COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI ICADE CIHS

Machine Learning I

- Prediction: Multiple Lineal Regression

comillas.edu

Multiple Linear Regression Statement

- The Multiple Linear Regression model allows estimate the continuous random variable Y from a set of $\boldsymbol{p}$ input variables

Deterministic component ($\boldsymbol{p}+\mathbf{1}$ coefficients)

Random component

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

HYPERPLANE
(unknown theoretical relation)

Sample

$(x_1, y_1) \ldots (x_N, y_N)$

$x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$

$$\epsilon \approx N(0, \sigma^2)$$

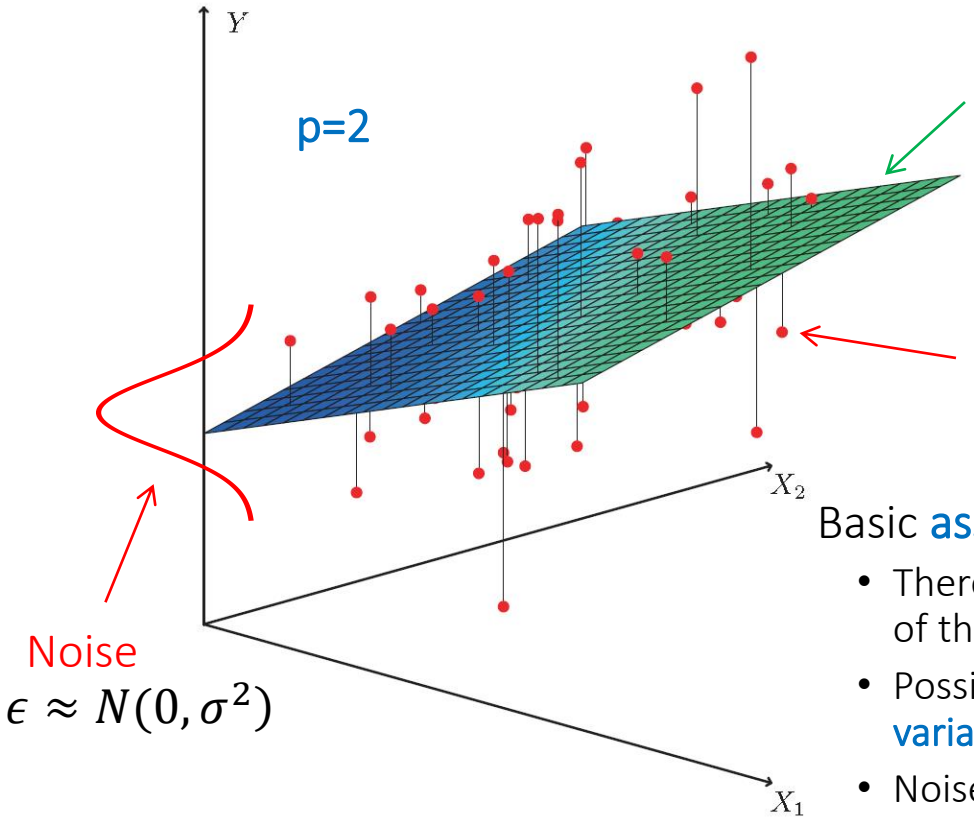
Random variable that considers the output variations with respect to the expected value of the deterministic component (error, noise)

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

ESTIMATED REGRESSION **HYPERPLANE**

Multiple Linear Regression Statement

- The Multiple Linear Regression model assumes a theoretical functional relation with these basic characteristics



(Hyper)plane

$$\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Cloud of points

$$(x_1, y_1) \ldots (x_N, y_N)$$

Noise

$$\epsilon \approx N(0, \sigma^2)$$

Basic assumptions

- There is a linear relation between Y and the rest of the variables
- Possible noise has zero mean and constant variance
- Noise follows a normal distribution
- Noise has no structure, there is independence between errors

Multiple Linear Regression Coefficient estimation

- Least squares approach (the common one):
  - The coefficients to determine will be those that minimize the sum of the squared errors

$$\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad \hat{y}_i \text{ is the estimation of } y_i$$

$$= \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2$$

  - RSS is a quadratic function for all the coefficients, with a clear minimum if the problem is well defined (there is a linear relation)
  - Deriving with respect to each parameter and setting equal to zero we obtain the analytical expression of the estimated coefficients minimizing RSS

MATRIX NOTATION

$$\frac{\partial \text{RSS}}{\partial \beta} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) \Rightarrow \mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) = 0$$

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

$\boldsymbol{X} \in R^{n\times(p+1)}$

$\boldsymbol{y} \in R^{n}$

$\boldsymbol{\beta} \in R^{p+1}$

Multiple Linear Regression Accuracy in the coefficient estimation

- Standard error of the estimators
  - It is possible to determine theoretically the joint distribution of the parameter estimators
  - From it, we can determine confidence intervals and do the hypothesis test
  - Usually it is expressed in matrix notation

Variance-covariance matrix:

$$\mathrm{Var}(\hat{\beta}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$$

Noise variance affects the estimator accuracy

$$\hat{\beta} \sim N(\beta, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2)$$

Normal multivariate with variance-covariance matrix

Centered in the theoretical value

Multiple Linear Regression Accuracy in the coefficient estimation

- Standard error of the estimators
  - Example

$$\hat{\beta} \sim N(\beta, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2)$$

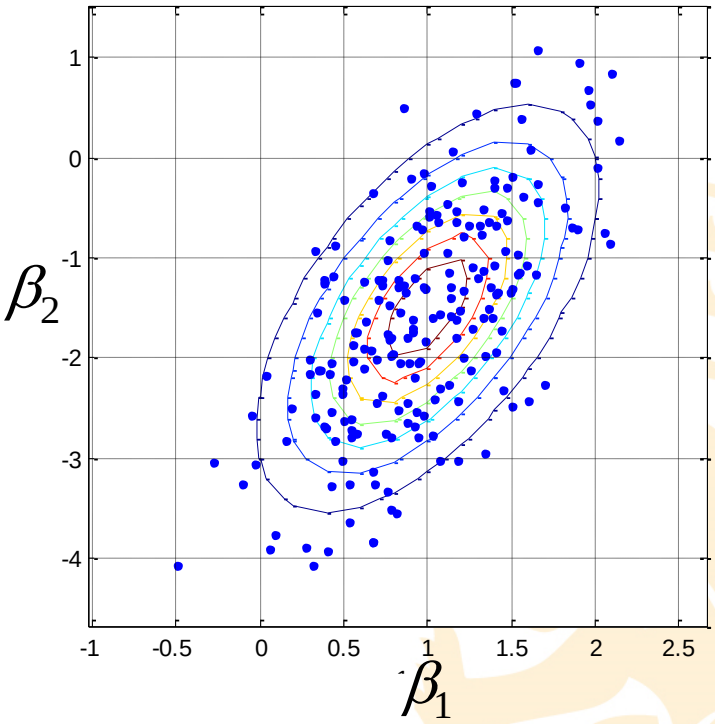
$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 1.0 \\ -1.5 \end{pmatrix}$$

  - Variance-covariance matrix

$$\sum = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\sigma^2 = \begin{pmatrix} 0.25 & 0.3 \\ 0.3 & 1.0 \end{pmatrix}$$

  - Correlation coefficient

$$\rho_{x1x2} = \frac{cov_{x1x2}}{\sqrt{var_{x1}}\sqrt{var_{x2}}} = \qquad \rho = \frac{0.3}{\sqrt{0.25 \cdot 1}} = 0.6$$



Multiple Linear Regression
Goodness of fit

- Some important concepts
  - Residual variance

$$S_{RES}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-(p+1)}$$

p is the number of independent variables

  - Root Mean Squared Error

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}}$$

  - Coefficient of determination $\boldsymbol{R^2}$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Insensitive to the number of variables

Multiple Linear Regression Goodness of fit

- Adjusted coefficient of determination
  - Specific for multivariate regression
  - Improves the coefficient of determination R^2 because includes a penalty associated to the model complexity
    - Fixes the natural decreases that happens in R^2 when increasing the number of input variables due to possible overfitting (over-parametrization of the model)
  - Bounded between 0 (a large dispersion) and 1 (data in a hyperplane)
    - Interpreted as R^2

$$R^2_{AJU} = 1 - \left(1 - R^2\right) \frac{n-1}{n-p-1}$$

Residual Variance

- Residual variance
  - The variability of the dependent variable Y can be decomposed in two components: variability *explained by the model* and *variability not explained* by the model due to random factors

  - The variability of the dependent variable will be:

$$SCT = n * \sigma^2 = \sum_{1}^{n} (y_i - \bar{y})^2$$

  - Adding and substracting the value estimated by the model:

$$(SCT)\sum_1^n (y_i - \bar{y})^2 = (SCR)\sum_1^n (y_i - \hat{y}_i)^2 + (SCE)\sum_1^n (\hat{y}_i - \bar{y})^2$$

SCR is the residual variability (not explained by the model)
SCE is the variability explained by the model

Residual Variance

- Residual variance
  - The variance of the variable Y is obtained dividing SCT by the freedom degrees (n-1) = $\sum_1^n (y_i - \bar{y})^2$/(n-1)
  - The variance of the explained variability is SCE divided by the freedom degrees (p): $\sum_1^n (\hat{y}_i - \bar{y})^2$ /p = SCE/p
  - The variance of the not explained variability is SCR divided by the freedom degrees (n-p-1): $\sum_1^n (y_i - \hat{y}_i)^2$ /(n-p-1) =SCR/(n-p-1)

Multiple Linear Regression
Is model significant?

- Is the model really a constant?
  - It can be stated as a hypothesis test

If all the coefficients (except the constant) can be zero, then the model is useless (there is no linear relation, no input variable is significant)

$$H_0: \beta_1 = \cdots = \beta_p = 0$$
$$H_1: \exists i / \beta_i \neq 0 \text{ (at least one)}$$

  - If the null hypothesis is true, then the ratio between the explained variance and the non explained variance should be similar and equal to 1 (if H_0 is true, variances should be similar)

  - If the residuals follow a normal distribution:

$$\frac{SCT}{n-1} \approx \chi_{n-1}\,, \frac{SCE}{p} \approx \chi_{p-1}\,, \qquad \frac{SCR}{n-p-1} \approx \chi_{n-p-1}$$

Multiple Linear Regression
Is model significant?

- Is the model really a constant?
  - Then the ratio between the explained variance and the non explained variance can be expressed as follows:

$$\frac{\frac{SCE}{p}}{\frac{SCR}{n-p-1}} = \frac{\chi_{p-1}}{\chi_{n-p-1}} = F_{p,n-p-1}$$

  - Being the F distribution, it is possible to assign a probability value (p-value) to the hypothesis of similarity between variance explained and not explained.

Multiple Linear Regression
Is model significant?

- Can a specific input variable be deleted?
  - If the confidence interval of the associated coefficient contains zero, then it is superfluous in the model
  - It can be stated as a hypothesis test

$$H_0 : \beta_i = 0$$
$$H_1 : \beta_i \neq 0$$

If this parameter is superfluous, it must be eliminated from the model and the fit repeated

  - It is possible to demonstrate that:

$$\frac{\hat{\beta}_i - \beta_i}{SE(\hat{\beta}_i)} \approx t_{n-p-1}$$

SE is the standard deviation of the coefficient:

$$SE(\hat{\beta}_i) = \sqrt{S_{RES}{}^2} \times \sqrt{Var(\hat{\beta}_i)}$$

  - If the null hypothesis is true (β_i=0), then

$$\frac{\hat{\beta}_i}{SE(\hat{\beta}_i)} \approx t_{n-p-1}$$

Multiple Linear Regression
Interpretation of the fitted model

- The coefficients show the marginal contribution of each variable to the output (assuming the others do not change)
- Usual information for diagnosis and model interpretation as a table
  - The joint $\boldsymbol{F}$ test and the adjusted $\boldsymbol{R^2}$
- Example with two input variables

```
Linear regression model:
    y ~ 1 + x1 + x2
Estimated Coefficients:
                   Estimate       SE       tStat       pValue
                   ________    ________    ______    __________

    (Intercept)     10.044     0.033416    300.57            0
    x1              1.9354     0.028088    68.904            0
    x2             0.98588     0.018831    52.353    2.32e-288

Number of observations: 1000, Error degrees of freedom: 997
Root Mean Squared Error: 0.877
R-squared: 0.898,  Adjusted R-Squared 0.898
F-statistic vs. constant model: 4.38e.03, p-value = 0
```

Multiple Linear Regression
Residual analysis

- Once the model is fitted it is important to check that residuals, what it is not explained by the model, satisfy reasonably the basic assumptions
  - Normality: follow a normal distribution with zero mean and constant variance
  - Independence: are independent among them, no internal clear structure

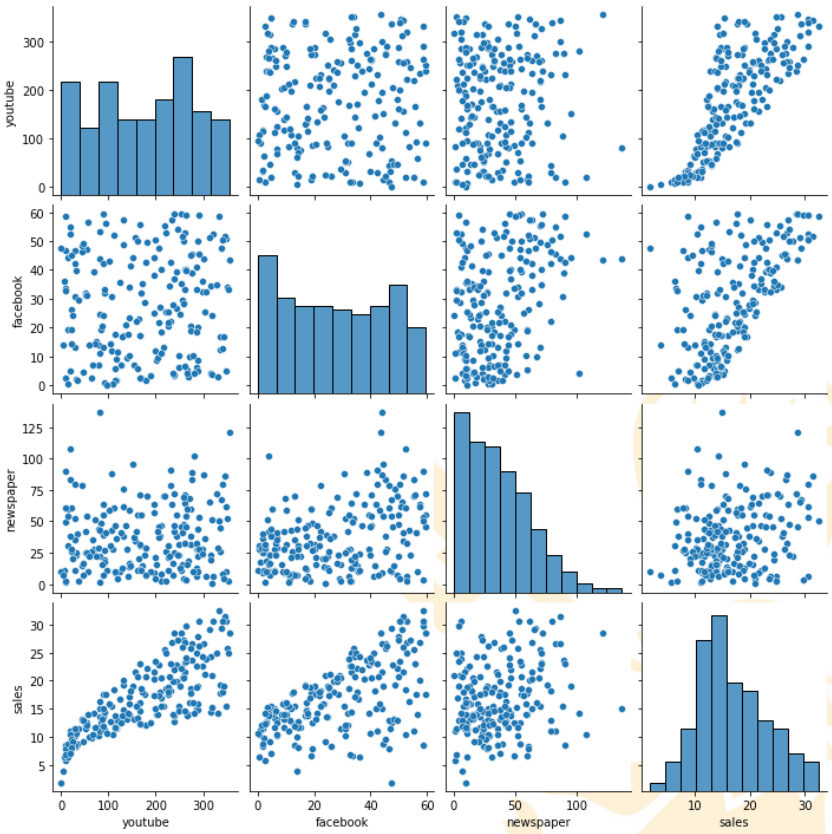
- Diagnosis of the model or residual analysis
  - Normality plot for the residuals (non parametric test)
  - Plot the residuals versus the estimated values
  - Independence plot among residuals, considering the different inputs

