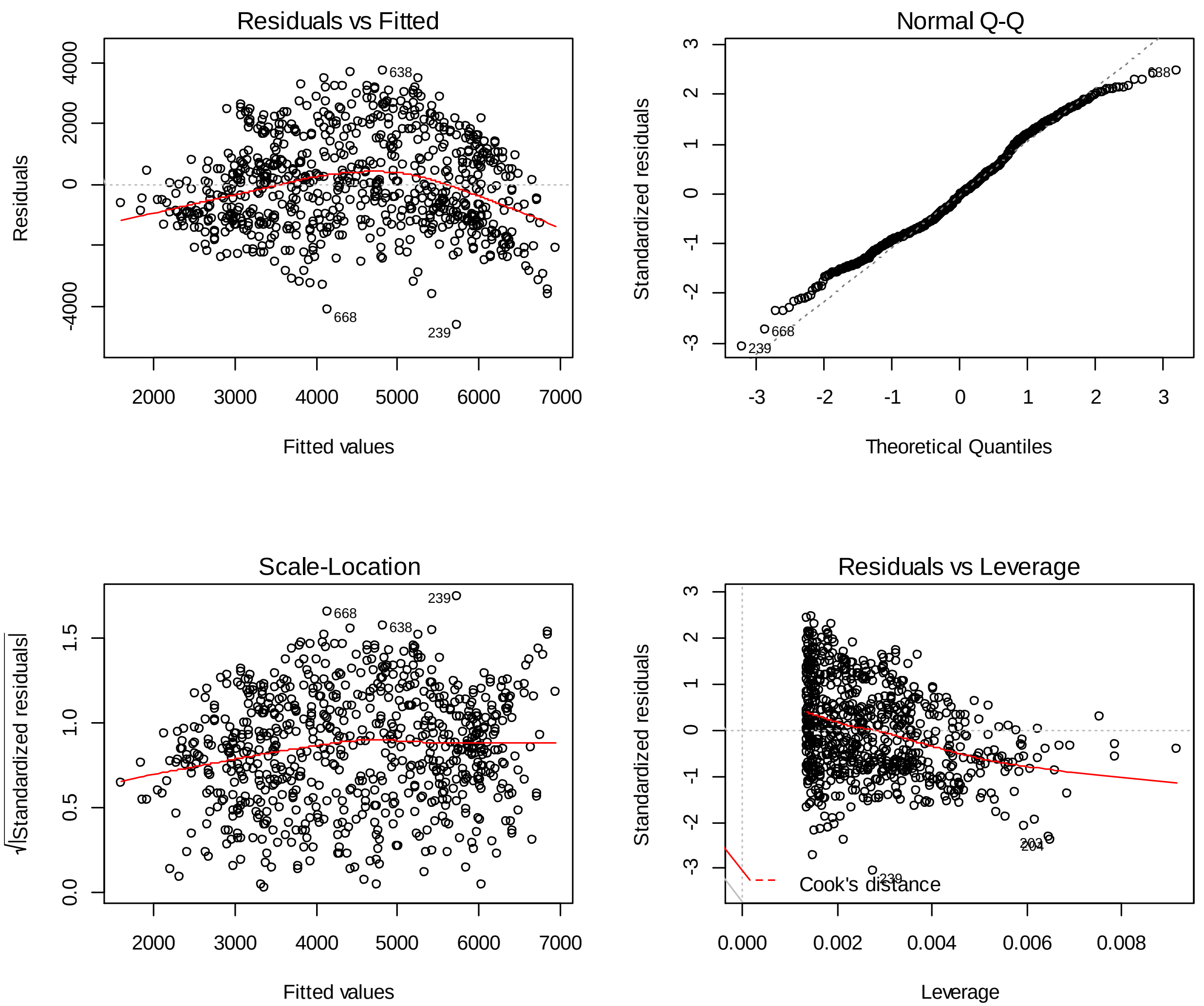
- Linear relation:
  - Check Residuals vs. Fitted (or predicted) plot
- Normality of residuals:
  - Check Normal Q-Q plot
- Homocedasticity:
  - Check Scale-Location plot

- No influential observations
  - Check the Residuals vs. Leverage plot

Basic diagnostics for simple LM

Code available at:

https://robert-alvarez.github.io/2018-06-04-diagnostic_plots/



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**
   e) **Model Diagnostics**

Regression II. 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**
   e) **Model Diagnostics**

Questions?

Thank you !

Alberto Sanz
asanz@edem.es