REDES NEURAIS E DEEP LEARNING

REGRESSÃO

DIEGO RODRIGUES DSC

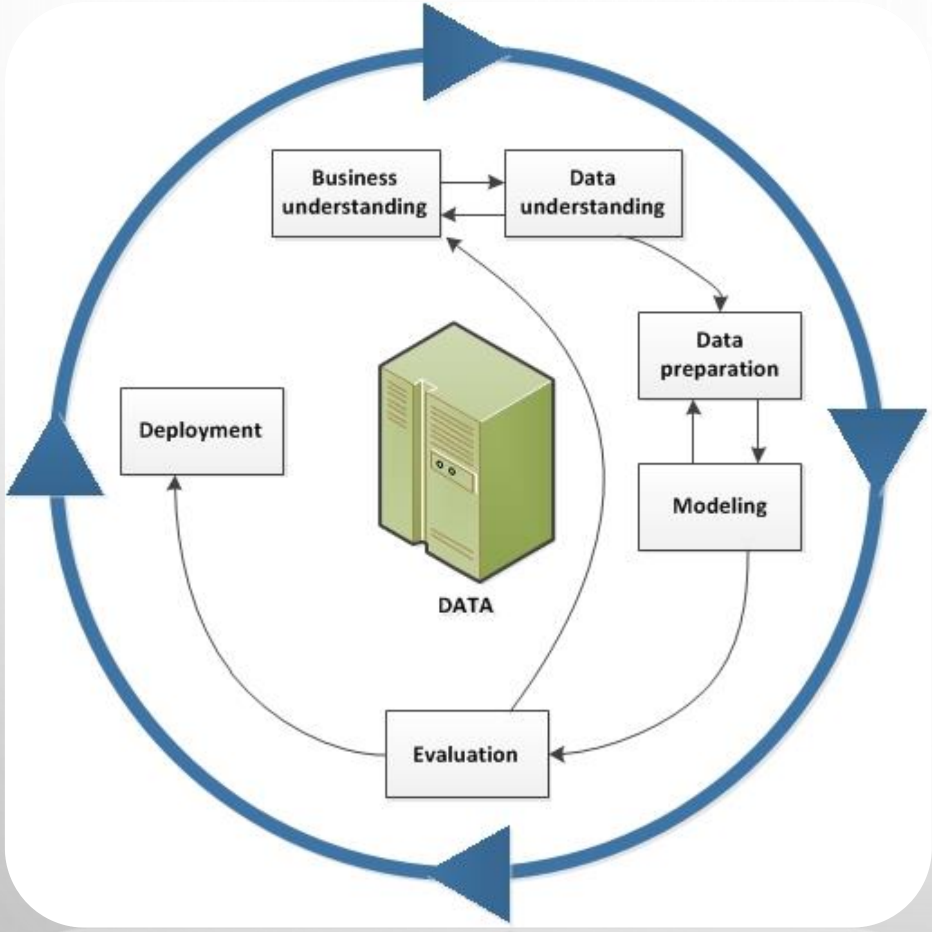
INFNET

Bloco	Matéria	Calendário	Avaliação
Treinamento Clássico	Introdução	06/10	
	Classificação	08, *13	
	Regressão	27, *29	
	Agrupamento	03/11, *05	
	Séries Temporais	10, *12	<Modelo Clássico>
Redes Profundas	Deep Feed Forward	17, *19	
	Visão Computacional	24, *26	
	Autoencoders	01/12, *03	<Modelo Profundo>
Treinamento Moderno	Transfer Learning	08, *10	
	Sequências	15, *17	<Modelo Avançado>
	Modelos Generativos	<COMBINAR>	

REGRESSÃO

- REGRESSÃO / APROXIMAÇÃO
- REGRESSÃO LINEAR
- REGRESSÃO COM REDE NEURAL
- FIGURAS DE MÉRITO

PARTE 1 : TEORIA



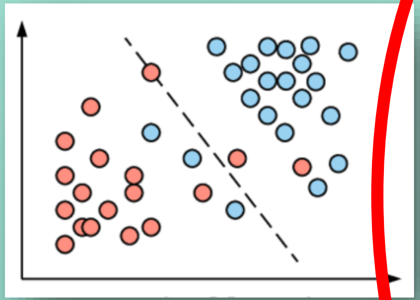
CROSS INDUSTRY PROCESS FOR DATA MINING (CRISP-DM)

BUSINESS UNDERSTANDING

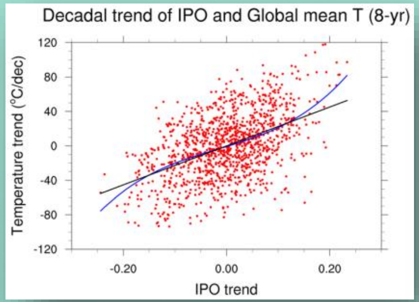
APRENDIZADO SUPERVISIONADO

APRENDIZADO NÃO-SUPERVISIONADO

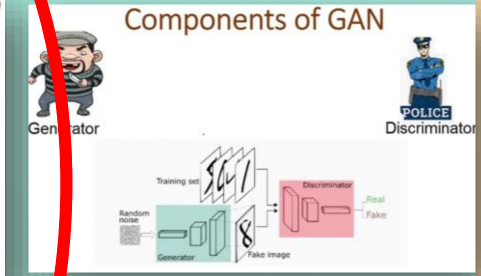
APRENDIZADO POR REFORÇO



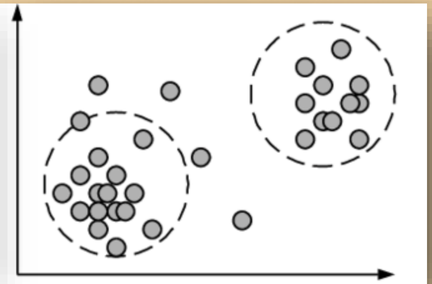
CLASSIFICAÇÃO



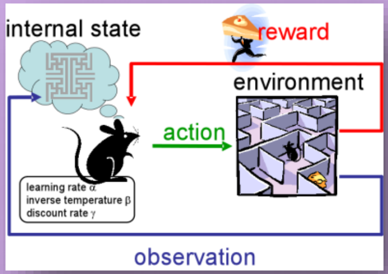
REGRESSÃO



GENERATIVO



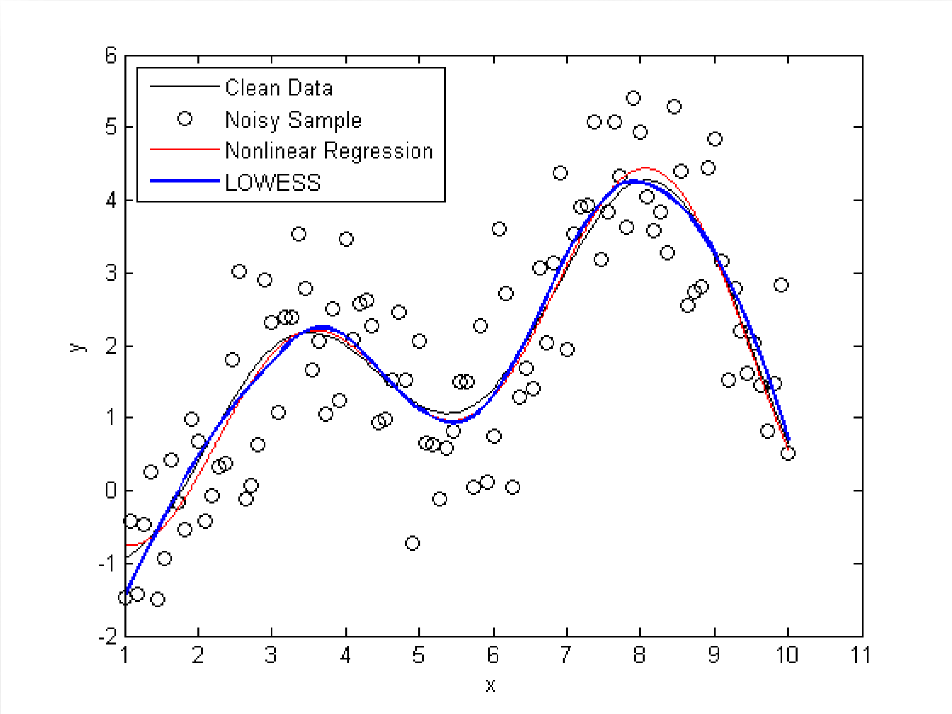
AGRUPAMENTO



REFORÇO

REGRESSÃO

O objetivo da regressão é **modelar as relações funcionais** entre dois conjuntos de variáveis.