Multiple Linear Regression. Cases

- MLR-1: Marketing dataset

```
# Data Loading
marketing = pd.read_csv('marketing.csv')
marketing.head()
```

Data knowledge is required before any analysis: plots, NaN/null detection, correlations, basic statistics, etc.



	youtube	facebook	newspaper	sales
count	200.000000	200.000000	200.000000	200.000000
mean	176.451000	27.916800	36.664800	16.827000
std	103.025084	17.816171	26.134345	6.260948
min	0.840000	0.000000	0.360000	1.920000
25%	89.250000	11.970000	15.300000	12.450000
50%	179.700000	27.480000	30.900000	15.480000
75%	262.590000	43.830000	54.120000	20.880000
max	355.680000	59.520000	136.800000	32.40000

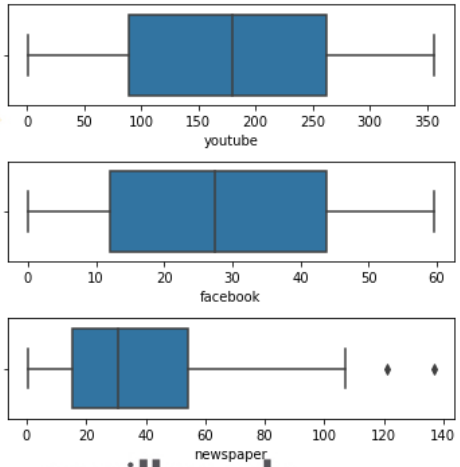
Multiple Linear Regression. Cases

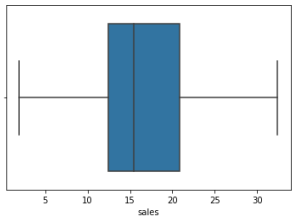
- MLR-1: Marketing dataset

```
# Cheking null values
marketing.isnull().sum()*100/marketing.shape[0]
```

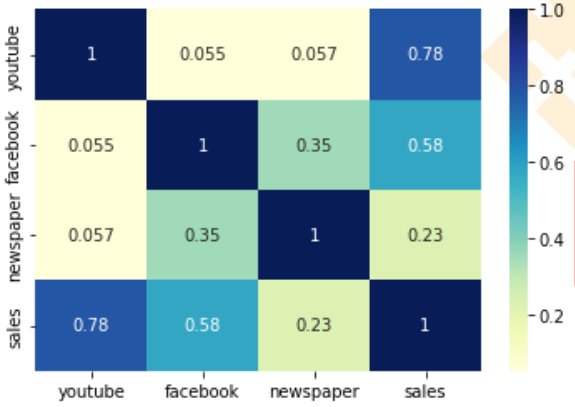
```
youtube      0.0
facebook     0.0
newspaper    0.0
sales        0.0
dtype: float64
```

```
# Outliers analysis
fig, axs = plt.subplots(3, figsize = (5,5))
plt1 = sns.boxplot(marketing['youtube'], ax = axs[0])
plt2 = sns.boxplot(marketing['facebook'], ax = axs[1])
plt3 = sns.boxplot(marketing['newspaper'], ax = axs[2])
plt.tight_layout()
```

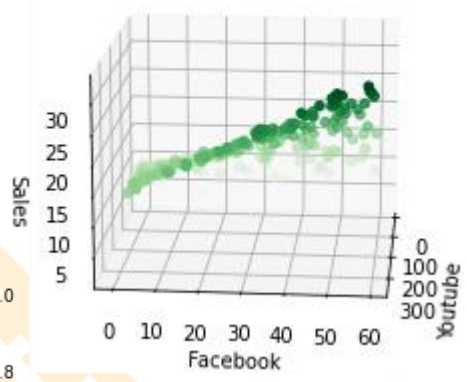




Correlations



Sales vs youtube & facebook



A linear relation between Sales and *Youtube* + *Facebook* seems to exist

Multiple Linear Regression. Cases

- MLR-1: Marketing dataset. MODEL

```
import statsmodels.api as sm
X = marketing[['youtube','facebook','newspaper']]
y = marketing['sales']

df= marketing[['youtube','facebook','newspaper', 'sales']]

# Add a constant to get an intercept
X_sm = sm.add_constant(X)

# Fit the resgression line using 'OLS'
model = sm.OLS(y, X_sm).fit()
```

Initial model

Multiple Linear Regression. Cases

- MLR-1: Marketing dataset. MODEL

Initial model

OLS Regression Results

Dep. Variable:	sales	R-squared:	0.897
Model:	OLS	Adj. R-squared:	0.896
Method:	Least Squares	F-statistic:	570.3
Date:	Mon, 12 Jun 2023	Prob (F-statistic):	1.58e-96
Time:	14:39:03	Log-Likelihood:	-422.65
No. Observations:	200	AIC:	853.3
Df Residuals:	196	BIC:	866.5
Df Model:	3		
Covariance Type:	nonrobust		

Confidence intervals (95%)

Coefficients estimated

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3.5267	0.374	9.422	0.000	2.789	4.265
youtube	0.0458	0.001	32.809	0.000	0.043	0.049
facebook	0.1885	0.009	21.893	0.000	0.172	0.206
newspaper	-0.0010	0.006	-0.177	0.860	-0.013	0.011

Omnibus:	60.414	Durbin-Watson:	2.084
Prob(Omnibus):	0.000	Jarque-Bera (JB):	151.241
Skew:	-1.327	Prob(JB):	1.44e-33
Kurtosis:	6.332	Cond. No.	545.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

High p-value

- p-value of ***F test*** very low, this means that at least one explanatory variable is significant to explain ***Sales***
- Adjusted R^2 is high, good fit to the data
- High p-value in coefficient $newspaper$ that could be superfluous (with a 95% we should delete)

This coefficient can be 0
If β=0 this variable is not explaining

Multiple Linear Regression. Cases

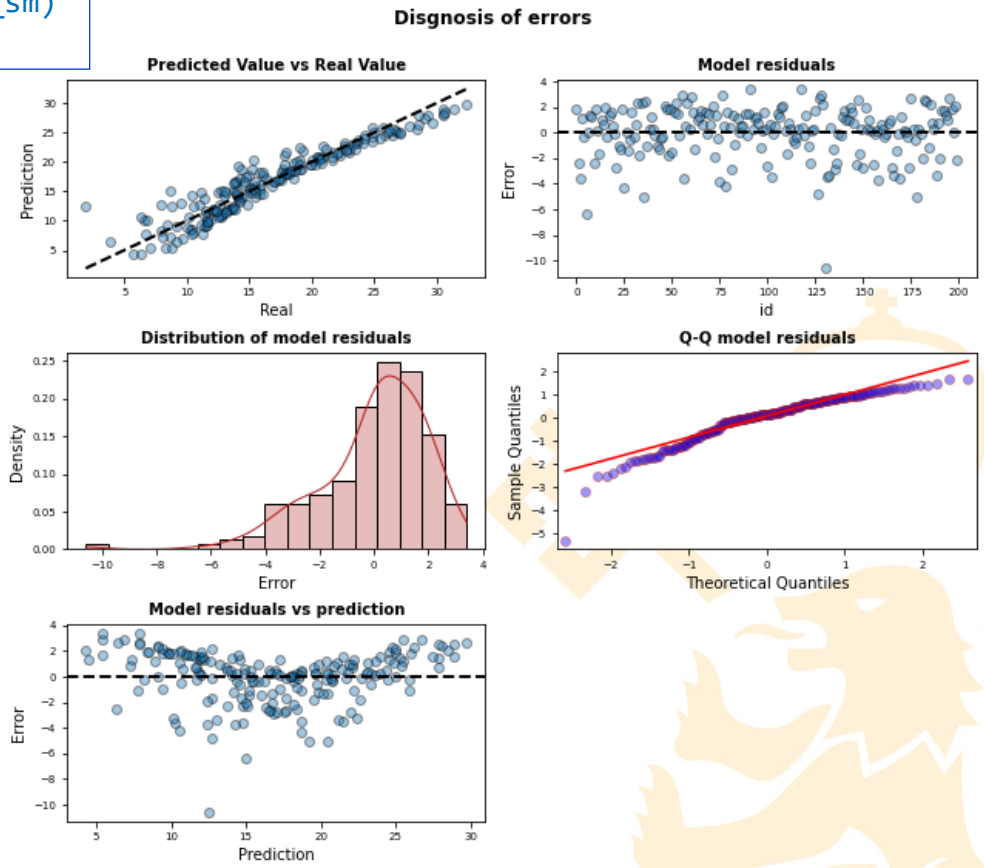
Initial model

- MLR-1: Marketing dataset (Analysis of residuals I)

```
# Estimation error
# ========================================
prediction_train = model.predict(exog = X_sm)
residuals_train   = y-prediction_train
```

- Residuals seem normal, although in the right tail something seems remarkable but reasonable
- Residuals seem independent of the estimated value

Residual analysis do not question the initial model



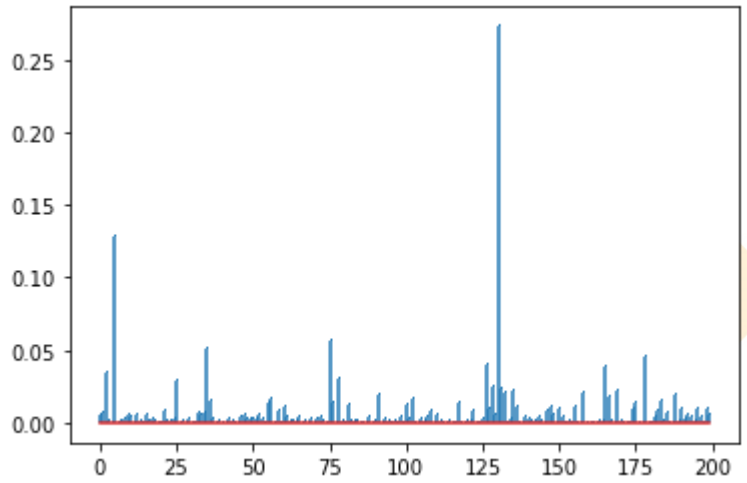
Multiple Linear Regression. Cases

Initial model

• MLR-1: Marketing dataset (Analysis of residuals II)

- Q-Q plot for testing normality. It compares two probability distributions. A point (x, y) on the plot corresponds to one of the quantiles of the second distribution (y-coordinate) plotted against the same quantile of the first distribution (x-coordinate). Comparison with a Normal distribution (The line in the plot is the ideal case)
- Cook's distance. Large residuals (outliers) and/or high leverage may distort the outcome and accuracy of a regression. Cook's distance measures the effect of deleting a given observation. Points with a large Cook's distance are considered to merit closer examination in the analysis.

```
#' ## Cooks distance

influence = model.get_influence()
#c is the distance and p is p-value
(c, p) = influence.cooks_distance

#PLotting Cook distance:
plt.stem(np.arange(len(c)), c, markerfmt=",")
```



Residual analysis do not question the initial model

Multiple Linear Regression. Cases

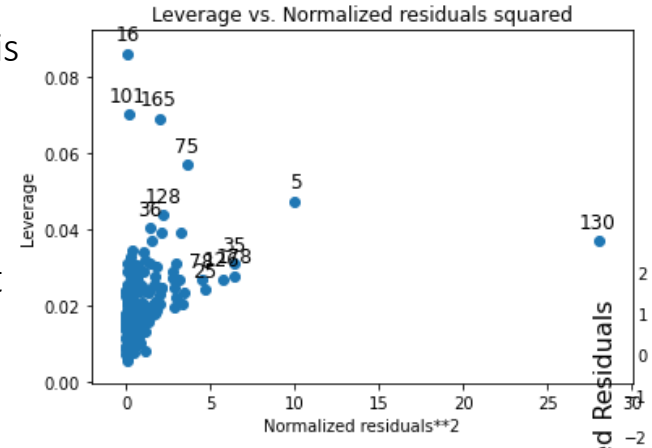
Initial model

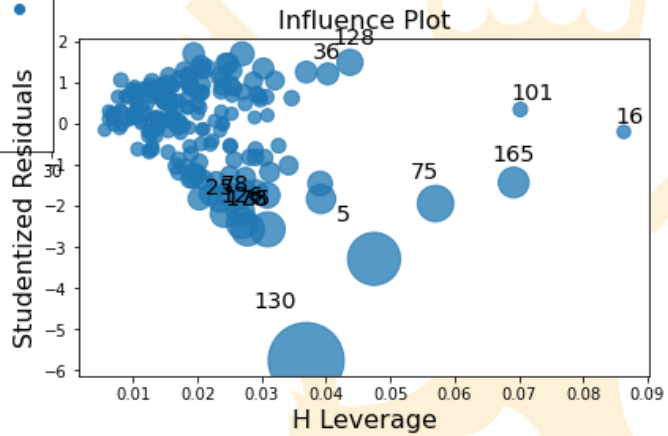
• MLR-1: Marketing dataset (Analysis of residuals III)

- Scale-Location plot. It takes the square root of the absolute residuals in order to diminish skewness (√|E| is much less skewed than |E|

```
# Leverage plots
from statsmodels.graphics.regressionplots import *
plot_leverage_resid2(model)

fig = sm.graphics.influence_plot(model, criterion="cooks")
fig.tight_layout(pad=1.0)
```

- Residual-Leverage plot. Leverage is a measure of how far away the residual of an observation is from those of the other observations. Also this plot shows contours of equal Cook's distance, (by default 0.5 and 1)..





Residual analysis do not question the initial model

Multiple Linear Regression. Cases

- MLR-1: Marketing dataset. SIMPLIFIED MODEL I

```
#Removing variable newspaper
Xnew = marketing[['youtube','facebook']]
ynew = marketing['sales']
# Add a constant to get an intercept
Xnew_sm = sm.add_constant(Xnew) # Fit the resgression line using 'OLS'
modelnew = sm.OLS(ynew, Xnew_sm).fit()
```

Simplified model I

OLS Regression Results

Dep. Variable:	sales	R-squared:	0.897
Model:	OLS	Adj. R-squared:	0.896
Method:	Least Squares	F-statistic:	859.6
Date:	Thu, 15 Jun 2023	Prob (F-statistic):	4.83e-98
Time:	12:44:51	Log-Likelihood:	-422.66
No. Observations:	200	AIC:	851.3
Df Residuals:	197	BIC:	861.2
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.5053	0.353	9.919	0.000	2.808	4.202
youtube	0.0458	0.001	32.909	0.000	0.043	0.048
facebook	0.1880	0.008	23.382	0.000	0.172	0.204

Omnibus:	60.022	Durbin-Watson:	2.081
Prob(Omnibus):	0.000	Jarque-Bera (JB):	148.679
Skew:	-1.323	Prob(JB):	5.19e-33
Kurtosis:	6.292	Cond. No.	510.

