Lisbon School
of Economics
& Management
Universidade de Lisboa

Carlos J. Costa

REGRESSIONS

Regression

- Is a set of statistical processes for estimating the relationships among variables.
- Dependent variable,outcome variable, target
- Independent variables, predictor, covariates, or features

Regression

- simple regression/multivariate regression

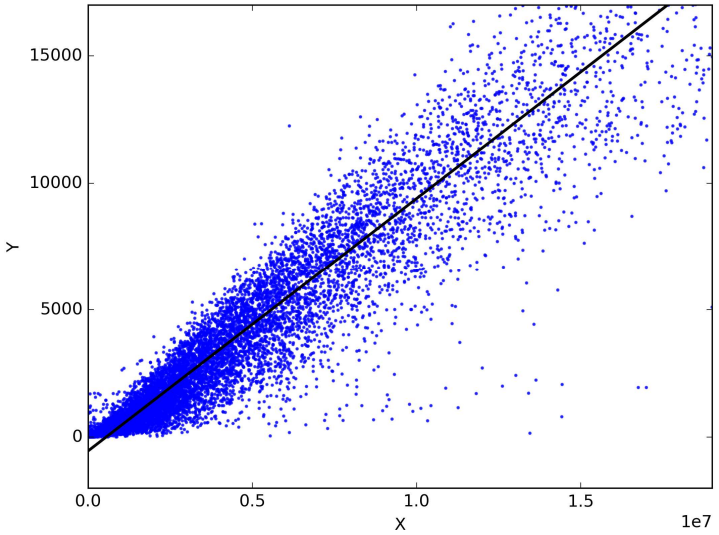
$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i.$$

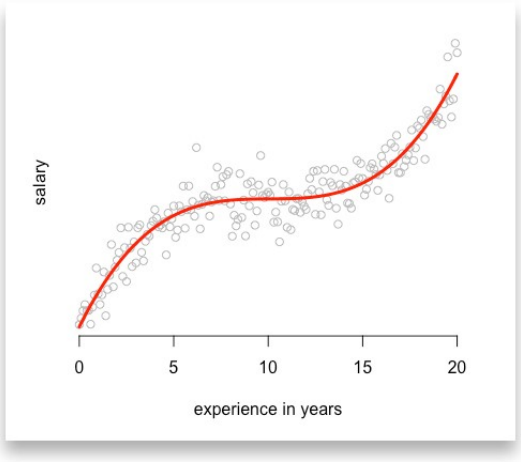
Regression

- .Linear/non linear

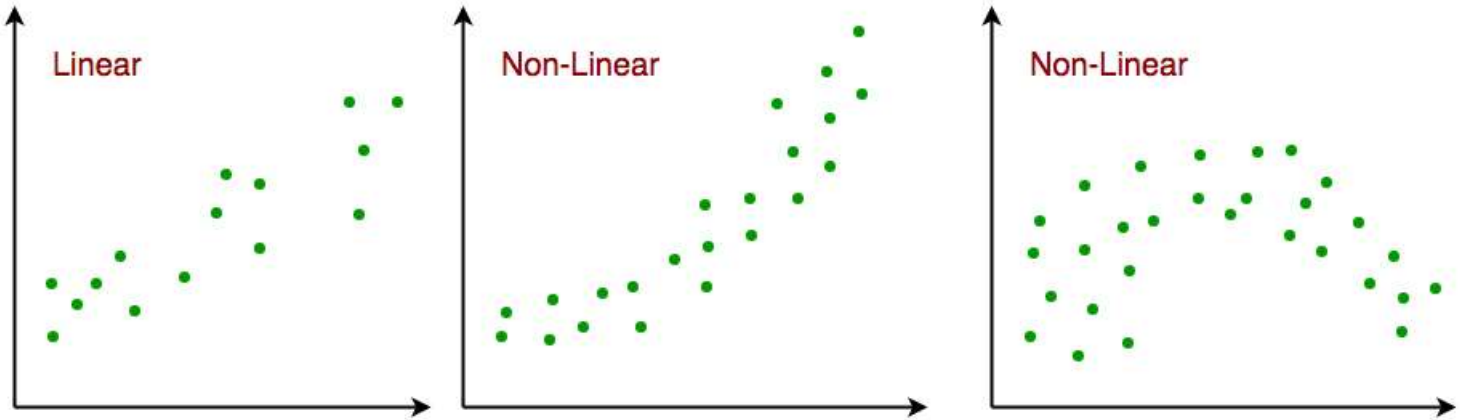
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n.$$