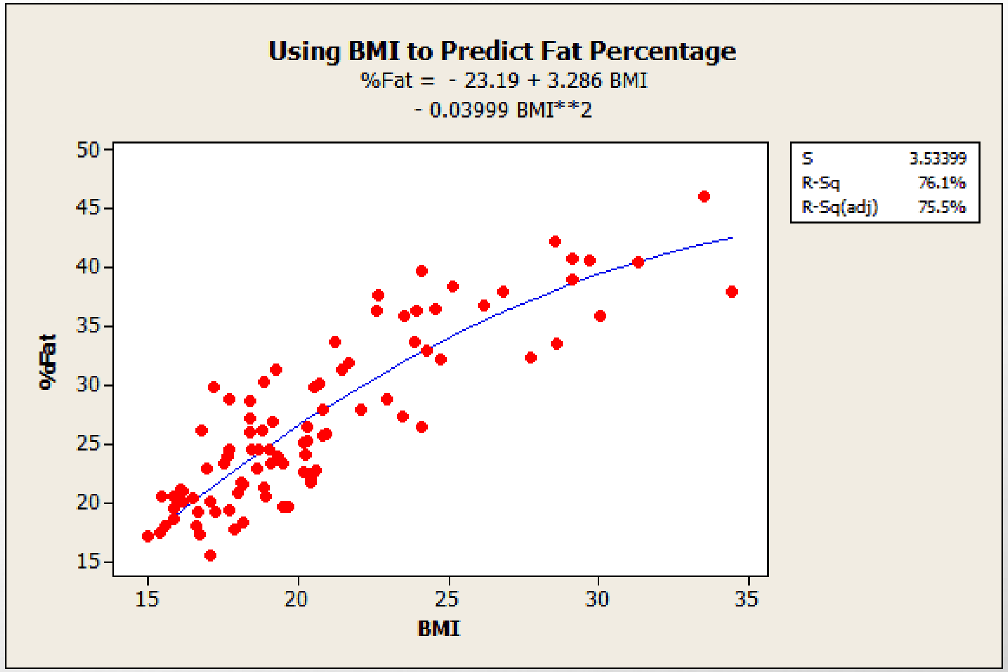
As vezes quando o mundo não é linear & gaussiano...

As variáveis que representam as causas são chamadas de **variáveis independentes,** e as variáveis cujo objetivo é prever, são chamadas **variáveis dependentes.**

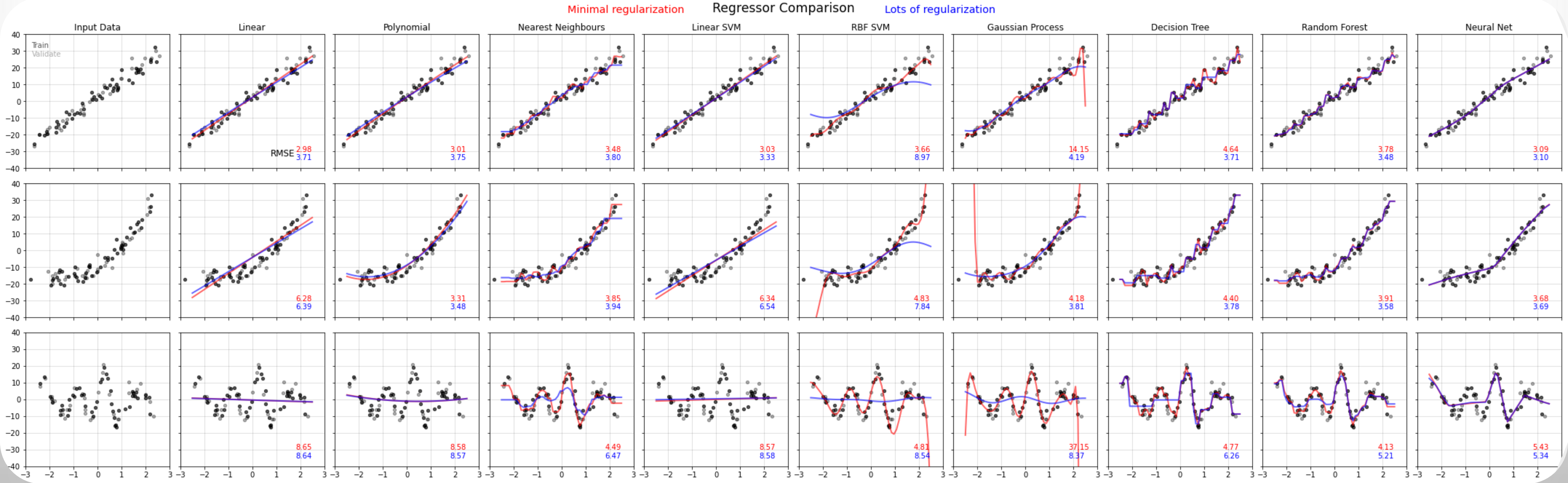
Então, uma **regressão** é um modelo utilizado para prever **uma ou mais variáveis dependentes,** baseado em causas, ou variáveis independentes.

Modelos de Regressão

1) **Regressão Linear**
2) Regressão Não-Linear
3) Processos Gaussianos
4) Máquina de Vetores Suporte
5) **Redes Neurais**

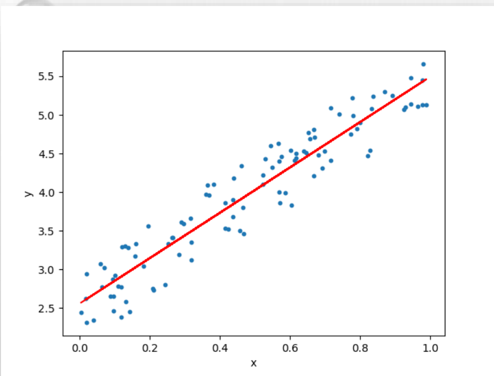


Algoritmos de regressão geralmente são modelados combinando uma **parte determinística e uma parte aleatória.** Os parâmetros correspondente à parte determinística são encontrados utilizando estimadores como máxima verossimilhança ou máximo a posteriori (MAP).

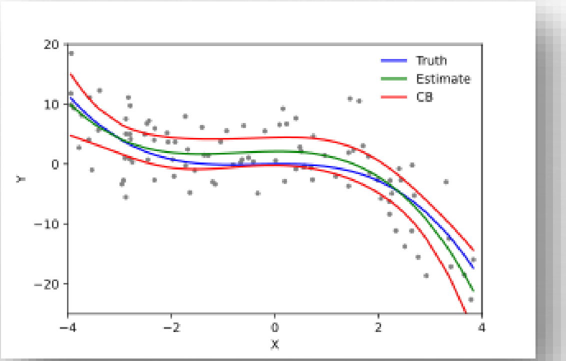

Minimal regularization
Regressor Comparison
Lots of regularization
Input Data
Linear
Polynomial
Nearest Neighbours
Linear SVM
RBF SVM
Gaussian Process
Decision Tree
Random Forest
Neural Net
Train
Validate
RMSE
2.98 3.71
3.01 3.75
3.48 3.80
3.03 3.33
3.66 8.97
14.15 4.19
4.64 3.71
3.78 3.48
3.09 3.10
6.28 6.39
3.31 3.48
3.85 3.94
6.34 6.54
4.83 7.84
4.18 3.81
4.40 3.78
3.91 3.58
3.68 3.69
8.65 8.64
8.58 8.57
4.49 6.47
8.57 8.58
4.81 8.54
37.15 8.37
4.77 6.26
4.13 5.21
5.43 5.34

O APROXIMADOR UNIVERSAL

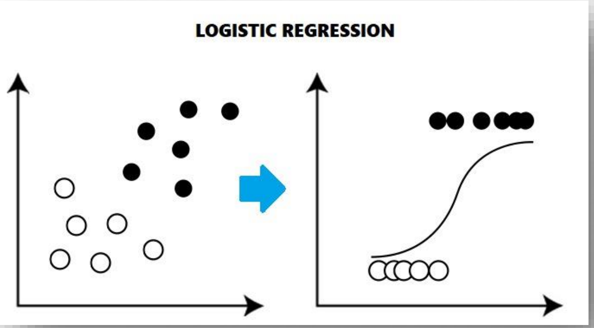
$$Y = \underbrace{F(X)}_{\text{Parte Determinística}} + \underbrace{\varepsilon}_{\text{Parte Estocástica}}$$