- The simplified model is significant (overall F test higher),
- Now all its coefficients are significant
- RMSE is similar or lower
- Adjusted R^2 is similar to the previous one

Simplified model is better than the initial one

Multiple Linear Regression. Cases

- MLR-1: Marketing dataset. SIMPLIFIED MODEL II

```
#Removing outlier case 130
marketing_sim =marketing
marketing_sim=marketing_sim.drop(index=130)

#Removing variable newspaper
Xnews = marketing_sim[['youtube','facebook']]
ynews = marketing_sim['sales']
```

Simplified model II

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  sales   R-squared:                       0.909
Model:                            OLS   Adj. R-squared:                  0.908
Method:                 Least Squares   F-statistic:                     982.2
Date:                Thu, 15 Jun 2023   Prob (F-statistic):          7.18e-103
Time:                        12:54:45   Log-Likelihood:                -405.71
No. Observations:                 199   AIC:                             817.4
Df Residuals:                     196   BIC:                             827.3
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          3.6226      0.329     11.022      0.000       2.974       4.271
youtube        0.0448      0.001     34.452      0.000       0.042       0.047
facebook       0.1917      0.007     25.588      0.000       0.177       0.206
==============================================================================
Omnibus:                       22.143   Durbin-Watson:                   2.148
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               25.935
Skew:                          -0.851   Prob(JB):                     2.33e-06
Kurtosis:                       3.479   Cond. No.                         511.
==============================================================================
```

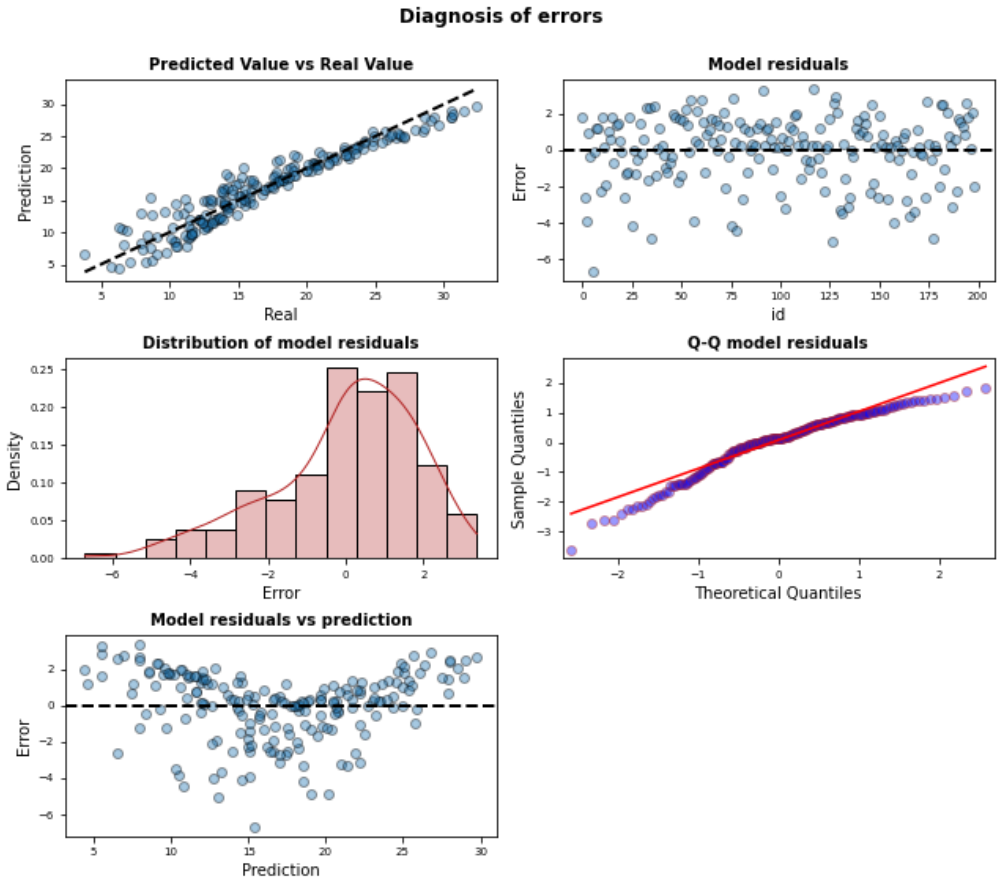
- The model has improved removing a big outlier

Simplified model II is even better than the initial one

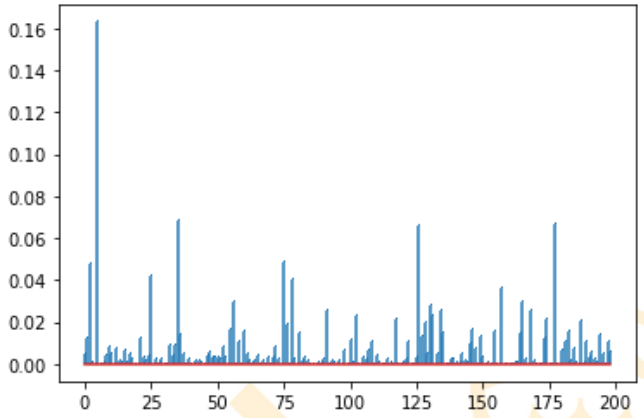
Multiple Linear Regression. Cases

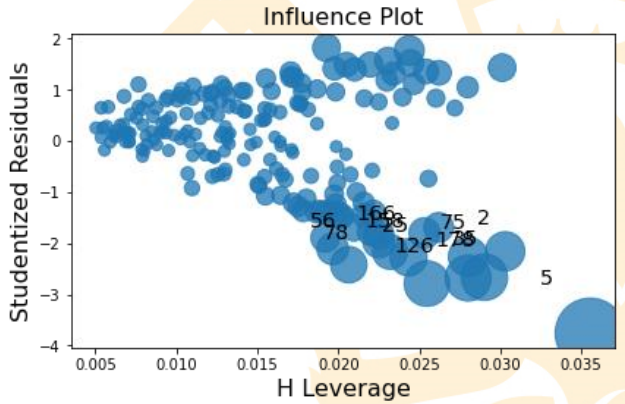
Simplified model II

- **MLR-1: Marketing dataset** (Analysis of residuals II)



Cook distances



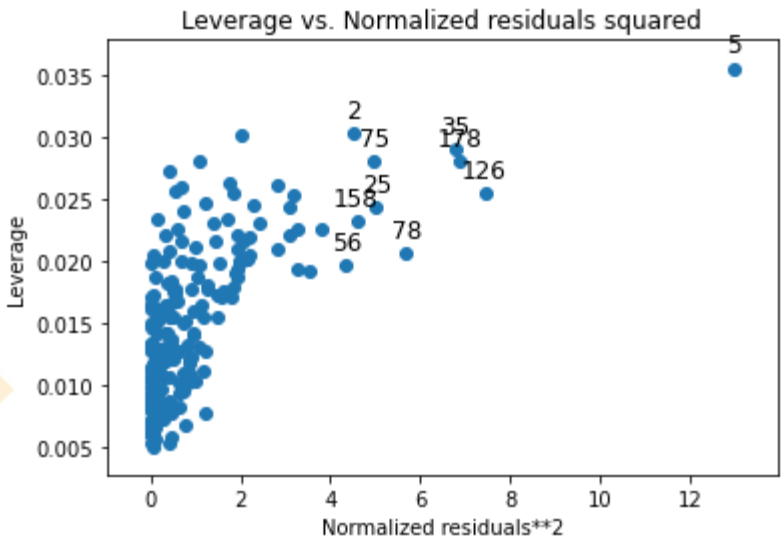


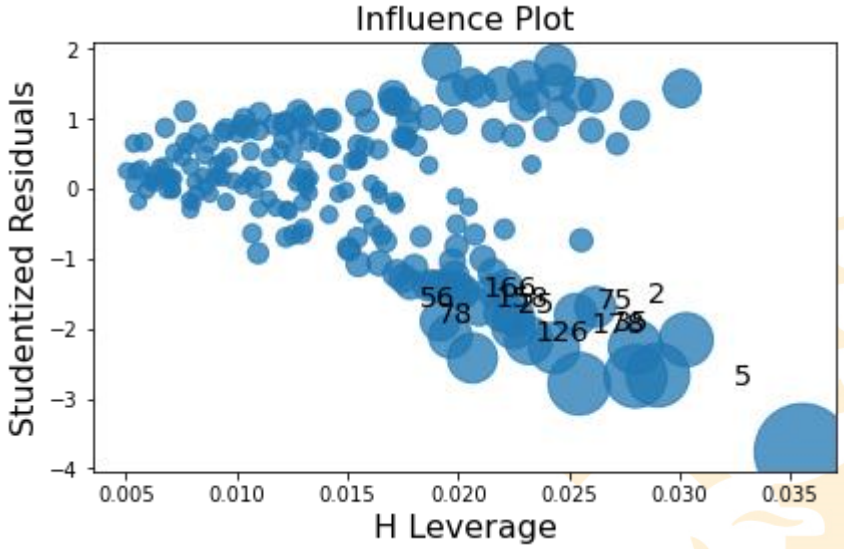
comillas.edu

Multiple Linear Regression. Cases

Simplified model II

- MLR-1: Marketing dataset (Analysis of residuals III)
  - Leverage vs Normalized residuals squared plot. It is better
  - Residual-Leverage plot. Better too.

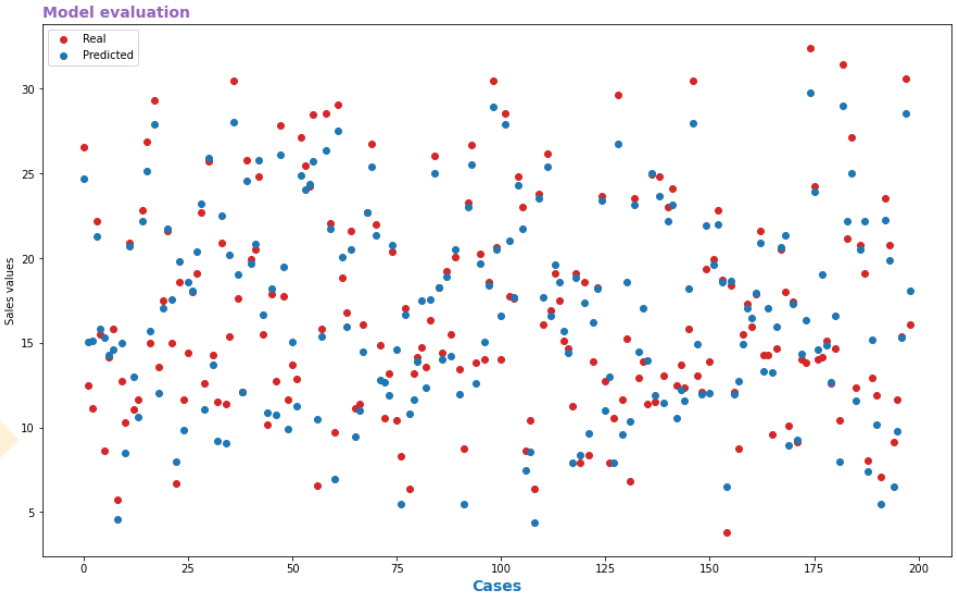


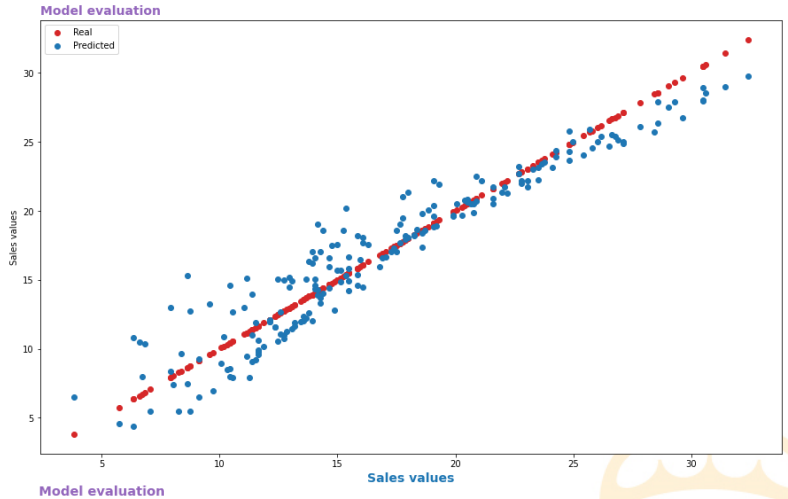


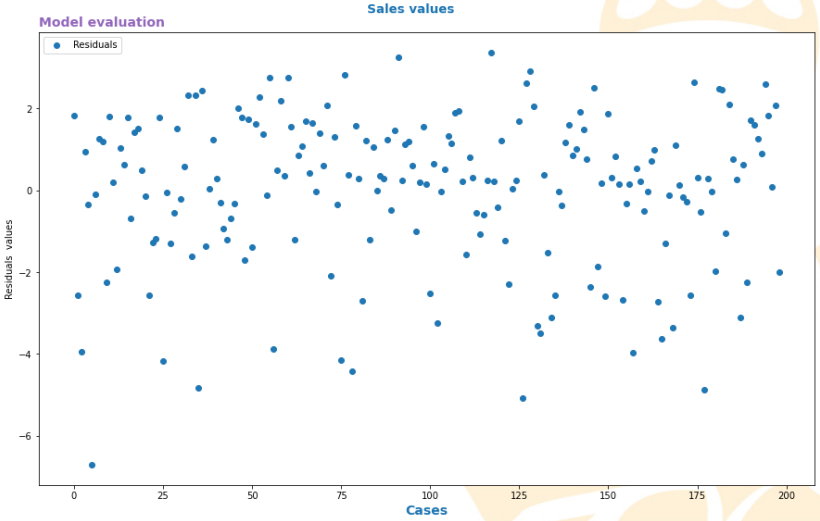
Multiple Linear Regression. Cases

Simplified model II

- **MLR-1: Marketing dataset** (Prediction)
  - Real and predicted values
  - Residuals plot







Multiple Linear Regression Collinearity

- It appears when there is a strong correlation among input variables

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Low determination of the problem causes an inaccurate computation of the inverse

- Consequences
  - Model coefficient estimation can be arbitrary (it is not clear which variable explains the output)
  - It is an unstable model that can behave bad with new data

Multiple Linear Regression
Collinearity

- How is it detected?
  - F test indicates that the model is clearly explanatory (with R^2 reasonable and small RMSE. They are not affected)
  - t individual tests do not show as clearly that, observed by the overall F test. Usually they are skewed to low value.
  - Parameter *variances* and *covariances* estimated are high (They increase with the collinearity degree)

- Other methods
  - *Correlation* among explanatory variables
  - Estimation of ***VIF's*** (**V**ariance **I**nflator **F**actor)

Multiple Linear Regression Collinearity

- Estimation of a Variance Inflator Factor (VIF)

A VIF can be estimated for each explanatory variable X_i

1. It is run an ordinary least square regression that has X_i as a function of all the other explanatory variables

2. The VIF_i is estimated for the coefficient β_i as

$$VIF_i = \frac{1}{1 - R_j^2}$$

Where the R_j^2 is the determination coefficient of the regression equation on step 1

3. A rule of thumb is that if $VIF_i > 10$ high collinearity (around 5 is also used as a symptom of collinearity)

Multiple Linear Regression. Cases

- MLR-2: LAozone

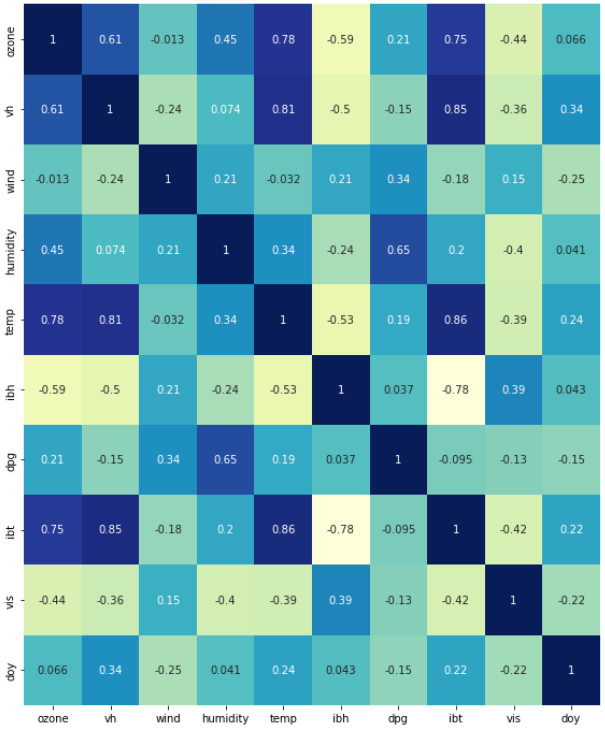
```
#Data Loading