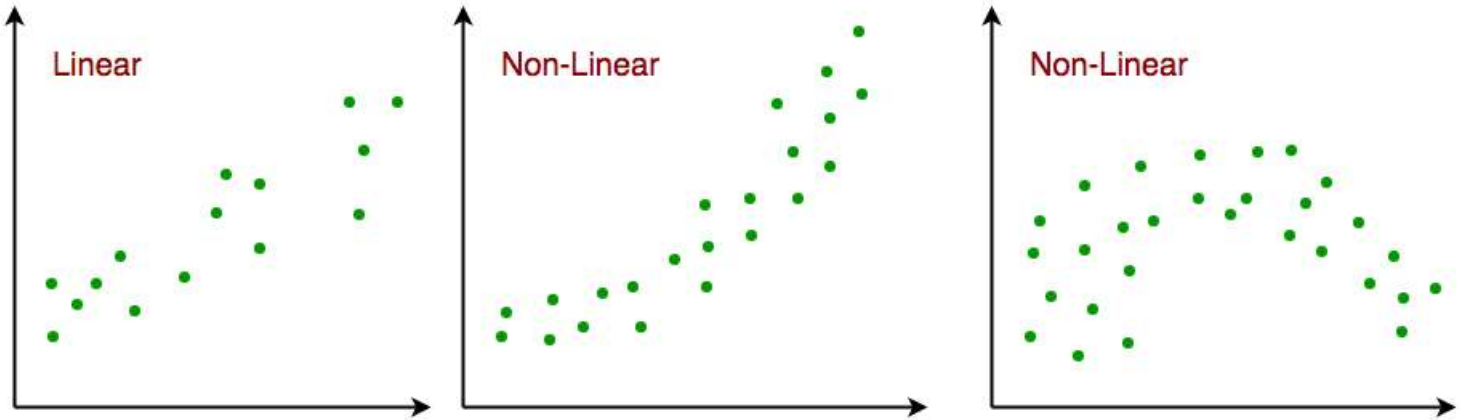
$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i, \; i = 1, \ldots, n.$$



Regression



Regression



```
import pandas as pd
import statsmodels.api as sm
import matplotlib.pyplot as plt
```

```
XY=pd.read_csv("dadosMundo.csv")
XY=XY.dropna()
```

```
X1=XY.drop(['Country'], axis=1)
X=X1.drop(['Internet users in the per 1000 people'], axis=1)
Y=XY['Internet users in the per 1000 people']
```

```
X=sm.add_constant(X)
model = sm.OLS(Y,X).fit()
```

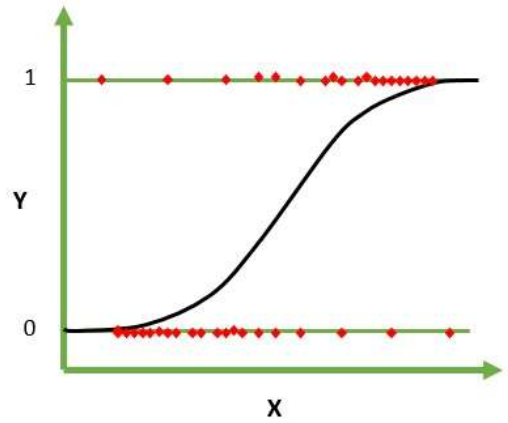
```
model.summary()
```

Dep. Variable:	Internet users in the per 1000 people	**R-squared:**	0.848
Model:	OLS	**Adj. R-squared:**	0.806
Method:	Least Squares	**F-statistic:**	20.12
Date:	Sun, 17 Oct 2021	**Prob (F-statistic):**	6.99e-22
Time:	17:36:59	**Log-Likelihood:**	-527.87
No. Observations:	93	**AIC:**	1098.
Df Residuals:	72	**BIC:**	1151.
Df Model:	20		
Covariance Type:	nonrobust		