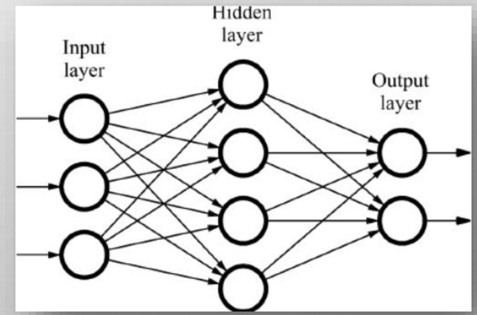
$$Y = \alpha^T x + \varepsilon$$



$$Y = X\alpha + \varepsilon$$



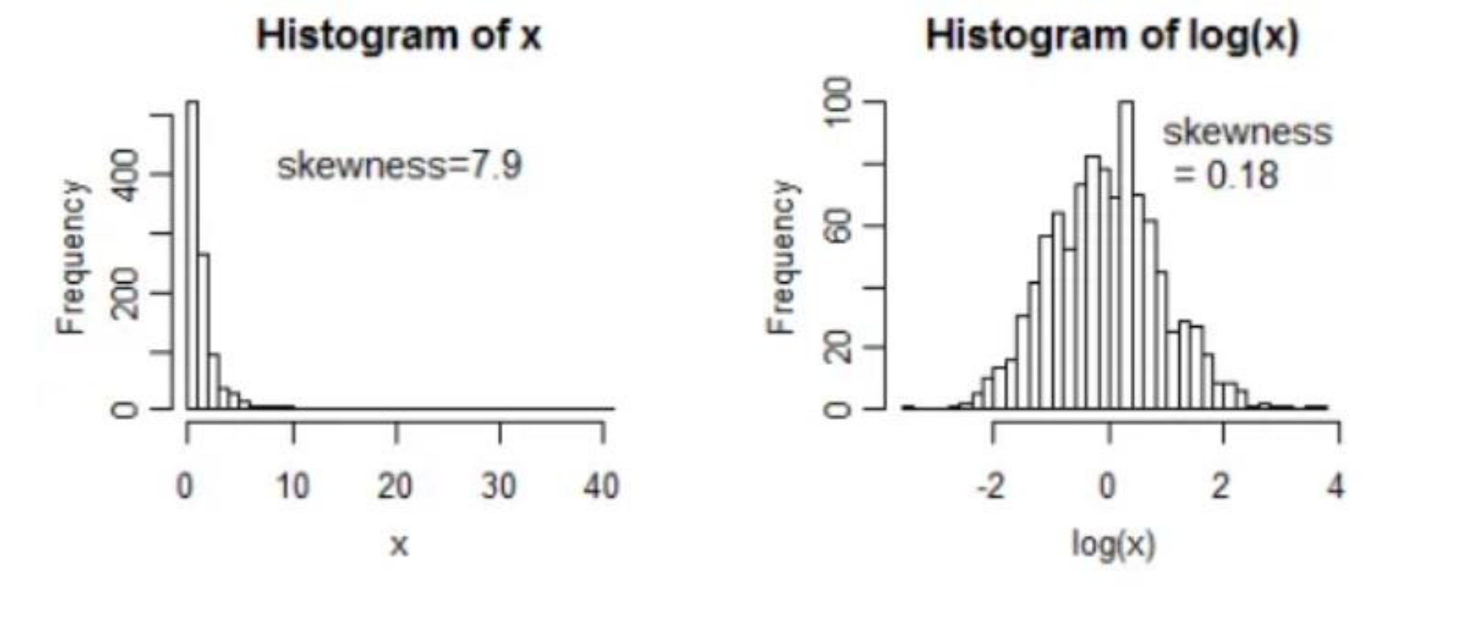
$$Y = \frac{1}{1 + e^{\alpha^t x + \varepsilon}}$$



$$Y = \varphi(x) + \varepsilon$$

DATA UNDERSTANDING & PREPARATION

NORMALIZAÇÃO



Example distribution before (left) and after (right) log transformation

https://medium.com/@isalindgren313/transformations-scaling-and-normalization-420b2be12300

Transformar as variáveis originais por funções, facilitando o problema numérico de otimização e ao mesmo tempo inserindo "não-linearidades" para resolver um problema não-linear de forma linear.

NORMALIZAÇÃO

TRANSFORMATION	USE IF	LIMITATIONS	SPSS EXAMPLES
Square/Cube Root	Variable shows positive skewness Residuals show positive heteroscedasticity Variable contains frequency counts	Square root only applies to positive values	compute newvar = sqrt(oldvar). compute newvar = oldvar**(1/3).
Logarithmic	Distribution is positively skewed	Ln and log10 only apply to positive values	compute newvar = ln(oldvar). compute newvar = lg10(oldvar).
Power	Distribution is negatively skewed	(None)	compute newvar = oldvar**3.
Inverse	Variable has platykurtic distribution	Can't handle zeroes	compute newvar = 1 / oldvar.
Hyperbolic Arcsine	Distribution is positively skewed	(None)	compute newvar = ln(oldvar + sqrt(oldvar**2 + 1)).
Arcsine	Variable contains proportions	Can't handle absolute values > 1	compute newvar = arsin(oldvar).

https://www.spss-tutorials.com/normalizing-variable-transformations/

MODELING

Regressão Linear : Modelo Matemático

Formulation [edit]

Given a data set $\{y_i,\, x_{i1}, \ldots, x_{ip}\}_{i=1}^{n}$ of *n* statistical units, a linear regression model assumes that the relationship between the dependent variable *y* and the vector of regressors **x** is linear. This relationship is modeled through a *disturbance term* or *error variable* *ε* — an unobserved random variable that adds "noise" to the linear relationship between the dependent variable and regressors. Thus the model takes the form

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^{\mathsf T} \boldsymbol\beta + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where $^{\mathsf T}$ denotes the transpose, so that $\mathbf{x}_i^{\mathsf T}\boldsymbol\beta$ is the inner product between vectors $\mathbf{x}_i$ and $\boldsymbol\beta$.

Often these *n* equations are stacked together and written in matrix notation as

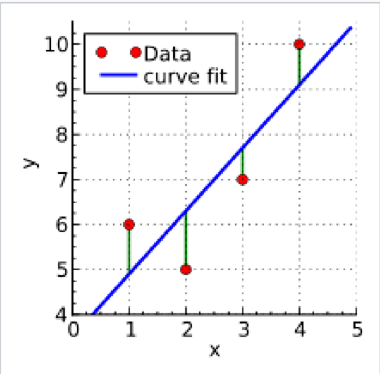
$$\mathbf{y} = \mathbf{X}\boldsymbol\beta + \boldsymbol\varepsilon,$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix},$$

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^{\mathsf T} \\ \mathbf{x}_2^{\mathsf T} \\ \vdots \\ \mathbf{x}_n^{\mathsf T} \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix},$$

$$\boldsymbol\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \boldsymbol\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$



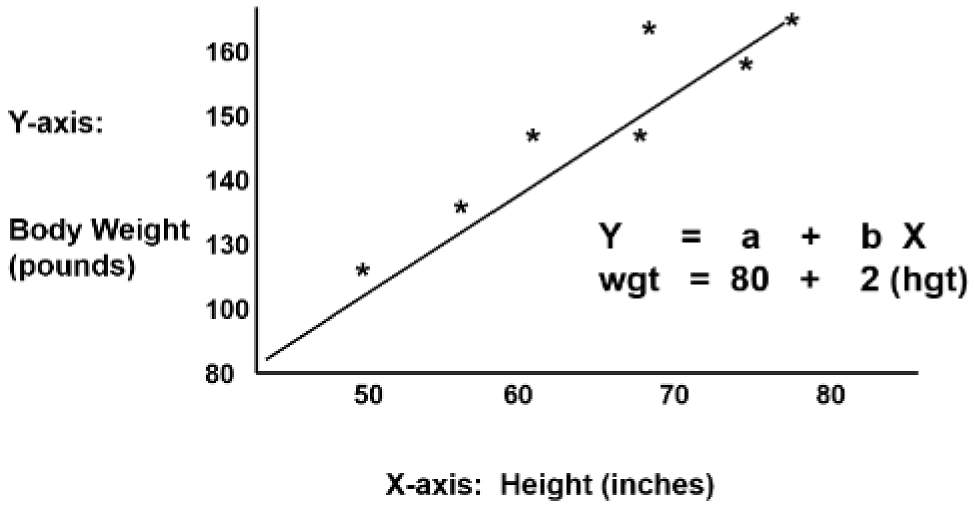
In linear regression, the observations (red) are assumed to be the result of random deviations (green) from an underlying relationship (blue) between a dependent variable (*y*) and an independent variable (*x*).

$$y = \sum_{i}^{V} \beta_i x_i + \varepsilon$$

Exemplo I: Altura e Peso

Simple Linear Regression

Regression analysis makes use of mathematical models to describe relationships. For example, suppose that height was the only determinant of body weight. If we were to plot height (the independent or 'predictor' variable) as a function of body weight (the dependent or 'outcome' variable), we might see a very linear relationship, as illustrated below.



We could also describe this relationship with the equation for a line, Y = a + b(x), where 'a' is the Y-intercept and 'b' is the slope of the line. We could use the equation to predict weight if we knew an individual's height. In this example, if an individual was 70 inches tall, we would predict his weight to be:

$$\text{Weight} = 80 + 2 \times (70) = 220 \text{ lbs.}$$

In this simple linear regression, we are examining the impact of one independent variable on the outcome. If height were the only determinant of body weight, we would expect that the points for individual subjects would lie close to the line. However, if there were other factors (independent variables) that influenced body weight besides height (e.g., age, calorie intake, and exercise level), we might expect that the points for individual subjects would be more loosely scattered around the line, since we are only taking height into account.

https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704-EP713_MultivariableMethods/

Premissas I

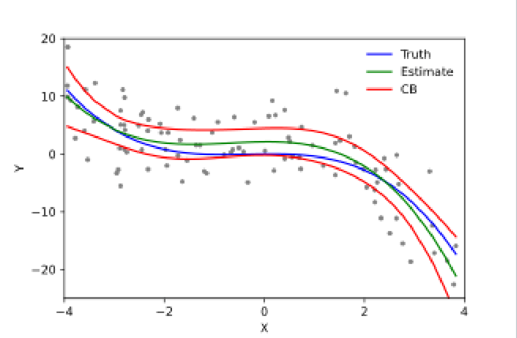
Assumptions [edit]

See also: Ordinary least squares § Assumptions

Standard linear regression models with standard estimation techniques make a number of assumptions about the predictor variables, the response variables and their relationship. Numerous extensions have been developed that allow each of these assumptions to be relaxed (i.e. reduced to a weaker form), and in some cases eliminated entirely. Generally these extensions make the estimation procedure more complex and time-consuming, and may also require more data in order to produce an equally precise model.