LAozone = pd.read_csv('LAozone.csv')
LAozone.head()
```

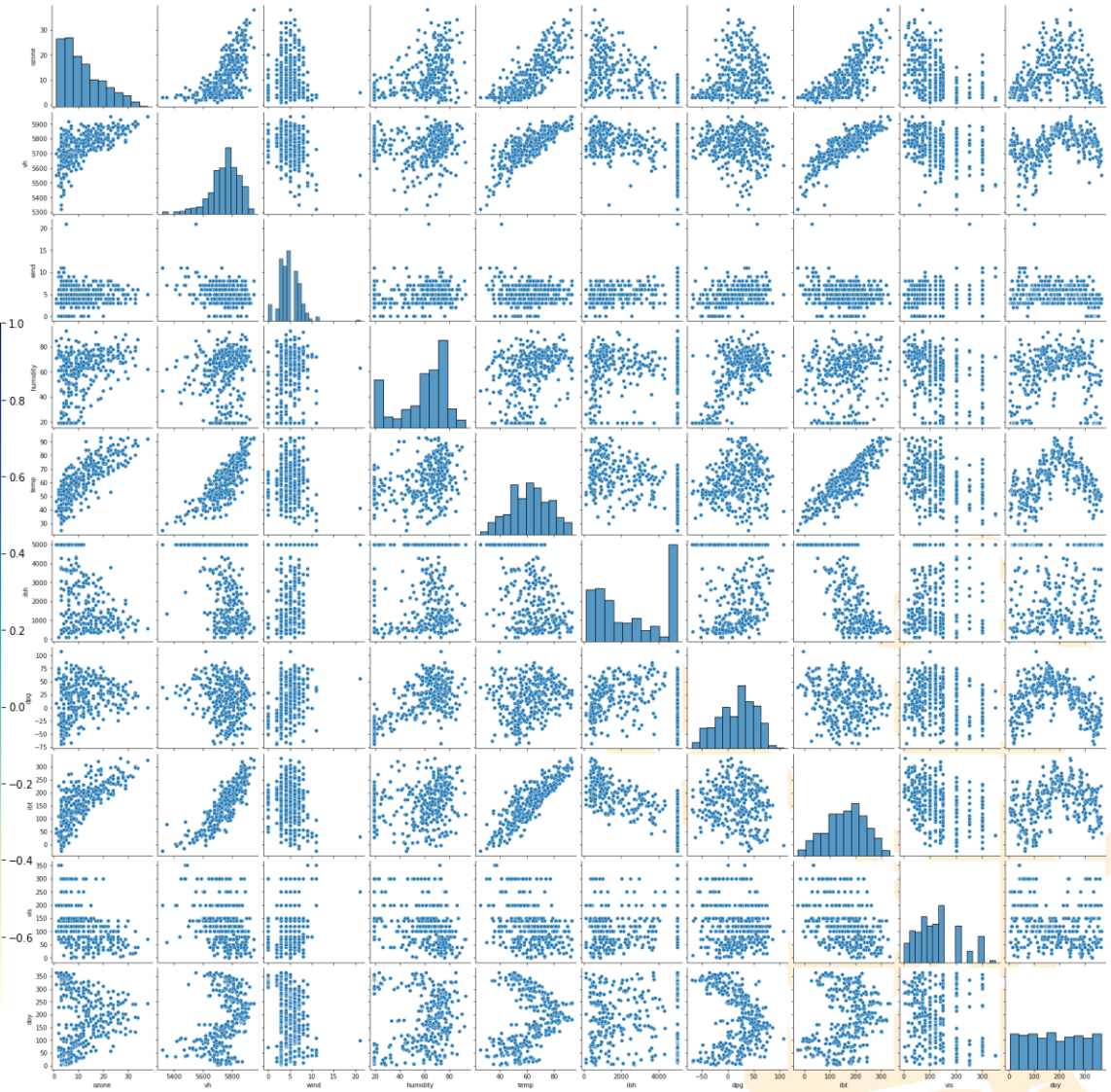
	ozone	vh	wind	humidity	temp	ibh	dpg	ibt	vis	doy
count	330.000000	330.000000	330.000000	330.000000	330.000000	330.000000	330.000000	330.000000	330.000000	330.000000
mean	11.775758	5750.484848	4.890909	58.130303	61.754545	2572.875758	17.369697	161.160606	124.533333	181.727273
std	8.011277	105.708241	2.293159	19.865000	14.458737	1803.885870	35.717181	76.679424	79.362393	106.060593
min	1.000000	5320.000000	0.000000	19.000000	25.000000	111.000000	-69.000000	-25.000000	0.000000	3.000000
25%	5.000000	5690.000000	3.000000	47.000000	51.000000	877.500000	-9.000000	107.000000	70.000000	90.250000
50%	10.000000	5760.000000	5.000000	64.000000	62.000000	2112.500000	24.000000	167.500000	120.000000	177.500000
75%	17.000000	5830.000000	6.000000	73.000000	72.000000	5000.000000	44.750000	214.000000	150.000000	275.750000
max	38.000000	5950.000000	21.000000	93.000000	93.000000	5000.000000	107.000000	332.000000	350.000000	365.000000

Multiple Linear Regression. Cases

- MLR-2: LAozone

```
# plot numerical data as pairs
sns.pairplot(LAozone);
```





Multiple Linear Regression. Cases

- MLR-2: LAozone

```
# Fit the resgression line using 'OLS'
model = sm.OLS(y, X_sm).fit()
```

OLS Regression Results

Dep. Variable:	ozone	R-squared:	0.701
Model:	OLS	Adj. R-squared:	0.693
Method:	Least Squares	F-statistic:	83.40
Date:	Thu, 15 Jun 2023	Prob (F-statistic):	1.65e-78
Time:	14:57:30	Log-Likelihood:	-955.16
No. Observations:	330	AIC:	1930.
Df Residuals:	320	BIC:	1968.
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	18.3793	29.505	0.623	0.534	-39.668	76.427
vh	-0.0051	0.005	-0.952	0.342	-0.016	0.005
wind	-0.0198	0.124	-0.160	0.873	-0.264	0.224
humidity	0.0805	0.019	4.274	0.000	0.043	0.118
temp	0.2743	0.050	5.516	0.000	0.176	0.372
ibh	-0.0002	0.000	-0.846	0.398	-0.001	0.000
dpg	-0.0037	0.011	-0.327	0.744	-0.026	0.019
ibt	0.0293	0.014	2.150	0.032	0.002	0.056
vis	-0.0081	0.004	-2.149	0.032	-0.015	-0.001
doy	-0.0088	0.003	-3.253	0.001	-0.014	-0.003

Omnibus:	1.694	Durbin-Watson:	1.470
Prob(Omnibus):	0.429	Jarque-Bera (JB):	1.649
Skew:	0.173	Prob(JB):	0.438
Kurtosis:	2.972	Cond. No.	7.66e+05

- p-value of ***F test*** very low, this means that at least one explanatory variable is significant to explain ***ozone***
- Adjusted R^2 is relatively high
- High p-value in several coefficients that could be superfluous (with a 95% we should delete)

These coefficients can be 0

If β=0 this variable is not explaining

Multiple Linear Regression. Cases

- MLR-2: LAozone

```
# VIF estimation
from statsmodels.stats.outliers_influence import variance_inflation_factor

# VIF dataframe
vif_data = pd.DataFrame()
vif_data["feature"] = X.columns

# calculating VIF for each feature
vif_data["VIF"] = [variance_inflation_factor(X.values, i)
                          for i in range(len(X.columns))]

print(vif_data)
```

```
Columns for VIF estimation Index(['vh', 'wind', 'humidity',
'temp', 'ibh', 'dpg', 'ibt', 'vis', 'doy'], dtype='object')
    feature         VIF
0        vh   73.698618
1      wind    7.177032
2  humidity   21.834728
3      temp  160.872648
4       ibh   13.251433
5       dpg    3.353920
6       ibt   79.227608
7       vis    5.095506
8       doy    5.421118
```

- Some variables present symptoms of collinearity
- Observe correlations and simplify the model

Multiple Linear Regression. Cases

vh, wind, ibh and dpg are not significant. We can try removing first these variables

- MLR-2: LAozone

```
# Simplified model removing= vh wind ibh dpg
X_simp = LAozone[['humidity', 'temp', 'ibt', 'vis', 'doy']]
y_simp = LAozone['ozone'])
```

Simplified model

OLS Regression Results

Dep. Variable:	ozone	R-squared:	0.699
Model:	OLS	Adj. R-squared:	0.694
Method:	Least Squares	F-statistic:	150.4
Date:	Fri, 16 Jun 2023	Prob (F-statistic):	3.24e-82
Time:	09:00:55	Log-Likelihood:	-956.37
No. Observations:	330	AIC:	1925.
Df Residuals:	324	BIC:	1948.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-10.3189	1.635	-6.309	0.000	-13.536	-7.101
humidity	0.0851	0.014	5.932	0.000	0.057	0.113
temp	0.2327	0.036	6.452	0.000	0.162	0.304
ibt	0.0349	0.007	5.208	0.000	0.022	0.048
vis	-0.0082	0.004	-2.222	0.027	-0.015	-0.001
doy	-0.0101	0.002	-4.185	0.000	-0.015	-0.005

Omnibus:	1.543	Durbin-Watson:	1.442
Prob(Omnibus):	0.462	Jarque-Bera (JB):	1.567
Skew:	0.164	Prob(JB):	0.457
Kurtosis:	2.922	Cond. No.	2.00e+03

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2e+03. This might indicate that there are strong multicollinearity or other numerical problems.

- p-value of ***F test*** very low, this means that at least one explanatory variable is significant to explain ***ozone***
- Adjusted R^2 is similar
- Low p-value in coefficients

Multiple Linear Regression. Cases

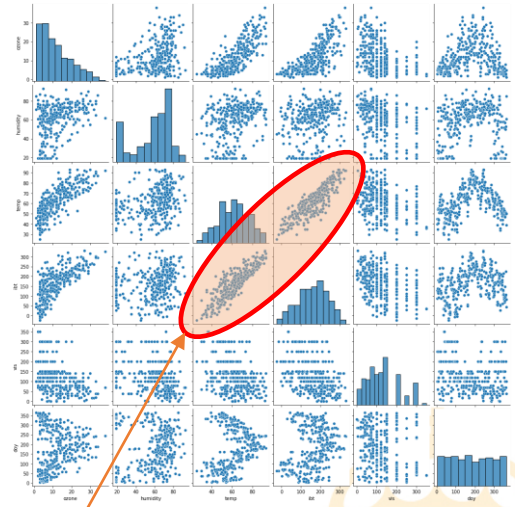
Simplified model

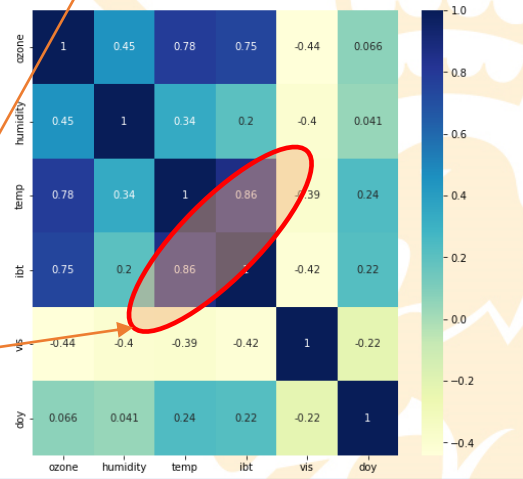
- MLR-2: LAozone

```
# VIF estimation

# VIF dataframe
vif_data = pd.DataFrame()
vif_data["feature"] = X_simp.columns

# calculating VIF for each feature
vif_data["VIF"] = [variance_inflation_factor(X_simp.values, i)
                          for i in range(len(X_simp.columns))]

print(vif_data)
```

	feature	VIF
0	humidity	11.501445
1	temp	59.858288
2	ibt	22.257528
3	vis	3.445520
4	doy	4.106155

- Some variables present some symptoms of collinearity to be investigated





Remove one

The model simplified is better than the original
Possible collinearity has to be studied

Multiple Linear Regression. Cases

- MLR-2: LAozone

```
# new Simplified model removing= ibt
X_simp2 = LAozone[['humidity', 'temp', 'vis', 'doy']]
y_simp2 = LAozone['ozone']
```