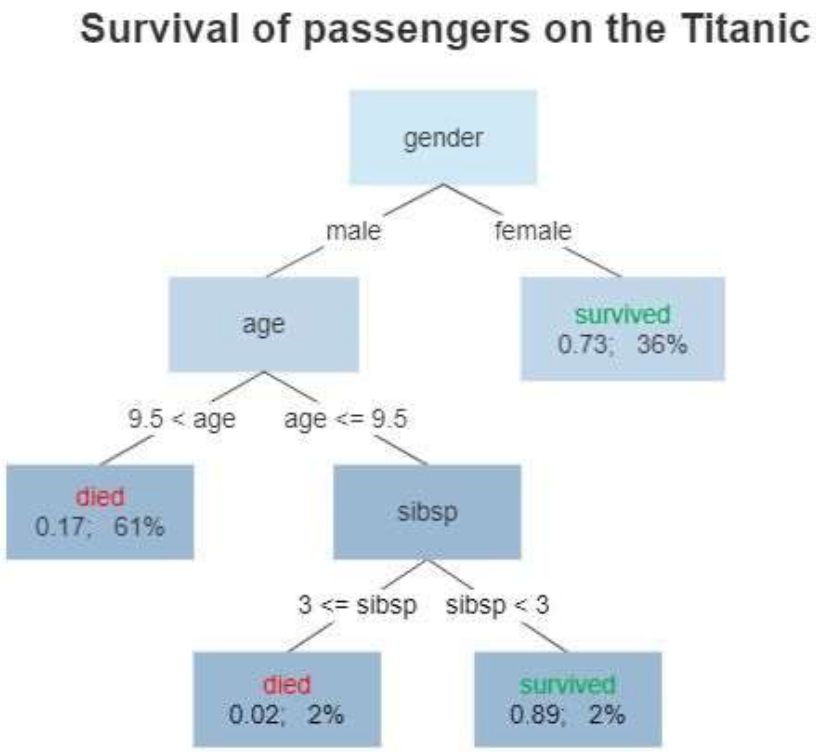
	coef	std err	t	P>|t|	[0.025	0.975]
const	333.4493	442.222	0.754	0.453	-548.105	1215.004
Area_km2	7.192e-08	6.32e-06	0.011	0.991	-1.25e-05	1.27e-05
Population	-9.251e-09	6.38e-08	-0.145	0.885	-1.37e-07	1.18e-07
Birth rate(births/1000 population)	-4.6573	8.398	-0.555	0.581	-21.399	12.085
Death rate(deaths/1000 population)	-4.6921	6.981	-0.672	0.504	-18.608	9.224
Electricity – consumption(kWh) per capita	0.0068	0.010	0.675	0.502	-0.013	0.027
Electricity – production(kWh) per capita	-0.0045	0.009	-0.524	0.602	-0.021	0.013
GDPpercapita	0.0021	0.003	0.686	0.495	-0.004	0.008
GDP - real growth rate(%)	-4.5051	4.864	-0.926	0.357	-14.202	5.191
Industrial production growth rate(%)	2.8341	2.795	1.014	0.314	-2.738	8.406

Logistics Regression

- A regression that having binary dependent variable
- in its basic form, uses a logistic function to model a binary dependent variable



Decision Tree



- Decision tree builds classification or regression models in the form of a tree structure.
- It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed.
- The final result is a tree with decision nodes and leaf nodes.

Random Forest

- are an ensemble learning method for classification, regression and other tasks
- operates by constructing a multitude of decision trees at training time
- outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.