The following are the major assumptions made by standard linear regression models with standard estimation techniques (e.g. ordinary least squares):

- **Weak exogeneity**. This essentially means that the predictor variables *x* can be treated as fixed values, rather than random variables. This means, for example, that the predictor variables are assumed to be error-free—that is, not contaminated with measurement errors. Although this assumption is not realistic in many settings, dropping it leads to significantly more difficult errors-in-variables models.
- **Linearity**. This means that the mean of the response variable is a linear combination of the parameters (regression coefficients) and the predictor variables. Note that this assumption is much less restrictive than it may at first seem. Because the predictor variables are treated as fixed values (see above), linearity is really only a restriction on the parameters. The predictor variables themselves can be arbitrarily transformed, and in fact multiple copies of the same underlying predictor variable can be added, each one transformed differently. This technique is used, for example, in polynomial regression, which uses linear regression to fit the response variable as an arbitrary polynomial function (up to a given degree) of a predictor variable. With this much flexibility, models such as polynomial regression often have "too much power", in that they tend to overfit the data. As a result, some kind of regularization must typically be used to prevent unreasonable solutions coming out of the estimation process. Common examples are ridge regression and lasso regression. Bayesian linear regression can also be used, which by its nature is more or less immune to the problem of overfitting. (In fact, ridge regression and lasso regression can both be viewed as special cases of Bayesian linear regression, with particular types of prior distributions placed on the regression coefficients.)

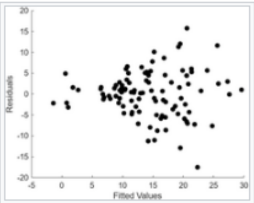


Example of a cubic polynomial regression, which is a type of linear regression. Although *polynomial regression* fits a nonlinear model to the data, as a statistical estimation problem it is linear, in the sense that the regression function E(*y* | *x*) is linear in the unknown parameters that are estimated from the data. For this reason, polynomial regression is considered to be a special case of multiple linear regression.

- Additivity: $f(x + y) = f(x) + f(y)$.
- Homogeneity of degree 1: $f(\alpha x) = \alpha f(x)$ for all α.

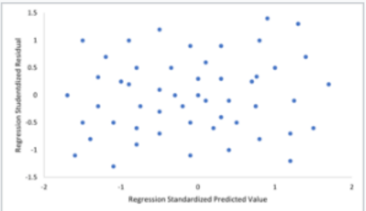
Premissas II

- **Constant variance** (a.k.a. **homoscedasticity**). This means that the variance of the errors does not depend on the values of the predictor variables. Thus the variability of the responses for given fixed values of the predictors is the same regardless of how large or small the responses are. This is often not the case, as a variable whose mean is large will typically have a greater variance than one whose mean is small. For example, a person whose income is predicted to be \$100,000 may easily have an actual income of \$80,000 or \$120,000—i.e., a standard deviation of around \$20,000—while another person with a predicted income of \$10,000 is unlikely to have the same \$20,000 standard deviation, since that would imply their actual income could vary anywhere between −\$10,000 and \$30,000. (In fact, as this shows, in many cases—often the same cases where the assumption of normally distributed errors fails—the variance or standard deviation should be predicted to be proportional to the mean, rather than constant.) The absence of homoscedasticity is called heteroscedasticity. In order to check this assumption, a plot of residuals versus predicted values (or the values of each individual predictor) can be examined for a "fanning effect" (i.e., increasing or decreasing vertical spread as one moves left to right on the plot). A plot of the absolute or squared residuals versus the predicted values (or each predictor) can also be examined for a trend or curvature. Formal tests can also be used; see Heteroscedasticity. The presence of heteroscedasticity will result in an overall "average" estimate of variance being used instead of one that takes into account the true variance structure. This leads to less precise (but in the case of ordinary least squares, not biased) parameter estimates and biased standard errors, resulting in misleading tests and interval estimates. The mean squared error for the model will also be wrong. Various estimation techniques including weighted least squares and the use of heteroscedasticity-consistent standard errors can handle heteroscedasticity in a quite general way. Bayesian linear regression techniques can also be used when the variance is assumed to be a function of the mean. It is also possible in some cases to fix the problem by applying a transformation to the response variable (e.g., fitting the logarithm of the response variable using a linear regression model, which implies that the response variable itself has a log-normal distribution rather than a normal distribution).

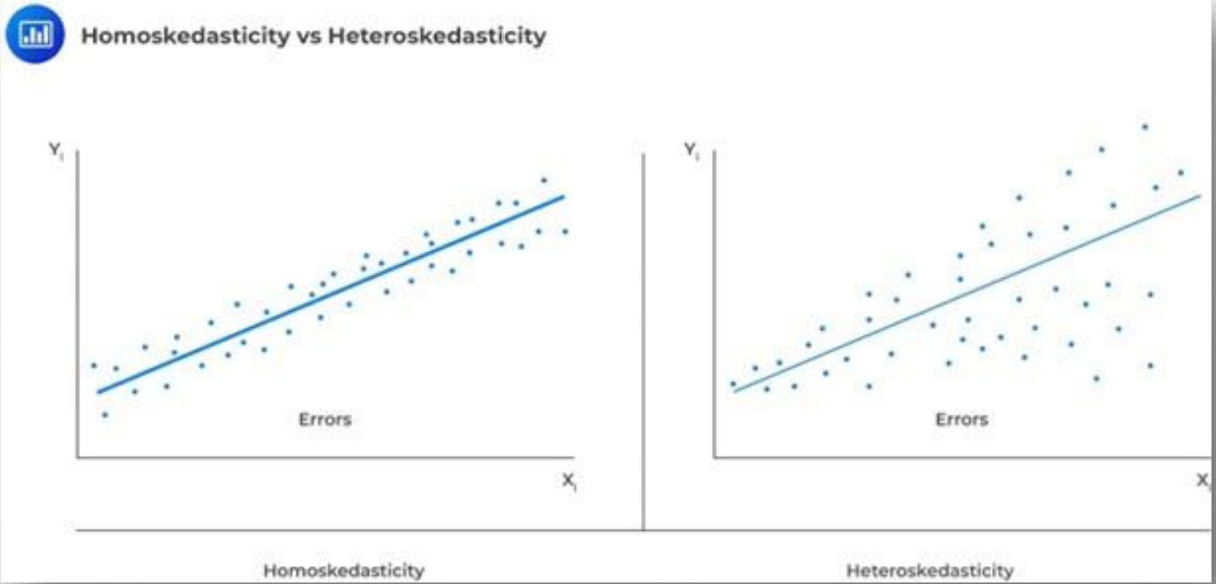


Visualization of heteroscedasticity in a scatter plot against 100 random fitted values using Matlab

- **Independence of errors**. This assumes that the errors of the response variables are uncorrelated with each other. (Actual statistical independence is a stronger condition than mere lack of correlation and is often not needed, although it can be exploited if it is known to hold.) Some methods such as generalized least squares are capable of handling correlated errors, although they typically require significantly more data unless some sort of regularization is used to bias the model towards assuming uncorrelated errors. Bayesian linear regression is a general way of handling this issue.
- **Lack of perfect multicollinearity** in the predictors. For standard least squares estimation methods, the design matrix X must have full column rank p; otherwise perfect multicollinearity exists in the predictor variables, meaning a linear relationship exists between two or more predictor variables. This can be caused by accidentally duplicating a variable in the data, using a linear transformation of a variable along with the original (e.g., the same temperature measurements expressed in Fahrenheit and Celsius), or including a linear combination of multiple variables in the model, such as their mean. It can also happen if there is too little data available compared to the number of parameters to be estimated (e.g., fewer data points than regression coefficients). Near violations of this assumption, where predictors are highly but not perfectly correlated, can reduce the precision of parameter estimates (see Variance inflation factor). In the case of perfect multicollinearity, the parameter vector $\boldsymbol{\beta}$ will be non-identifiable—it has no unique solution. In such a case, only some of the parameters can be identified (i.e., their values can only be estimated within some linear subspace of the full parameter space $\mathbf{R}^p$). See partial least squares regression. Methods for fitting linear models with multicollinearity have been developed,[5][6][7][8] some of which require additional assumptions such as "effect sparsity"—that a large fraction of the effects are exactly zero. Note that the more computationally expensive iterated algorithms for parameter estimation, such as those used in generalized linear models, do not suffer from this problem.



To check for violations of the assumptions of linearity, constant variance, and independence of errors within a linear regression model, the residuals are typically plotted against the predicted values (or each of the individual predictors). An apparently random scatter of points about the horizontal midline at 0 is ideal, but cannot rule out certain kinds of violations such as autocorrelation in the errors or their correlation with one or more covariates.



Encontrando os Coeficientes : Mínimos Quadrados Ordinários

Pseudo-inversa de Moore-Penrose

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

Linear model [edit]

Main article: Linear regression model

Suppose the data consists of n observations $\{\mathbf{x}_i, y_i\}_{i=1}^n$. Each observation i includes a scalar response y_i and a column vector $\mathbf{x}_i$ of p parameters (regressors), i.e., $\mathbf{x}_i = [x_{i1}, x_{i2}, \dots, x_{ip}]^{\mathrm{T}}$. In a linear regression model, the response variable, y_i, is a linear function of the regressors:

$$y_i = \beta_1\, x_{i1} + \beta_2\, x_{i2} + \cdots + \beta_p\, x_{ip} + \varepsilon_i,$$

or in vector form,

$$y_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \varepsilon_i,$$

where $\mathbf{x}_i$, as introduced previously, is a column vector of the i-th observation of all the explanatory variables; $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters; and the scalar ε_i represents unobserved random variables (errors) of the i-th observation. ε_i accounts for the influences upon the responses y_i from sources other than the explanatory variables $\mathbf{x}_i$. This model can also be written in matrix notation as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\mathbf{y}$ and $\boldsymbol{\varepsilon}$ are $n \times 1$ vectors of the response variables and the errors of the n observations, and $\mathbf{X}$ is an $n \times p$ matrix of regressors, also sometimes called the design matrix, whose row i is $\mathbf{x}_i^{\mathrm{T}}$ and contains the i-th observations on all the explanatory variables.