```
 OLS Regression Results
==============================================================================
Dep. Variable:                  ozone   R-squared:                       0.674
Model:                            OLS   Adj. R-squared:                  0.670
Method:                 Least Squares   F-statistic:                     167.8
Date:                Fri, 16 Jun 2023   Prob (F-statistic):           1.01e-77
Time:                        09:15:03   Log-Likelihood:                -969.63
No. Observations:                 330   AIC:                             1949.
Df Residuals:                     325   BIC:                             1968.
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        -12.6613      1.634     -7.747      0.000     -15.877      -9.446
humidity       0.0650      0.014      4.524      0.000       0.037       0.093
temp           0.3920      0.020     19.725      0.000       0.353       0.431
vis           -0.0133      0.004     -3.601      0.000      -0.021      -0.006
doy           -0.0104      0.002     -4.151      0.000      -0.015      -0.005
==============================================================================
Omnibus:                        5.296   Durbin-Watson:                   1.427
Prob(Omnibus):                  0.071   Jarque-Bera (JB):                5.235
Skew:                           0.308   Prob(JB):                       0.0730
Kurtosis:                       3.016   Cond. No.                     1.62e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 1.62e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Simplified model 2
Removing ibt

- p-value of ***F test*** very low, this means that at least one explanatory variable is significant to explain ***ozone***
- Adjusted R^2 is similar
- Low p-value in coefficients

	feature	VIF
0	humidity	9.625785
1	temp	14.223059
2	vis	2.496666
3	doy	4.073537

Multiple Linear Regression. Cases

- MLR-2: LAozone

```
# new Simplified model removing= ibt and humidity
X_simp3 = LAozone[['temp', 'vis', 'doy']]
y_simp3 = LAozone['ozone']
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  ozone   R-squared:                       0.653
Model:                            OLS   Adj. R-squared:                  0.650
Method:                 Least Squares   F-statistic:                     204.6
Date:                Fri, 16 Jun 2023   Prob (F-statistic):           1.28e-74
Time:                        09:22:33   Log-Likelihood:                -979.71
No. Observations:                 330   AIC:                             1967.
Df Residuals:                     326   BIC:                             1983.
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -9.3142      1.500     -6.209      0.000     -12.266      -6.363
temp           0.4128      0.020     20.743      0.000       0.374       0.452
vis           -0.0187      0.004     -5.176      0.000      -0.026      -0.012
doy           -0.0114      0.003     -4.460      0.000      -0.016      -0.006
==============================================================================
Omnibus:                        5.505   Durbin-Watson:                   1.382
Prob(Omnibus):                  0.064   Jarque-Bera (JB):                5.639
Skew:                           0.310   Prob(JB):                       0.0596
Kurtosis:                       2.843   Cond. No.                     1.41e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 1.41e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Simplified model 3
Removing ibt, humidity

- p-value of ***F test*** very low, this means that at least one explanatory variable is significant to explain ***ozone***
- Adjusted R^2 is lower, but close
- Low p-value in coefficients

- No symptoms of collinearity t

	feature	VIF
0	temp	5.597619
1	vis	2.495784
2	doy	4.072839

Multiple Linear Regression
Interactions between inputs

- The standard linear regression model provides interpretable results and works quite well on many real-world problems
- However, it makes several highly restrictive assumptions that are often violated in practice
  - One of the most important assumption states that the relationship between the predictors and response is additive

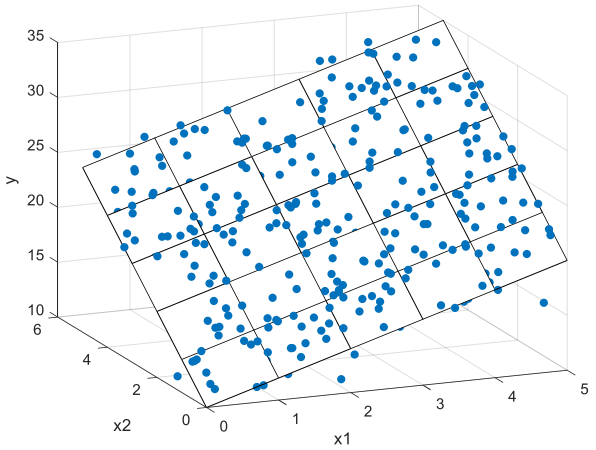
$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

The **additive assumption** means that the **effect of changes in an input X_i on the response Y** is **independent of the values of the other predictors**

Other considerations in regression
Interactions between inputs

- The interaction between inputs can have a synergy effect in the output
  - The combined effect or impact of inputs is greater than the sum of their individual effects on the output (e.g. a multiplicative effect)

WITHOUT Interaction

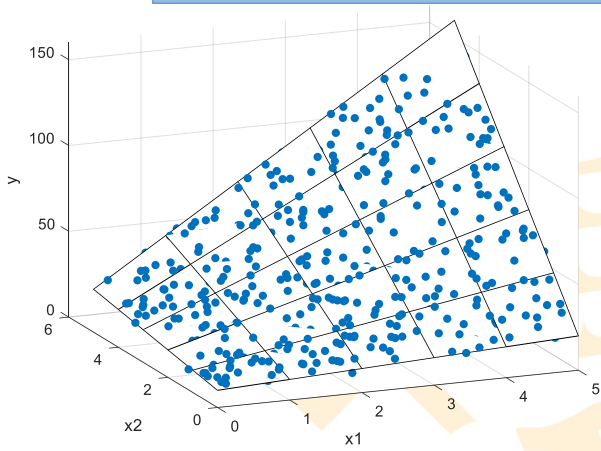


$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

Regardless of the value of X_2, a one-unit increase in X_1 will lead to a β_1-unit increase in Y

Note that the **interaction** effect has **nothing to do with the correlation between inputs**

WITH Interaction



$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \boxed{\beta_3 X_1 X_2} + \epsilon$$

The whole is greater than the sum of the parts

Interaction term

Other considerations in regression
Interactions between inputs

- The standard way to allow for interaction effects is to include a third input variable, called an interaction term, which is constructed by computing the product of inputs

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

- Note that this model is equivalent to

$$\begin{aligned} Y &= \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon \\ &= \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \epsilon \end{aligned}$$

where

$$\tilde{\beta}_1 = \beta_1 + \beta_3 X_2$$

The effect of X_1 on Y is no longer constant: varying X_2 will change the impact of X_1 on Y

Other considerations in regression Interactions between inputs

- The hierarchical principle states that:

> if we include an interaction in a model, we should also include the main effects, even if the p-values associated with their coefficients are not significant

- In other words, if the interaction between X_1 and X_2 seems important, then we should include both X_1 and X_2 in the model even if their coefficient estimates have large p-values.
- The rationale for this principle is that if $X_1 \times X_2$ is related to the response, then whether or not the coefficients of X_1 or X_2 are exactly zero is of little interest.
- Also $X_1 \times X_2$ is typically correlated with X_1 and X_2, and so leaving them out tends to alter the meaning of the interaction.

Multiple Linear Regression. Cases

- MLR-2: LAozone

```
model_interac = smf.ols(formula = 'ozone ~ humidity+temp+vis+doy+
temp*humidity', data = LAozone)
```

Simplified model with interaction

```
 OLS Regression Results
==============================================================================
Dep. Variable:                  ozone   R-squared:                       0.704
Model:                            OLS   Adj. R-squared:                  0.699
Method:                 Least Squares   F-statistic:                     153.8
Date:                Fri, 16 Jun 2023   Prob (F-statistic):           2.58e-83
Time:                        10:43:00   Log-Likelihood:                -953.78
No. Observations:                 330   AIC:                             1920.
Df Residuals:                     324   BIC:                             1942.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      8.6376      4.039      2.138      0.033       0.691      16.584
humidity      -0.2899      0.064     -4.560      0.000      -0.415      -0.165
temp           0.0140      0.069      0.203      0.839      -0.121       0.149
vis           -0.0117      0.004     -3.304      0.001      -0.019      -0.005
doy           -0.0114      0.002     -4.780      0.000      -0.016      -0.007
temp:humidity  0.0061      0.001      5.716      0.000       0.004       0.008
==============================================================================
Omnibus:                        6.509   Durbin-Watson:                   1.483
Prob(Omnibus):                  0.039   Jarque-Bera (JB):                6.458
Skew:                           0.281   Prob(JB):                       0.0396
Kurtosis:                       3.393   Cond. No.                     6.79e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 6.79e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

- p-value of ***F test*** very low, this means that the model is OK
- ***Significant*** coefficient of the interaction term
- Adjusted R^2 is better
- Low p-value in coefficients (except temp but we keep it)

As expected, there are symptoms of collinearity. They do not have to take into account

	feature	VIF
0	humidity	31.095243
1	temp	27.957021
2	vis	4.504093
3	doy	4.216351
4	temp_hum	48.560084

The simplified model with interaction term is better

Ozone=8.637-0,29*humidity-0.0117*vis-0.0114*doy+0.014*temp+0.006*humidity*temp

Multiple Linear Regression. Cases

- MLR-2: LAozone

Creation of two data sets: TRAINING and TEST

```
from sklearn.model_selection import train_test_split

# Dividing the dataset in training and test
# ============================================================================
X = LAozone[['humidity', 'temp', 'vis', 'doy']]
y = LAozone['ozone']


X_train, X_test, y_train, y_test = train_test_split(
                                        X,
                                        #y.values.reshape(-1,1),
                                        y,
                                        train_size   = 0.7,  # 70% for training
                                        random_state = 1234,
                                        shuffle      = True
                                    )
```

Final model Using training data

```
# Interaction of variables

# Model creation with interactions
# ================================================
# Adding new column including the interaction
X_train['hum_temp'] = X_train['humidity'] * X_train['temp']
X_test['hum_temp']  = X_test['humidity'] * X_test['temp']

# A la matriz de predictores se le tiene que añadir una columna de 1s
# para el intercept del modelo

# Add a constant to get an intercept
X_train_c = sm.add_constant(X_train, prepend=True)
model_train  = sm.OLS(endog=y_train, exog=X_train_c,)

# Fit the resgression line using 'OLS'
model_train  = model_train.fit()
print(model_train.summary())
```

Multiple Linear Regression. Cases

- MLR-2: LAozone

Results from the training dataset

OLS Regression Results

Dep. Variable:	ozone	R-squared:	0.714
Model:	OLS	Adj. R-squared:	0.707
Method:	Least Squares	F-statistic:	111.6
Date:	Fri, 16 Jun 2023	Prob (F-statistic):	8.33e-59
Time:	11:26:03	Log-Likelihood:	-667.20
No. Observations:	230	AIC:	1346.
Df Residuals:	224	BIC:	1367.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	9.0452	5.089	1.777	0.077	-0.983	19.073
humidity	-0.3301	0.081	-4.073	0.000	-0.490	-0.170
temp	-0.0073	0.087	-0.083	0.934	-0.179	0.164
vis	-0.0087	0.004	-2.040	0.043	-0.017	-0.000
doy	-0.0110	0.003	-3.733	0.000	-0.017	-0.005
hum_temp	0.0068	0.001	5.009	0.000	0.004	0.010

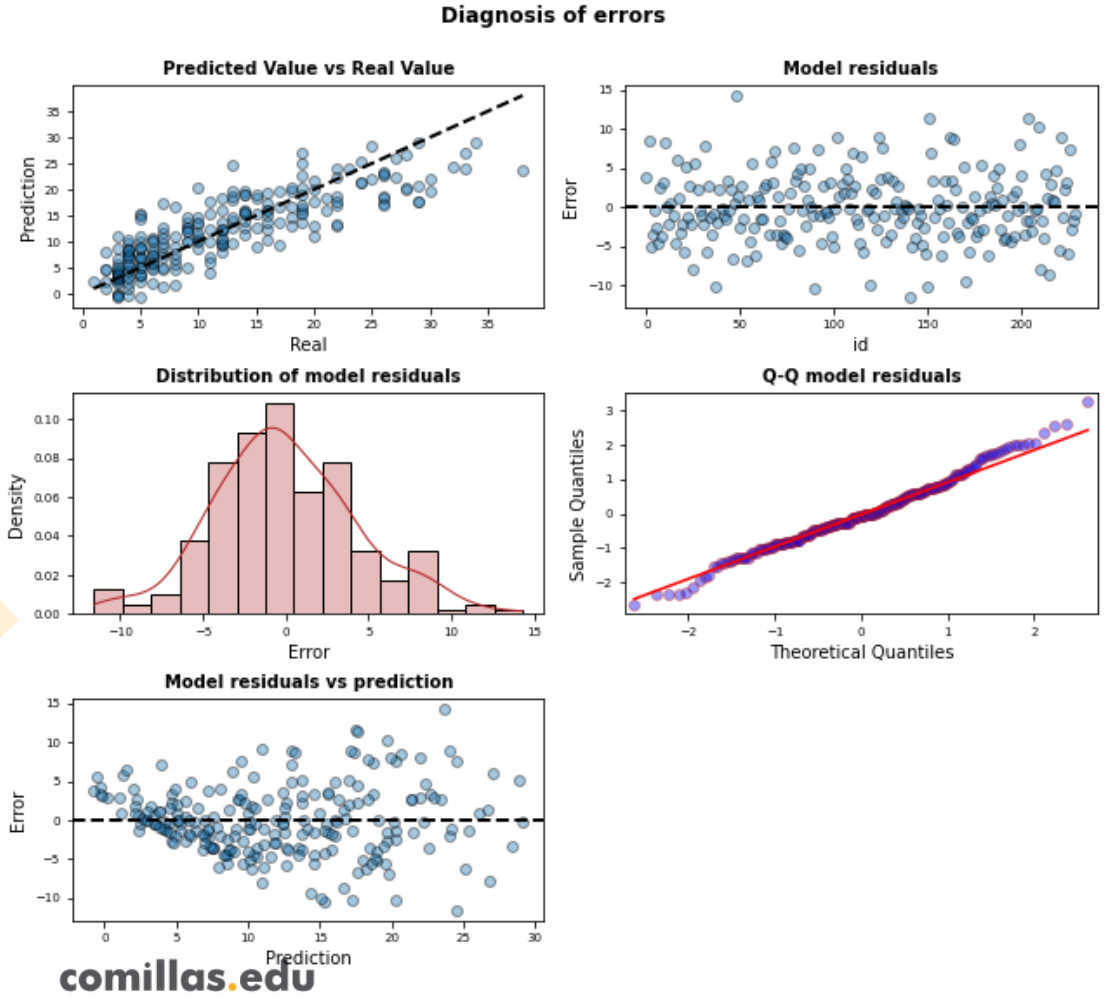
Omnibus:	3.979	Durbin-Watson:	2.111
Prob(Omnibus):	0.137	Jarque-Bera (JB):	3.605
Skew:	0.265	Prob(JB):	0.165
Kurtosis:	3.309	Cond. No.	7.23e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.23e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

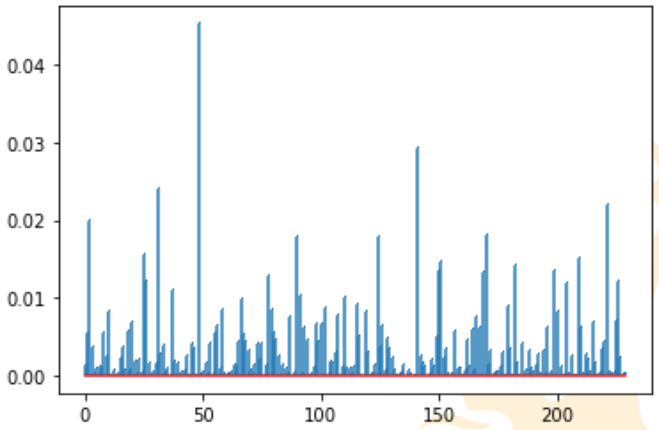
Multiple Linear Regression. Cases

- MLR-2: LAozone

Results from the training dataset



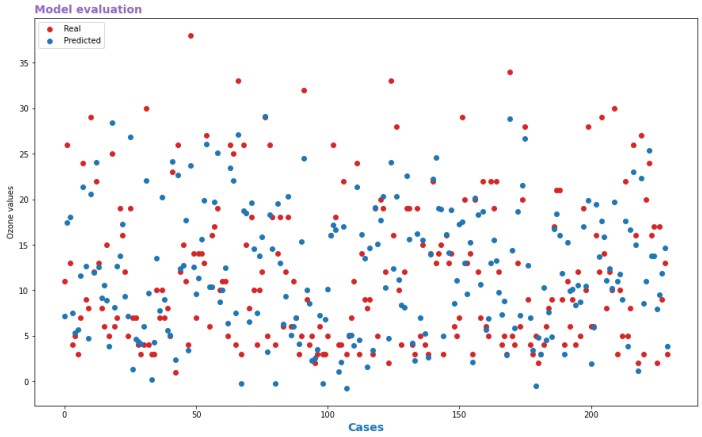

```
 Examples with higher Cook distance
[ 48 141   31 221    2 170   90 124   25 209]
```

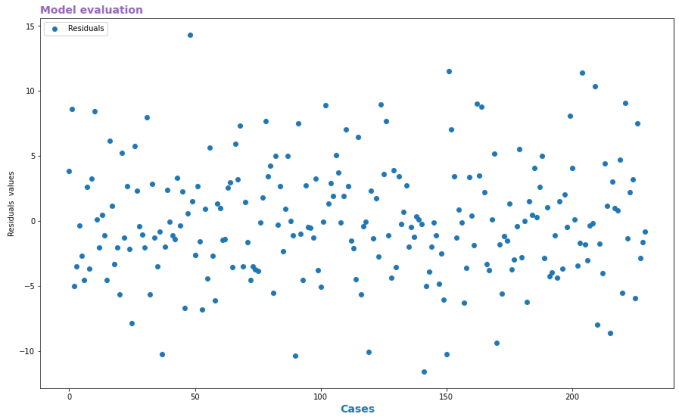


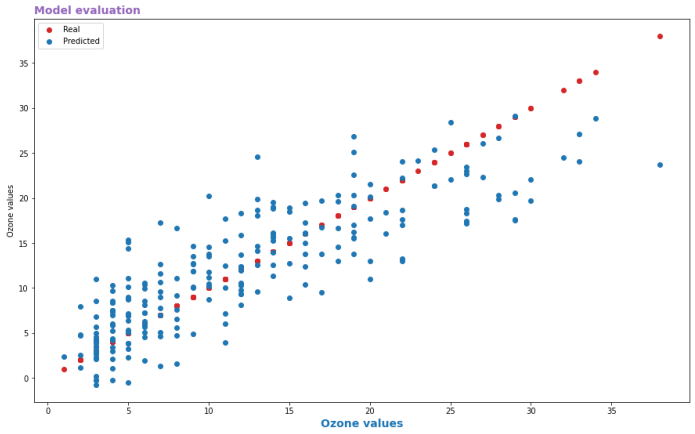
Multiple Linear Regression. Cases

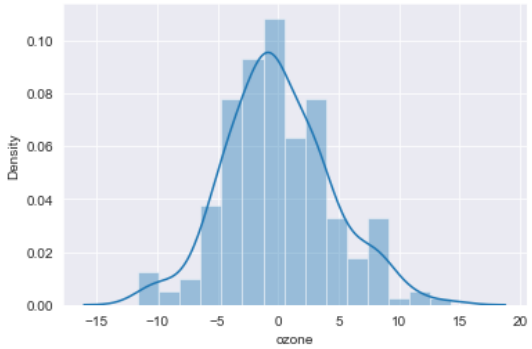
Results from the training dataset

- MLR-2: LAozone (TRAINING – Real vs Predicted)









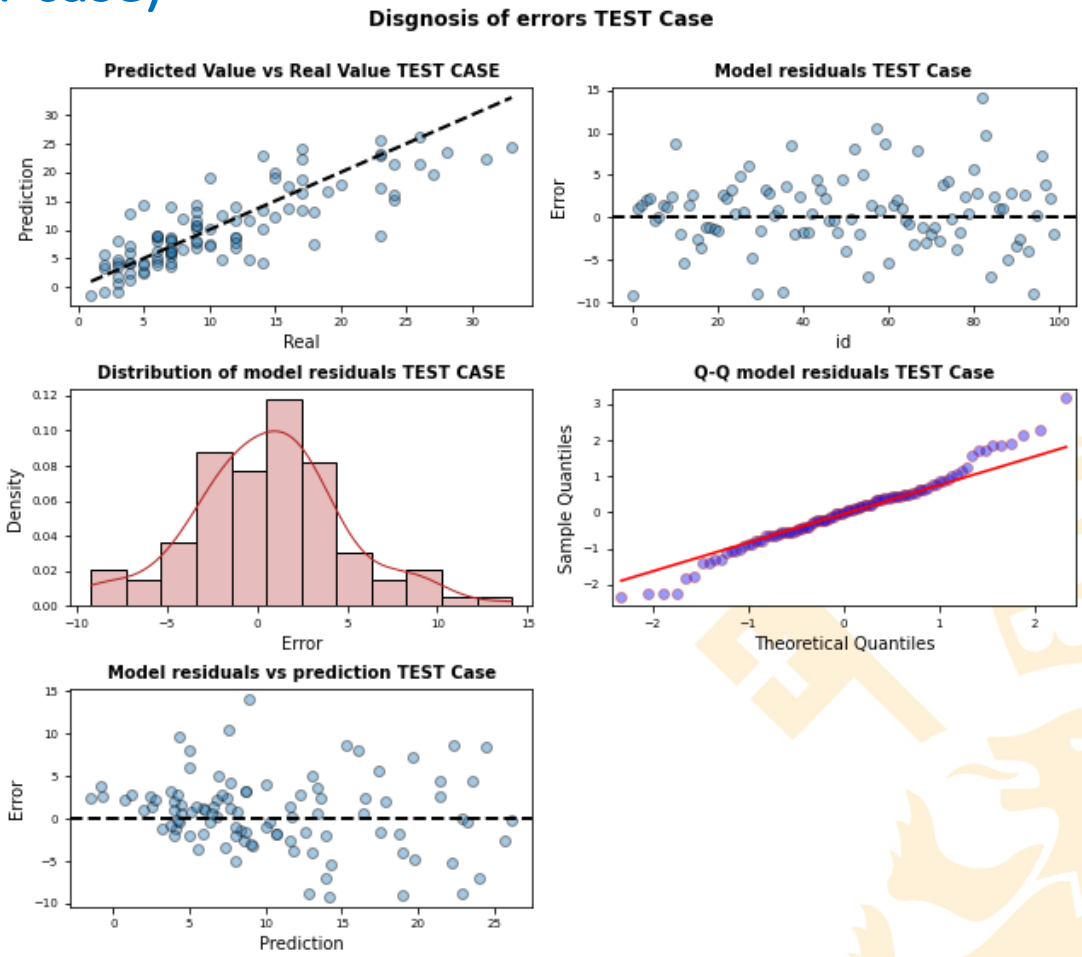

```
Mean Absolute Error - training: 3.415019480242933
Mean Squared Error - training: 19.372471143703752
Root Mean Squared Error - training: 4.4014169472686575
```

Multiple Linear Regression. Cases

Testing the dataset

- MLR-2: LAozone (TEST case)

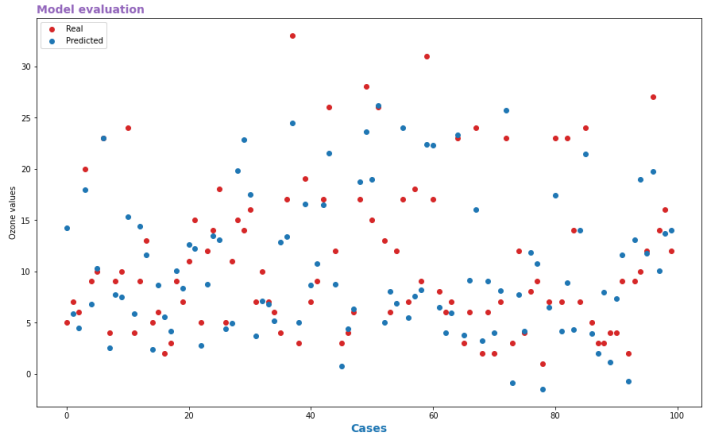
- Errors in test seem close to normal,
- Prediction follows Real values

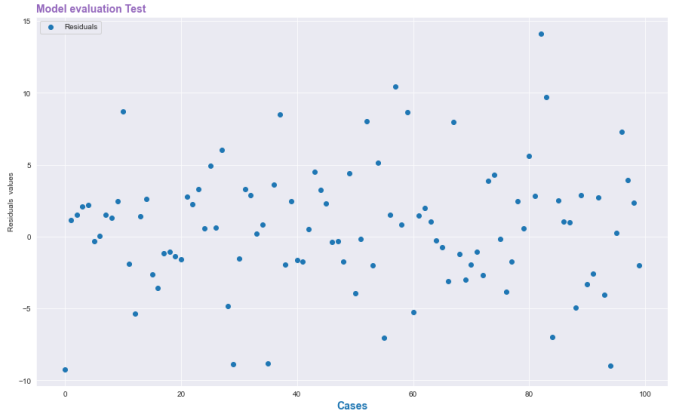


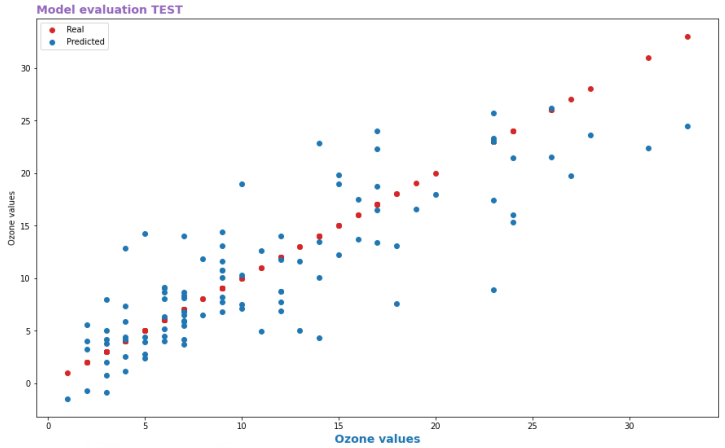
Multiple Linear Regression. Cases

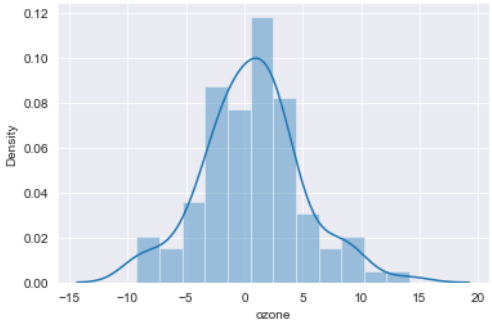
Testing the dataset

- MLR-2: LAozone (TEST – Real vs Predicted









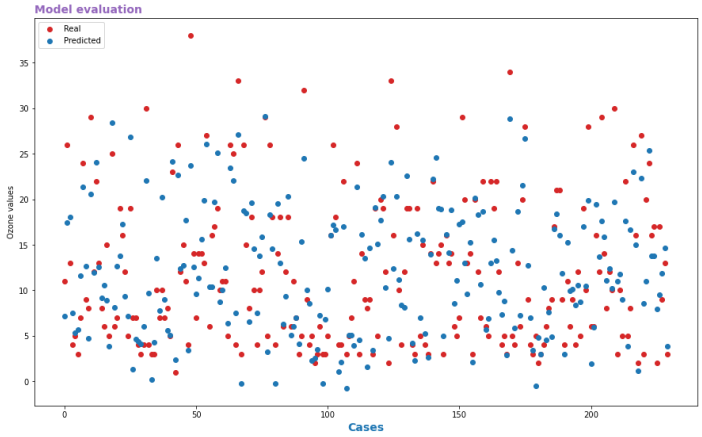
Similar values to training

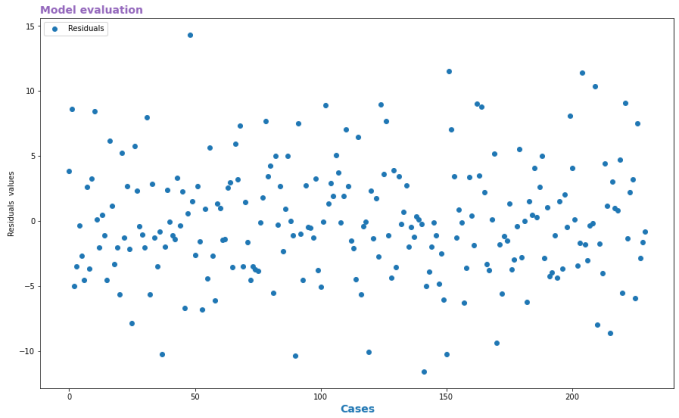
```
Mean Absolute Error - test: 3.2781792626338673
Mean Squared Error - test: 18.579964112878738
Root Mean Squared Error - test: 4.310448249646287
```