Typically, a constant term is included in the set of regressors $\mathbf{X}$, say, by taking $x_{i1} = 1$ for all $i = 1, \dots, n$. The coefficient β_1 corresponding to this regressor is called the *intercept*. Without the intercept, the fitted line is forced to cross the origin when $x_i = \vec{0}$.

Regressors do not have to be independent: there can be any desired relationship between the regressors (so long as it is not a linear relationship). For instance, we might suspect the response depends linearly both on a value and its square; in which case we would include one regressor whose value is just the square of another regressor. In that case, the model would be *quadratic* in the second regressor, but none-the-less is still considered a *linear* model because the model *is* still linear in the parameters ($\boldsymbol{\beta}$).

Matrix/vector formulation [edit]

Consider an overdetermined system

$$\sum_{j=1}^{p} x_{ij}\beta_j = y_i,\ (i = 1, 2, \dots, n),$$

of n linear equations in p unknown coefficients, $\beta_1, \beta_2, \dots, \beta_p$, with $n > p$. This can be written in matrix form as

$$\mathbf{X}\boldsymbol{\beta} = \mathbf{y},$$

where

$$\mathbf{X} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}, \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}.$$

(Note: for a linear model as above, not all elements in $\mathbf{X}$ contains information on the data points. The first column is populated with ones, $X_{i1} = 1$. Only the other columns contain actual data. So here p is equal to the number of regressors plus one).

Such a system usually has no exact solution, so the goal is instead to find the coefficients $\boldsymbol{\beta}$ which fit the equations "best", in the sense of solving the quadratic minimization problem

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\arg\min}\, S(\boldsymbol{\beta}),$$

where the objective function S is given by

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left| y_i - \sum_{j=1}^{p} X_{ij}\beta_j \right|^2 = \left\| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \right\|^2.$$

A justification for choosing this criterion is given in Properties below. This minimization problem has a unique solution, provided that the p columns of the matrix $\mathbf{X}$ are linearly independent, given by solving the so-called *normal equations*:

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})\,\hat{\boldsymbol{\beta}} = \mathbf{X}^{\mathrm{T}}\mathbf{y}\,.$$

The matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is known as the *normal matrix* or Gram matrix and the matrix $\mathbf{X}^{\mathrm{T}}\mathbf{y}$ is known as the moment matrix of regressand by regressors.[2] Finally, $\hat{\boldsymbol{\beta}}$ is the coefficient vector of the least-squares hyperplane, expressed as

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}.$$

or

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{\varepsilon}.$$

Mínimos Quadrados Ordinários: Premissas

Classical linear regression model [edit]

The classical model focuses on the "finite sample" estimation and inference, meaning that the number of observations *n* is fixed. This contrasts with the other approaches, which study the asymptotic behavior of OLS, and in which the number of observations is allowed to grow to infinity.

- **Correct specification**. The linear functional form must coincide with the form of the actual data-generating process.
- **Strict exogeneity**. The errors in the regression should have conditional mean zero:[16]

  $$\operatorname{E}[\,\varepsilon \mid X\,] = 0.$$

  The immediate consequence of the exogeneity assumption is that the errors have mean zero: $\operatorname{E}[\varepsilon] = 0$ (for the law of total expectation), and that the regressors are uncorrelated with the errors: $\operatorname{E}[X^{\mathsf T}\varepsilon] = 0$.

  The exogeneity assumption is critical for the OLS theory. If it holds then the regressor variables are called *exogenous*. If it doesn't, then those regressors that are correlated with the error term are called *endogenous*,[17] and the OLS estimator becomes biased. In such case the method of instrumental variables may be used to carry out inference.
- **No linear dependence**. The regressors in X must all be linearly independent. Mathematically, this means that the matrix X must have full column rank almost surely:[18]

  $$\Pr\!\big[\,\operatorname{rank}(X) = p\,\big] = 1.$$

  Usually, it is also assumed that the regressors have finite moments up to at least the second moment. Then the matrix $Q_{xx} = \operatorname{E}[X^{\mathsf T}X / n]$ is finite and positive semi-definite.

  When this assumption is violated the regressors are called linearly dependent or perfectly multicollinear. In such case the value of the regression coefficient β cannot be learned, although prediction of y values is still possible for new values of the regressors that lie in the same linearly dependent subspace.
- **Spherical errors**:[18]

  $$\operatorname{Var}[\,\varepsilon \mid X\,] = \sigma^2 I_n,$$

  where I_n is the identity matrix in dimension n, and σ^2 is a parameter which determines the variance of each observation. This σ^2 is considered a nuisance parameter in the model, although usually it is also estimated. If this assumption is violated then the OLS estimates are still valid, but no longer efficient.

  It is customary to split this assumption into two parts:

  - **Homoscedasticity**: $\operatorname{E}[\,\varepsilon_i^2 \mid X\,] = \sigma^2$, which means that the error term has the same variance σ^2 in each observation. When this requirement is violated this is called heteroscedasticity, in such case a more efficient estimator would be weighted least squares. If the errors have infinite variance then the OLS estimates will also have infinite variance (although by the law of large numbers they will nonetheless tend toward the true values so long as the errors have zero mean). In this case, robust estimation techniques are recommended.
  - **No autocorrelation**: the errors are uncorrelated between observations: $\operatorname{E}[\,\varepsilon_i\varepsilon_j \mid X\,] = 0$ for $i \neq j$. This assumption may be violated in the context of time series data, panel data, cluster samples, hierarchical data, repeated measures data, longitudinal data, and other data with dependencies. In such cases generalized least squares provides a better alternative than the OLS. Another expression for autocorrelation is *serial correlation*.
- **Normality**. It is sometimes additionally assumed that the errors have normal distribution conditional on the regressors:[19]

  $$\varepsilon \mid X \sim \mathcal{N}(0, \sigma^2 I_n).$$

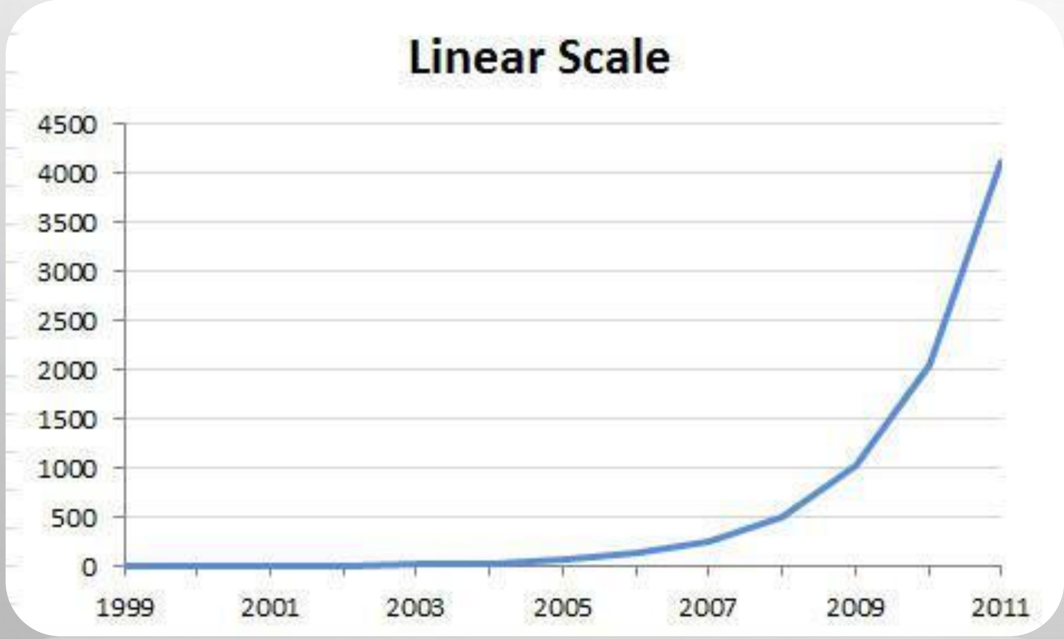
  This assumption is not needed for the validity of the OLS method, although certain additional finite-sample properties can be established in case when it does (especially in the area of hypotheses testing). Also when the errors are normal, the OLS estimator is equivalent to the maximum likelihood estimator (MLE), and therefore it is asymptotically efficient in the class of all regular estimators. Importantly, the normality assumption applies only to the error terms; contrary to a popular misconception, the response (dependent) variable is not required to be normally distributed.[20]

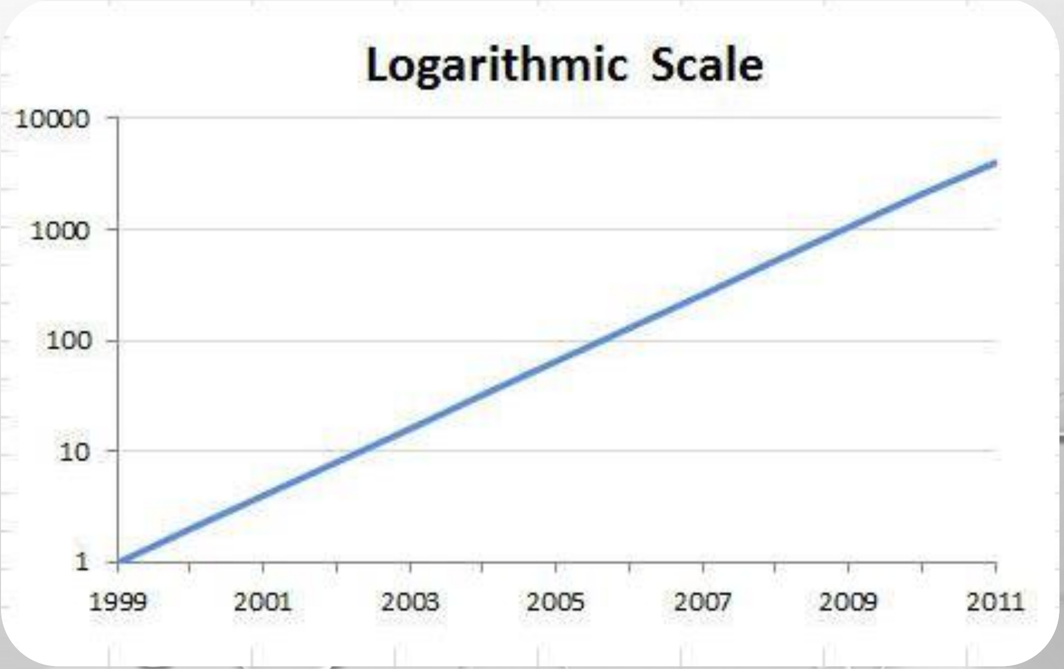
MODELO LOG-LINEAR

A **log-linear model** is a mathematical model that takes the form of a function whose logarithm equals a linear combination of the parameters of the model, which makes it possible to apply (possibly multivariate) linear regression. That is, it has the general form

$$\exp\left(c + \sum_i w_i f_i(X)\right),$$

in which the $f_i(X)$ are quantities that are functions of the variable X, in general a vector of values, while c and the w_i stand for the model parameters.





MODELO MÍNIMOS QUADRADOS PONDERADO

Weighted least squares (**WLS**), also known as **weighted linear regression**,[1][2] is a generalization of ordinary least squares and linear regression in which knowledge of the unequal variance of observations (*heteroscedasticity*) is incorporated into the regression. WLS is also a specialization of generalized least squares, when all the off-diagonal entries of the covariance matrix of the errors, are null.

Formulation [edit]

The fit of a model to a data point is measured by its residual, r_i, defined as the difference between a measured value of the dependent variable, y_i and the value predicted by the model, $f(x_i, \boldsymbol{\beta})$:

$$r_i(\boldsymbol{\beta}) = y_i - f(x_i, \boldsymbol{\beta}).$$

If the errors are uncorrelated and have equal variance, then the function

$$S(\boldsymbol{\beta}) = \sum_i r_i(\boldsymbol{\beta})^2,$$

is minimised at $\hat{\boldsymbol{\beta}}$, such that $\frac{\partial S}{\partial \beta_j}(\hat{\boldsymbol{\beta}}) = 0$.

The Gauss–Markov theorem shows that, when this is so, $\hat{\boldsymbol{\beta}}$ is a best linear unbiased estimator (BLUE). If, however, the measurements are uncorrelated but have different uncertainties, a modified approach might be adopted. Aitken showed that when a weighted sum of squared residuals is minimized, $\hat{\boldsymbol{\beta}}$ is the BLUE if each weight is equal to the reciprocal of the variance of the measurement

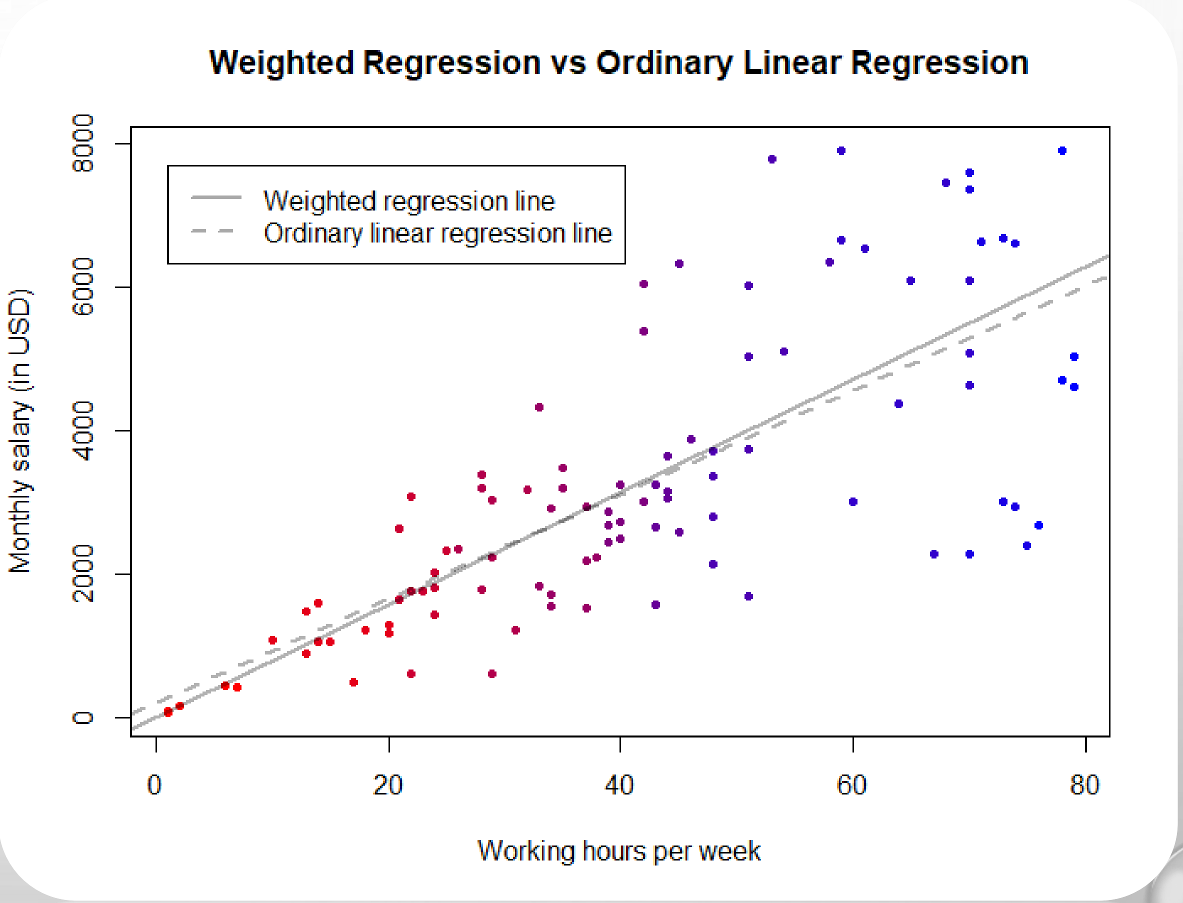
$$S = \sum_{i=1}^{n} W_{ii} {r_i}^2, \qquad W_{ii} = \frac{1}{{\sigma_i}^2}$$

The gradient equations for this sum of squares are

$$-2 \sum_i W_{ii} \frac{\partial f(x_i, \boldsymbol{\beta})}{\partial \beta_j} r_i = 0, \quad j = 1, \ldots, m$$

which, in a linear least squares system give the modified normal equations,

$$\sum_{i=1}^{n} \sum_{k=1}^{m} X_{ij} W_{ii} X_{ik} \hat{\beta}_k = \sum_{i=1}^{n} X_{ij} W_{ii} y_i, \quad j = 1, \ldots, m.$$