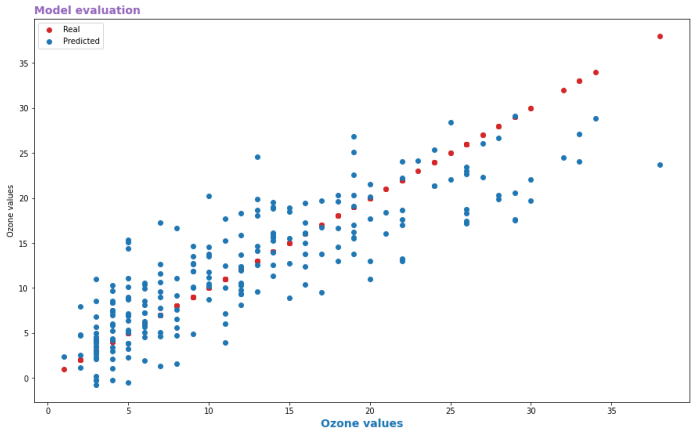
Multiple Linear Regression. Cases

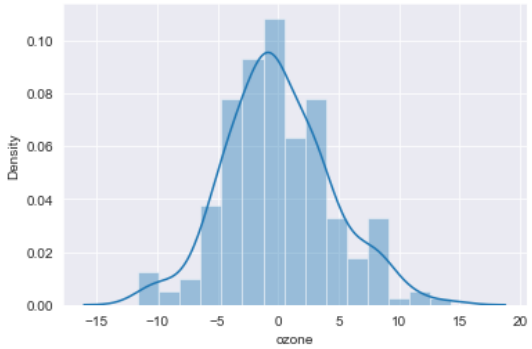
Results from the training dataset

- MLR-2: LAozone (TEST – Real vs Predicted)










```
Mean Absolute Error - training: 3.415019480242933
Mean Squared Error - training: 19.372471143703752
Root Mean Squared Error - training: 4.4014169472686575
```

Multiple Linear Regression
Qualitative input variables

- In the regression model we have assumed that inputs are quantitative
- In practice, often some inputs are qualitative
- Approach
  - Create a set of new dummy variables that represent the same information of the qualitative variable
  - Fit the regression model using the dummy variables instead of the qualitative one
- A **dummy** variable is an indicator variable that has two possible numerical values (typically 0 and 1)
- Two different cases
  - Input variables with two possible values
  - Input variables with more than two possible values

Other considerations in regression
Qualitative input variable with 2 levels

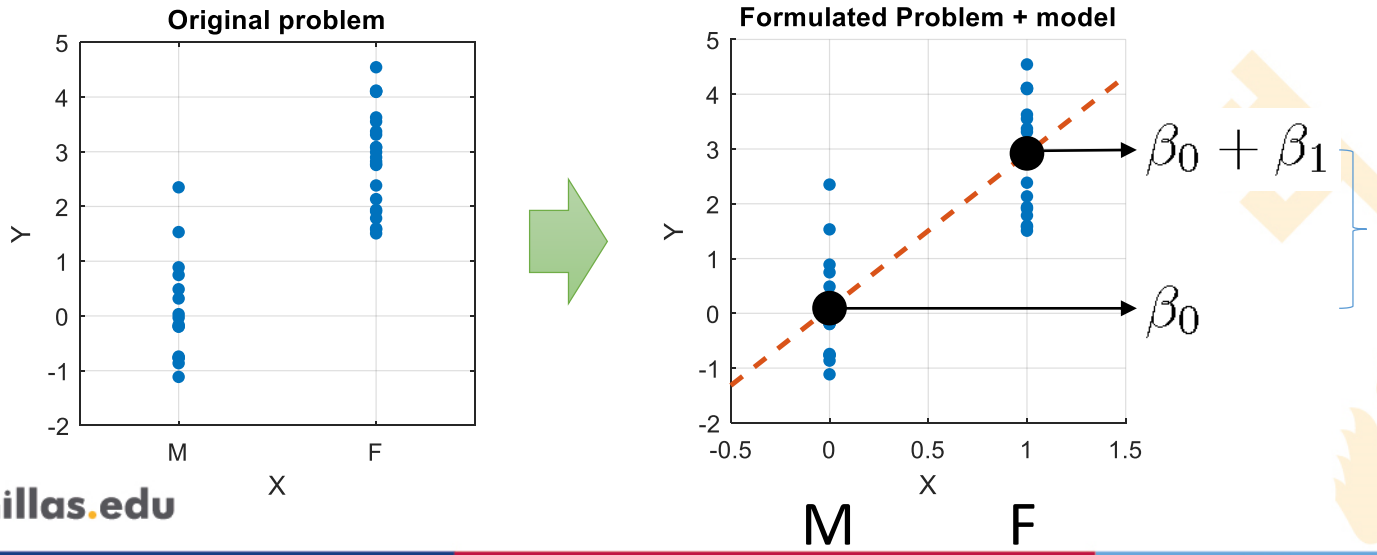
- If a qualitative variable has only two levels (or possible values), then incorporating it into a regression model is very simple
- Example

Dummy variable

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male,} \end{cases}$$

Regression model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male.} \end{cases}$$



Other considerations in regression
Qualitative input variable with 2 levels

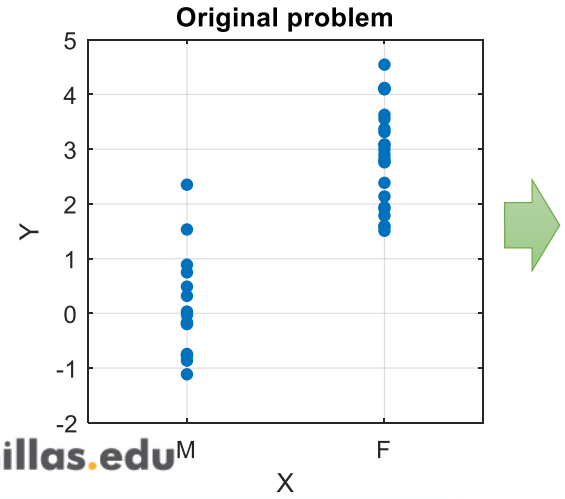
- Coding females as 1 and males as 0 is arbitrary
  - Although it has no effect on the regression fit
  - It alters the interpretation of the coefficients
- A different coding (values -1 and 1)

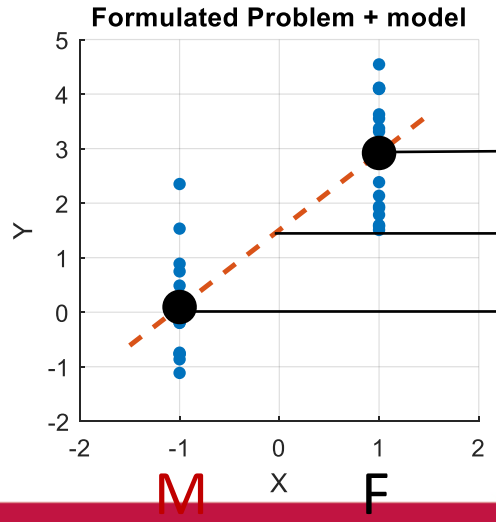
Dummy variable

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ -1 & \text{if } i\text{th person is male} \end{cases}$$

Regression model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 - \beta_1 + \epsilon_i & \text{if } i\text{th person is male.} \end{cases}$$



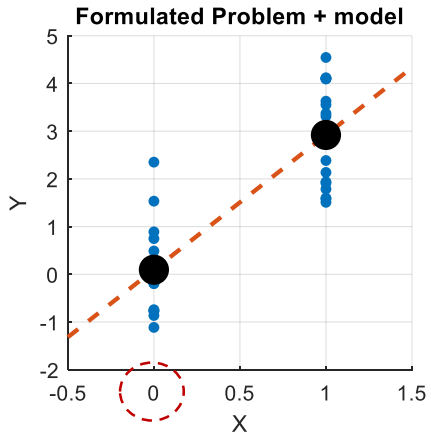


Average output ignoring the gender effect

Now the slope can be interpreted as the amount that females/males are above/below the average

Other considerations in regression
Qualitative input variable with 2 levels

- Compare both models (the estimated coeffs. are different)

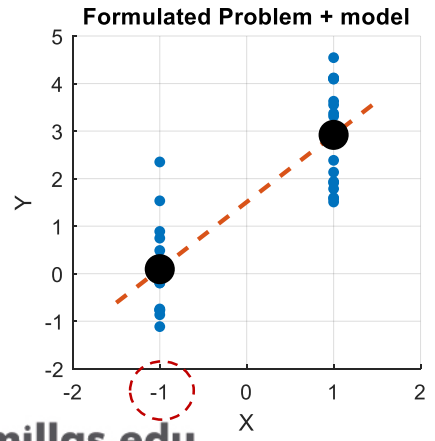


x is a **dummy variable with values {0, 1}**

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	0.095147	0.23149	0.41102	0.68336
x	2.8247	0.29885	9.452	1.5965e-11

Number of observations: 40, Error degrees of freedom: 38
Root Mean Squared Error: 0.926
R-squared: 0.702, Adjusted R-Squared 0.694
F-statistic vs. constant model: 89.3, p-value = 1.6e-11



x is a **dummy variable with values {-1, 1}**

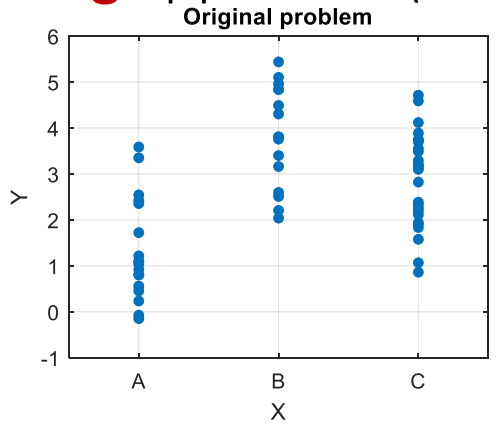
Estimated Coefficients:

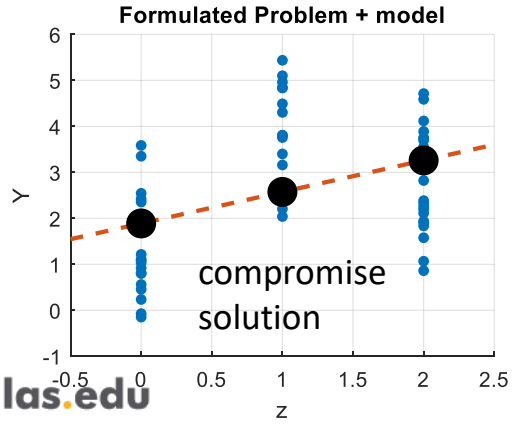
	Estimate	SE	tStat	pValue
(Intercept)	1.5075	0.14943	10.089	2.6692e-12
x	1.4124	0.14943	9.452	1.5965e-11

Number of observations: 40, Error degrees of freedom: 38
Root Mean Squared Error: 0.926
R-squared: 0.702, Adjusted R-Squared 0.694
F-statistic vs. constant model: 89.3, p-value = 1.6e-11

Other considerations in regression
Qualitative input variable with >2 levels

- When a qualitative input has more than two levels, a single dummy variable cannot represent all possible values
- **Wrong** approach (example with 3 levels)





- Use only one quantitative variable with numeric values representing the different possible values of the original qualitative variable

Typical mistake when >2 levels

This model is quite bad because of the wrong modelling approach

```
z is a variable with values {0, 1, 2}

Estimated Coefficients:
                   Estimate       SE       tStat       pValue
                   ________    ________    ______    __________

    (Intercept)     1.8889      0.28259    6.6842    9.9212e-09
    z               0.68543     0.19982    3.4302     0.0011173

Number of observations: 60, Error degrees of freedom: 58
Root Mean Squared Error: 1.31
R-squared: 0.169,  Adjusted R-Squared 0.154
F-statistic vs. constant model: 11.8, p-value = 0.00112
```

Other considerations in regression
Qualitative input variable with >2 levels

- The correct approach consists in creating several dummy variables (for M levels $M-1$ dummies are required)
- Example with 3 levels

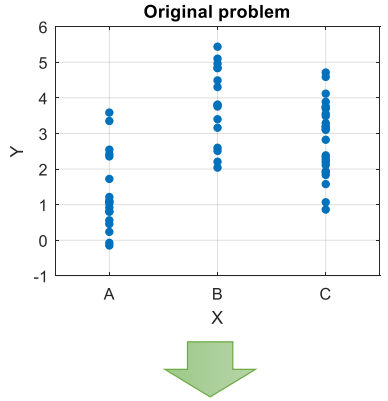
x	y
C	4.1174
C	1.9109
A	1.0326
C	3.5525
B	5.1006
A	2.5442

A	B	
x1	x2	y
0	0	4.1174
0	0	1.9109
1	0	1.0326
0	0	3.5525
0	1	5.1006
1	0	2.5442

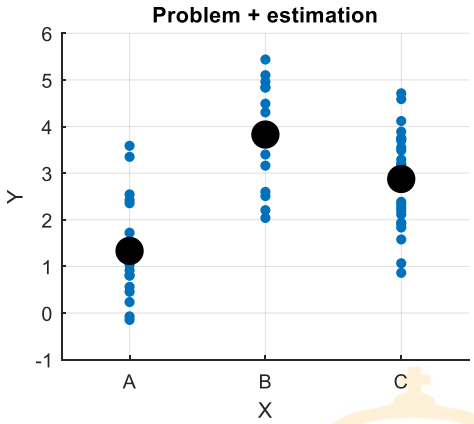
$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th observation is A} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th observation is B} \\ \beta_0 + \epsilon_i & \text{if } i\text{th observation is C} \end{cases}$$

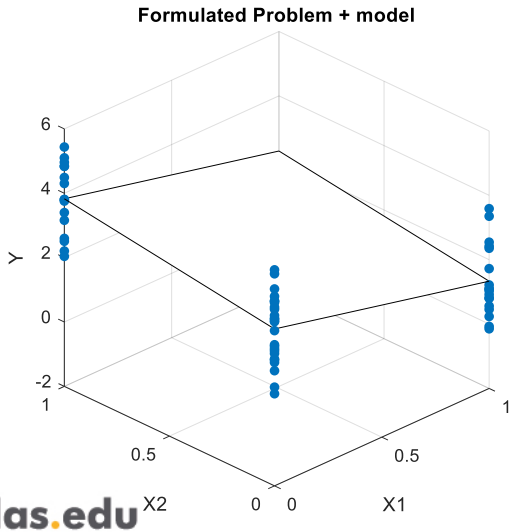
Other considerations in regression
Qualitative input variable with >2 levels

- The correct approach consists in fitting the regression model using the dummy variables instead of the qualitative one



The original problem has one input variable, but a multiple linear regression model is fitted to get the correct solution





x1 and x2 are dummy variables with values {0,1}

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	2.8741	0.20916	13.741	9.7126e-20
x1	-1.5423	0.32702	-4.7162	1.5957e-05
x2	0.95377	0.33888	2.8145	0.0066955

Number of observations: 60, Error degrees of freedom: 57
Root Mean Squared Error: 1.07
R-squared: 0.458, Adjusted R-Squared 0.439
F-statistic vs. constant model: 24.1, p-value = 2.64e-08

Multiple Linear Regression. Cases

- MLR-3: Salaries

```
# Data Loading

salaries = pd.read_csv('salaries.csv')
salaries.head()
```

There are 3 categorical variables

	rank	discipline	yrs.since.phd	yrs.service	sex	salary
0	Prof	B	19	18	Male	139750
1	Prof	B	20	16	Male	173200
2	AsstProf	B	4	3	Male	79750
3	Prof	B	45	39	Male	115000
4	Prof	B	40	41	Male	141500

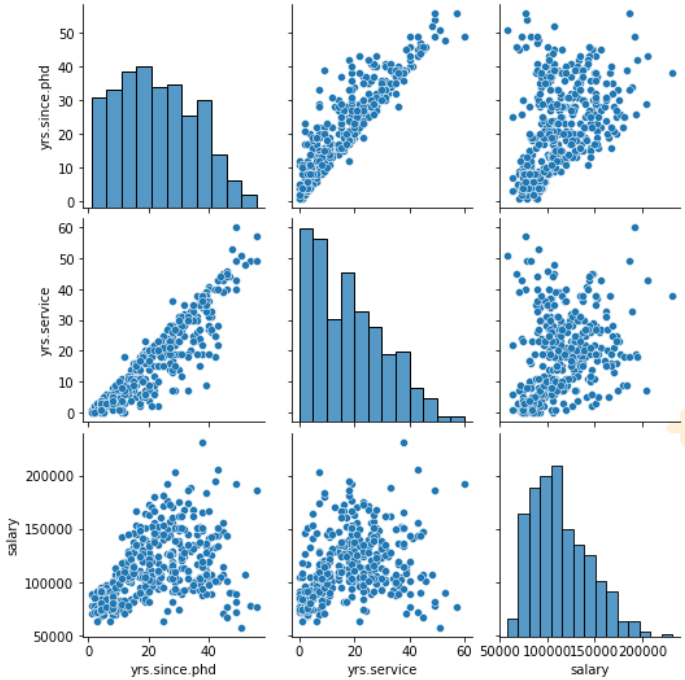
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 397 entries, 0 to 396
Data columns (total 6 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   rank            397 non-null    object
 1   discipline      397 non-null    object
 2   yrs.since.phd   397 non-null    int64
 3   yrs.service     397 non-null    int64
 4   sex             397 non-null    object
 5   salary          397 non-null    int64
dtypes: int64(3), object(3)
memory usage: 18.7+ KB
```

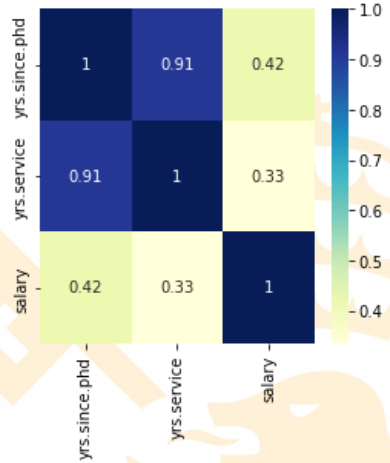
Multiple Linear Regression. Cases

- MLR-3: Salaries

Only numerical variables should be used. How can we use categorical variables?

	yrs.since.phd	yrs.service	salary
count	397.000000	397.000000	397.000000
mean	22.314861	17.614610	113706.458438
std	12.887003	13.006024	30289.038695
min	1.000000	0.000000	57800.000000
25%	12.000000	7.000000	91000.000000
50%	21.000000	16.000000	107300.000000
75%	32.000000	27.000000	134185.000000
max	56.000000	60.000000	231545.000000





Multiple Linear Regression. Cases

- Suppose that, we want to investigate differences in salaries between males and females.
- Based on the gender variable, we can create a new dummy variable that takes the value:

  1 if a person is male; 0 if a person is female

We will use this variable as a predictor in the regression equation, leading to the following model:

b0 + b1 if person is male & b0 if person is female

- The coefficients can be interpreted as follow:

  - b0 is the average salary among females

  - b0 + b1 is the average salary among males,

  - b1 is the average difference in salary between males and females

Multiple Linear Regression. Cases

- MLR-3: Salaries

```
# Creation of dummies variables for sex

# First we observe the levels of the categoric variables

# count of each category value
print(salaries["sex"].value_counts())


# Alternative:   Creation of dummy binary variables
#sexMale = pd.get_dummies(salaries['sex'], drop_first = True) # drop_first creats k-1 levels
#sexMale.head()



s = pd.Series(list(salaries['sex'])) # extraction of variable sex

sexMale=pd.get_dummies(s, drop_first=True)  # drop_first removes 1 values leaving k-1 levels

print("\n \n sexMale", sexMale[1:20])
```

Multiple Linear Regression. Cases

- MLR-3: Salaries

	rank	discipline	yrs.since.phd	yrs.service	sex	salary
0	Prof	B	19	18	Male	139750
1	Prof	B	20	16	Male	173200
2	AsstProf	B	4	3	Male	79750
3	Prof	B	45	39	Male	115000
4	Prof	B	40	41	Male	141500
5	AssocProf	B	6	6	Male	97000
6	Prof	B	30	23	Male	175000
7	Prof	B	45	45	Male	147765
8	Prof	B	21	20	Male	119250
9	Prof	B	18	18	Female	129000
10	AssocProf	B	12	8	Male	119800
11	AsstProf	B	7	2	Male	79800
12	AsstProf	B	1	1	Male	77700
13	AsstProf	B	2	0	Male	78000
14	Prof	B	20	18	Male	104800
15	Prof	B	12	3	Male	117150
16	Prof	B	19	20	Male	101000
17	Prof	A	38	34	Male	103450
18	Prof	A	37	23	Male	124750
19	Prof	A	39	36	Female	137000

```
Male      358
Female     39
Name: sex, dtype: int64


 sexMale     Male
1       1
2       1
3       1
4       1
5       1
6       1
7       1
8       1
9       0
10      1
11      1
12      1
13      1
14      1
15      1
16      1
17      1
18      1
19      0
```

Multiple Linear Regression. Cases

How affects the salary depending on Male or Female?

- MLR-3: Salaries

```
# Variables fot the Linear model
import statsmodels.api as sm
X=sexMale
y = salaries['salary']

# Add a constant to get an intercept
X_sm = sm.add_constant(X)
```

```
const     101002.410256
sexMale    14088.008738
dtype: float64
```

```
 OLS Regression Results
==============================================================================
Dep. Variable:                 salary   R-squared:                       0.019
Model:                            OLS   Adj. R-squared:                  0.017
Method:                 Least Squares   F-statistic:                     7.738
Date:                Fri, 16 Jun 2023   Prob (F-statistic):            0.00567
Time:                        14:25:15   Log-Likelihood:                -4655.4
No. Observations:                 397   AIC:                             9315.
Df Residuals:                     395   BIC:                             9323.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        1.01e+05   4809.386     21.001      0.000    9.15e+04     1.1e+05
sexMale     1.409e+04   5064.579      2.782      0.006    4131.107     2.4e+04
==============================================================================
Omnibus:                       28.630   Durbin-Watson:                   1.971
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               33.087
Skew:                           0.703   Prob(JB):                     6.53e-08
Kurtosis:                       3.152   Cond. No.                         6.23
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is
correctly specified.
```

- In this case ***b0*** is the average salary for female is estimated to be 101002
- For male is estimated as a total of 101002 (***b0***)+ 14088 (***b1***)= 115090
- p-value for ***Male*** is significant suggesting that there is a statistical evidence of a difference in average salary between the genders

Multiple Linear Regression. Cases

How affects the salary depending on Male or Female?

- MLR-3: Salaries

```
model_m <- lm(salary ~ sex, data=salaries)
summary(model_m)

The reference dummy variable in R has changed
```

```
lm(formula = salary ~ sex, data = salaries)

Residuals:
   Min     1Q Median     3Q    Max
-57290 -23502  -6828  19710 116455

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   115090       1587  72.503  < 2e-16 ***
sexFemale     -14088       5065  -2.782  0.00567 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 30030 on 395 degrees of freedom
Multiple R-squared:  0.01921,    Adjusted R-squared:  0.01673
F-statistic: 7.738 on 1 and 395 DF,  p-value: 0.005667
```

- In this case ***b0*** is the average salary for male is estimated to be 115090
- For female is estimated as a total of 115090 (***b0***) - 14088 (***b1***)= 101002
- p-value for ***sexFemale*** is very significant suggesting that there is a statistical evidence of a difference in average salary between the genders
- *SAME RESULT*

Multiple Linear Regression. Cases

- Suppose that, we want to investigate differences in salaries between males and females and also according to the rank.

- Also, the variable rank is categorical defined by three categories including three categories: *AssocProf*, *AsstProf* and Prof

- What's happen if we use ***rank*** as predictor added to the previous variable ***sex?***

Multiple Linear Regression. Cases

How affects the salary depending on sex and rank

- MLR-3: Salaries

```
# Añadimos las variables binarias al DataFrame
sex_rank = pd.concat([sexMale, rankd], axis = 1)
# Data for the model
X_simp = sex_rank
y_simp = salaries['salary']

# Add a constant to get an intercept
X_simp_sm = sm.add_constant(X_simp)

# Fit the resgression line using 'OLS'
model_simp = sm.OLS(y_simp, X_simp_sm).fit()
```

```
 OLS Regression Results
==============================================================================
Dep. Variable:                 salary   R-squared:                       0.397
Model:                            OLS   Adj. R-squared:                  0.392
Method:                 Least Squares   F-statistic:                     86.09
Date:                Fri, 16 Jun 2023   Prob (F-statistic):           7.84e-43
Time:                        14:49:00   Log-Likelihood:                -4559.0
No. Observations:                 397   AIC:                             9126.
Df Residuals:                     393   BIC:                             9142.
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       8.971e+04   4500.671     19.932      0.000    8.09e+04    9.86e+04
sexMale     4942.9511   4026.127      1.228      0.220   -2972.489    1.29e+04
AsstProf   -1.306e+04   4128.317     -3.164      0.002   -2.12e+04   -4944.911
Prof        3.246e+04   3307.632      9.813      0.000     2.6e+04     3.9e+04
==============================================================================
Omnibus:                       36.134   Durbin-Watson:                   1.737
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               54.229
Skew:                           0.623   Prob(JB):                     1.68e-12
Kurtosis:                       4.314   Cond. No.                         7.96
==============================================================================
```

- The ***adjusted R*** has increased
- The ***p-value*** of the model is significant
- The variable ***sex*** is not so important.

Multiple Linear Regression. Cases

The Whole model

- MLR-3: Salaries

```
# Creation of a new dataframe including dummies in sex and rank (using prefix)
data=pd.get_dummies(salaries, columns=['discipline','sex','rank'], prefix="dmy",drop_first=True)
```

```
 OLS Regression Results
==============================================================================
Dep. Variable:                 salary   R-squared:                       0.455
Model:                            OLS   Adj. R-squared:                  0.446
Method:                 Least Squares   F-statistic:                     54.20
Date:                Fri, 16 Jun 2023   Prob (F-statistic):           1.79e-48
Time:                        14:55:36   Log-Likelihood:                -4538.9
No. Observations:                 397   AIC:                             9092.
Df Residuals:                     390   BIC:                             9120.
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          7.886e+04   4990.326     15.803      0.000    6.91e+04    8.87e+04
yrs.since.phd   535.0583    240.994      2.220      0.027      61.248    1008.869
yrs.service    -489.5157    211.938     -2.310      0.021    -906.199     -72.833
dmy_B          1.442e+04   2342.875      6.154      0.000    9811.380     1.9e+04
dmy_Male       4783.4928   3858.668      1.240      0.216   -2802.901    1.24e+04
dmy_AsstProf  -1.291e+04   4145.278     -3.114      0.002   -2.11e+04   -4757.700
dmy_Prof       3.216e+04   3540.647      9.083      0.000    2.52e+04    3.91e+04
==============================================================================
Omnibus:                       46.385   Durbin-Watson:                   1.919
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               82.047
Skew:                           0.699   Prob(JB):                     1.53e-18
Kurtosis:                       4.733   Cond. No.                         183.
==============================================================================
```

```
B    216
A    181
Name: discipline, dtype: int64
```

	yrs.since.phd	yrs.service	salary	dmy_B	dmy_Male	dmy_AsstProf	dmy_Prof
0	19	18	139750	1	1	0	1
1	20	16	173200	1	1	0	1
2	4	3	79750	1	1	1	0
3	45	39	115000	1	1	0	1
4	40	41	141500	1	1	0	1

- The ***p-value*** of the model is significant
- The variable ***sex*** is not so important.

Multiple Linear Regression. Cases

- MLR-3: Salaries

The Whole model, but without sex

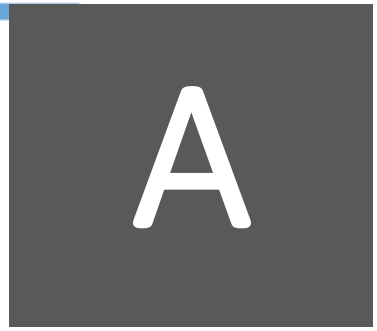
OLS Regression Results

Dep. Variable:	salary	R-squared:	0.453
Model:	OLS	Adj. R-squared:	0.446
Method:	Least Squares	F-statistic:	64.64
Date:	Fri, 16 Jun 2023	Prob (F-statistic):	4.51e-49
Time:	14:57:50	Log-Likelihood:	-4539.7
No. Observations:	397	AIC:	9091.
Df Residuals:	391	BIC:	9115.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	8.27e+04	3916.719	21.115	0.000	7.5e+04	9.04e+04
yrs.since.phd	534.6313	241.159	2.217	0.027	60.500	1008.762
yrs.service	-476.7179	211.831	-2.250	0.025	-893.189	-60.247
dmy_B	1.451e+04	2343.418	6.190	0.000	9897.875	1.91e+04
dmy_AsstProf	-1.283e+04	4147.669	-3.094	0.002	-2.1e+04	-4677.015
dmy_Prof	3.246e+04	3534.915	9.182	0.000	2.55e+04	3.94e+04

Omnibus:	47.406	Durbin-Watson:	1.900
Prob(Omnibus):	0.000	Jarque-Bera (JB):	83.795
Skew:	0.713	Prob(JB):	6.37e-19
Kurtosis:	4.741	Cond. No.	163.

- The ***adjusted R*** is the same
- The ***p-value*** of the model is significant
- The variable ***sex*** is not related to salary



Dictionary

Linear Regression - Dictionary

- **Bias** – sesgo
- **Bound** - cota
- **Dummy variable –** variable ficticia
- **Fit** - ajuste
- **Goodness of fit** – bondad de ajuste
- **Intercept –** ordenada en el origen
- **Joint distribution** – distribución conjunta
- **Least squares** – Mínimos cuadrados
- **Mean squared error** – error cuadrático medio
- **Sample** – muestra
- **Scatterplot-** nube de puntos
- **Slope** - pendiente
- **Straight line** – línea recta

comillas.edu