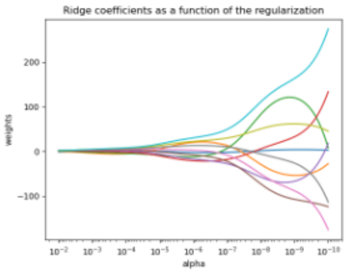
MODELO RIDGE

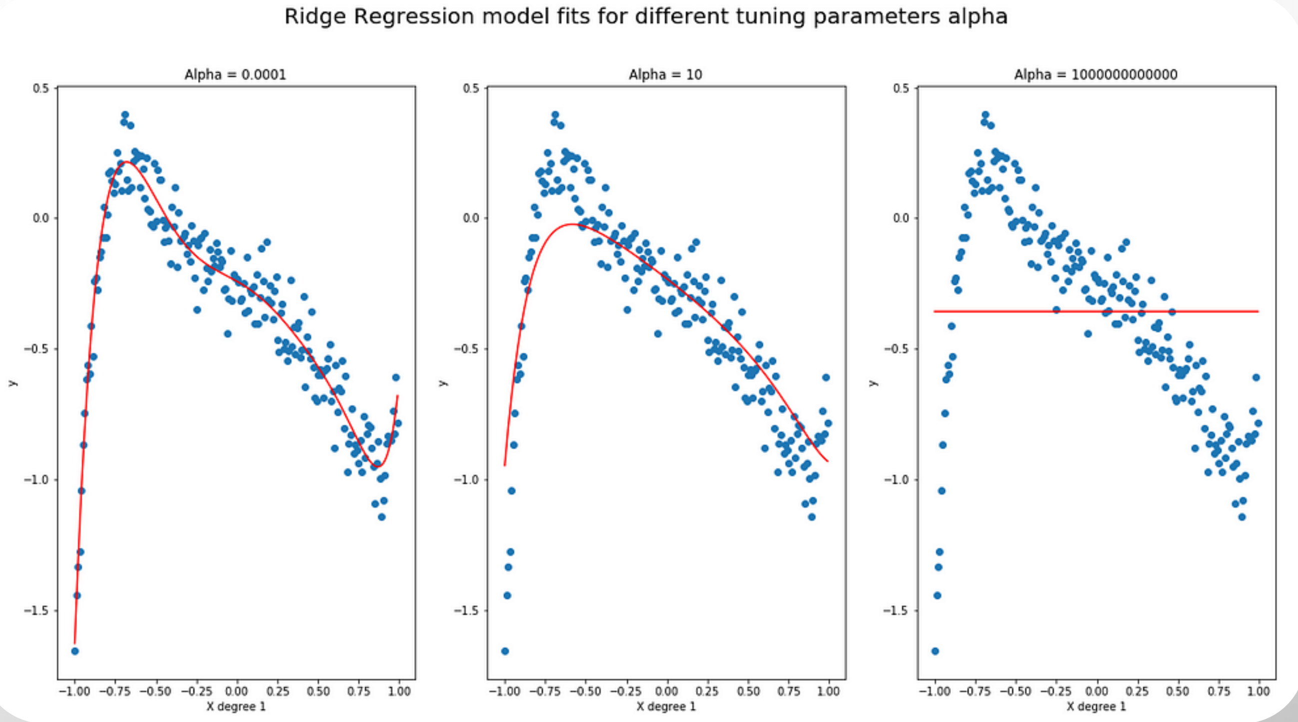
1.1.2.1. Regression

`Ridge` regression addresses some of the problems of Ordinary Least Squares by imposing a penalty on the size of the coefficients. The ridge coefficients minimize a penalized residual sum of squares:

$$\min_{w} ||Xw - y||_2^2 + \alpha ||w||_2^2$$

The complexity parameter $\alpha \geq 0$ controls the amount of shrinkage: the larger the value of α, the greater the amount of shrinkage and thus the coefficients become more robust to collinearity.





MODELO LASSO

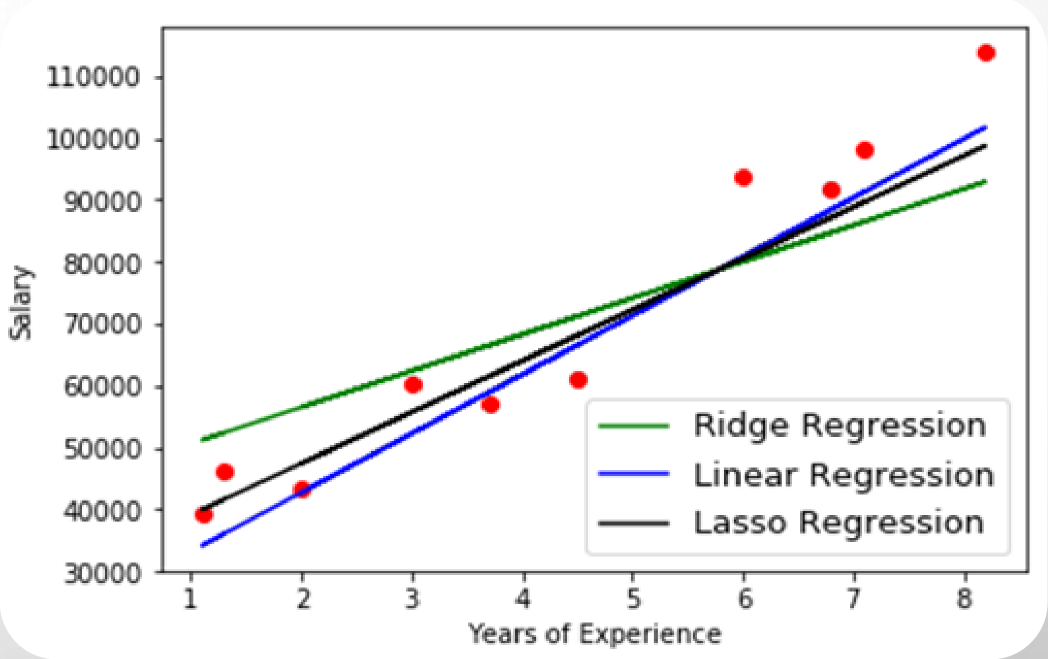
1.1.3. Lasso

The `Lasso` is a linear model that estimates sparse coefficients. It is useful in some contexts due to its tendency to prefer solutions with fewer non-zero coefficients, effectively reducing the number of features upon which the given solution is dependent. For this reason, Lasso and its variants are fundamental to the field of compressed sensing. Under certain conditions, it can recover the exact set of non-zero coefficients (see Compressive sensing: tomography reconstruction with L1 prior (Lasso)).

Mathematically, it consists of a linear model with an added regularization term. The objective function to minimize is:

$$\min_{w} \frac{1}{2n_{\text{samples}}} ||Xw - y||_2^2 + \alpha ||w||_1$$

The lasso estimate thus solves the minimization of the least-squares penalty with $\alpha ||w||_1$ added, where α is a constant and $||w||_1$ is the ℓ_1-norm of the coefficient vector.



MODELO ELASTIC-NET

1.1.5. Elastic-Net

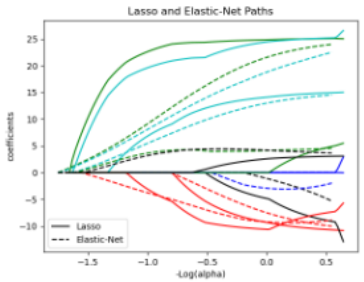
`ElasticNet` is a linear regression model trained with both ℓ_1 and ℓ_2-norm regularization of the coefficients. This combination allows for learning a sparse model where few of the weights are non-zero like `Lasso`, while still maintaining the regularization properties of `Ridge`. We control the convex combination of ℓ_1 and ℓ_2 using the `l1_ratio` parameter.

Elastic-net is useful when there are multiple features that are correlated with one another. Lasso is likely to pick one of these at random, while elastic-net is likely to pick both.

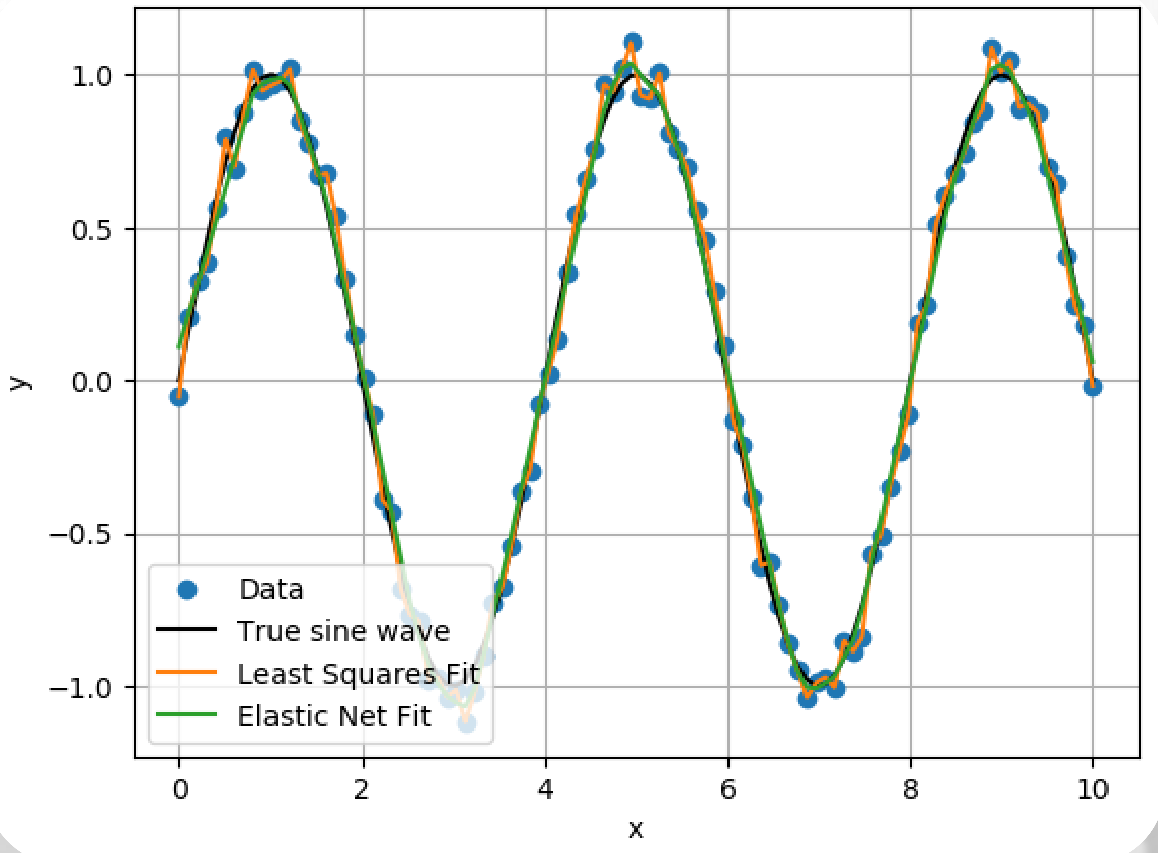
A practical advantage of trading-off between Lasso and Ridge is that it allows Elastic-Net to inherit some of Ridge's stability under rotation.

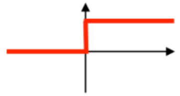
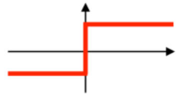
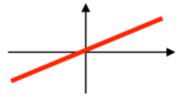
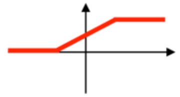
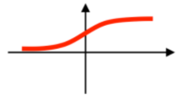
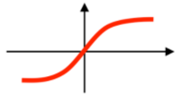
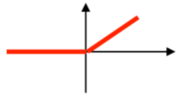
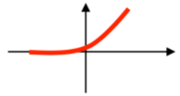
The objective function to minimize is in this case

$$\min_{w} \frac{1}{2n_{\text{samples}}} ||Xw - y||_2^2 + \alpha\rho||w||_1 + \frac{\alpha(1-\rho)}{2}||w||_2^2$$



The class `ElasticNetCV` can be used to set the parameters `alpha` (α) and `l1_ratio` (ρ) by cross-validation.



Activation function	Equation	Example	1D Graph
Unit step (Heaviside)	$\phi(z) = \begin{cases} 0, & z < 0, \\ 0.5, & z = 0, \\ 1, & z > 0, \end{cases}$	Perceptron variant	
Sign (Signum)	$\phi(z) = \begin{cases} -1, & z < 0, \\ 0, & z = 0, \\ 1, & z > 0, \end{cases}$	Perceptron variant	
Linear	$\phi(z) = z$	Adaline, linear regression	
Piece-wise linear	$\phi(z) = \begin{cases} 1, & z \geq \frac{1}{2}, \\ z + \frac{1}{2}, & -\frac{1}{2} < z < \frac{1}{2}, \\ 0, & z \leq -\frac{1}{2}, \end{cases}$	Support vector machine	
Logistic (sigmoid)	$\phi(z) = \frac{1}{1 + e^{-z}}$	Logistic regression, Multi-layer NN	
Hyperbolic tangent	$\phi(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$	Multi-layer Neural Networks	
Rectifier, ReLU (Rectified Linear Unit)	$\phi(z) = max(0, z)$	Multi-layer Neural Networks	
Rectifier, softplus	$\phi(z) = \ln(1 + e^z)$	Multi-layer Neural Networks	

SAÍDA LINEAR

VALIDATION

REDE NEURAL REGRESSÃO

ERRO MÉDIO QUADRÁTICO

MINIMIZAÇÃO DO MSE NO CONJUNTO DE TREINO, CONTROLADO PELO CONJUNTO DE VALIDAÇÃO.

ESTRATÉGIA DE BUSCA IDÊNTICA A DE CLASSIFICAÇÃO.

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (f_i - y_i)^2$$

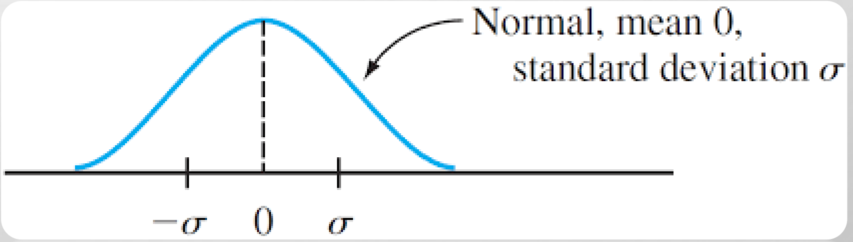
where N is the number of data points, f_i the value returned by the model and y_i the actual value for data point i.

FIGURAS DE MÉRITO - REGRESSÃO

- R QUADRADO

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

- RESÍDUO NORMAL DE MÉDIA ZERO E VARIÂNCIA CONSTANTE



VALIDAÇÃO : STATSMODELS

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 Ticker   R-squared:                       0.782
Model:                            OLS   Adj. R-squared:                  0.776
Method:                 Least Squares   F-statistic:                     128.4
Date:                Sun, 11 Sep 2022   Prob (F-statistic):           3.25e-79
Time:                        15:28:31   Log-Likelihood:                -844.20
No. Observations:                 259   AIC:                             1704.
Df Residuals:                     251   BIC:                             1733.
Df Model:                           7
Covariance Type:            nonrobust
============================================================================================
                               coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
const                      -45.0907     27.593     -1.634      0.103     -99.434       9.253
UK FTSE                      0.0227      0.002      9.162      0.000       0.018       0.028
ASX200 (yesterday)           0.0358      0.003     12.808      0.000       0.030       0.041
Japan Nikkei (y'day)        -0.0028      0.001     -4.576      0.000      -0.004      -0.002
AUD TWI 4pm                 -2.1353      0.413     -5.175      0.000      -2.948      -1.323
Iron Ore futures ($US/t)    -0.3426      0.034    -10.080      0.000      -0.409      -0.276
Uranium, weekly ($US/lb)     0.5427      0.092      5.879      0.000       0.361       0.725
Copper ($US/t)               0.0041      0.001      6.149      0.000       0.003       0.005
==============================================================================
Omnibus:                        0.789   Durbin-Watson:                   0.618
Prob(Omnibus):                  0.674   Jarque-Bera (JB):                0.853
Skew:                           0.001   Prob(JB):                        0.653
Kurtosis:                       2.719   Cond. No.                     2.16e+06
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.16e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Coeficiente de Determinação R^2

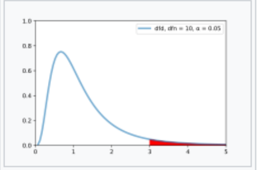
P Valor da Estatística F

***F*-test** 21 languages

Article Talk Tools

From Wikipedia, the free encyclopedia

An ***F*-test** is any statistical test in which the test statistic has an *F*-distribution under the null hypothesis. It is most often used when comparing statistical models that have been fitted to a data set, in order to identify the model that best fits the population from which the data were sampled. Exact "*F*-tests" mainly arise when the models have been fitted to the data using least squares. The name was coined by George W. Snedecor, in honour of Ronald Fisher. Fisher initially developed the statistic as the variance ratio in the 1920s.[1]

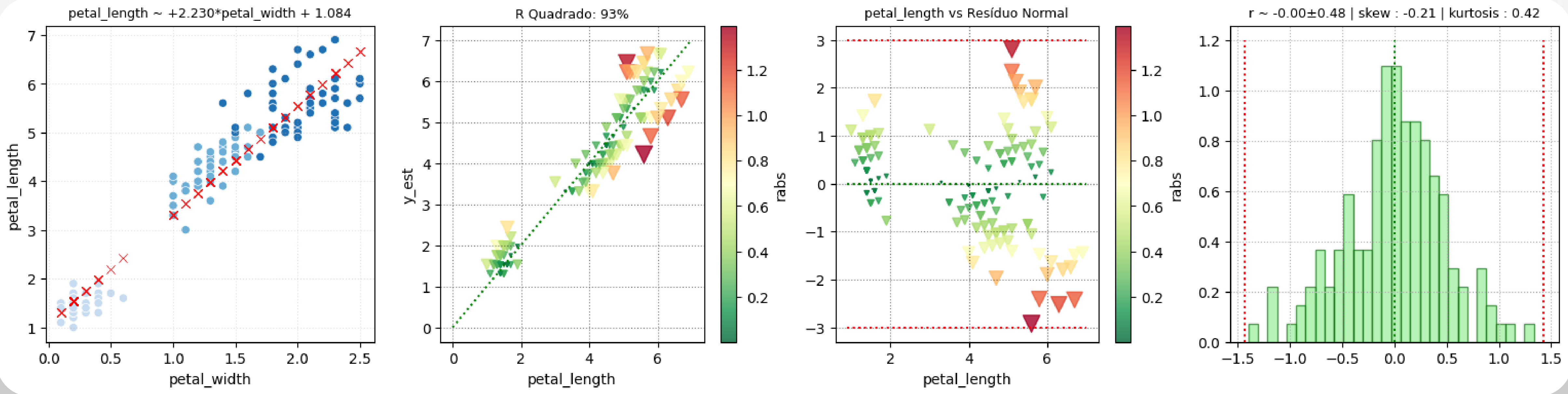


An f-test pdf with d1 and d2 = 10, at a significance level of 0.05.

P Valor dos Coeficientes

Número de Condicionamento

VALIDAÇÃO : GRÁFICOS DE APOIO



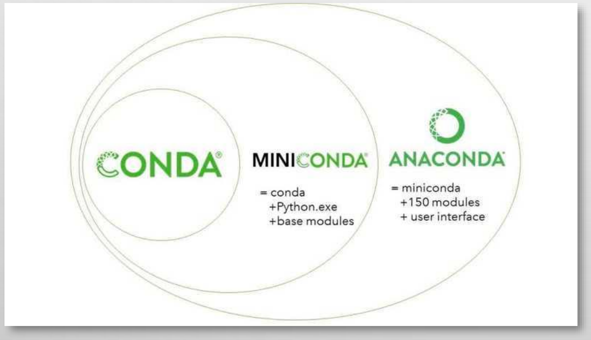
PARTE 2 : PRÁTICA

AMBIENTE PYTHON

5. Visualização

6. Machine
Learning

4. Variáveis
Aleatórias



1. Editor de Código

2. Gestor de Ambiente

3. Ambiente
Python do Projeto

3. Notebook
Dinâmico

PROBLEMA DE NEGÓCIO

Características das flores

Largura & comprimento da pétala

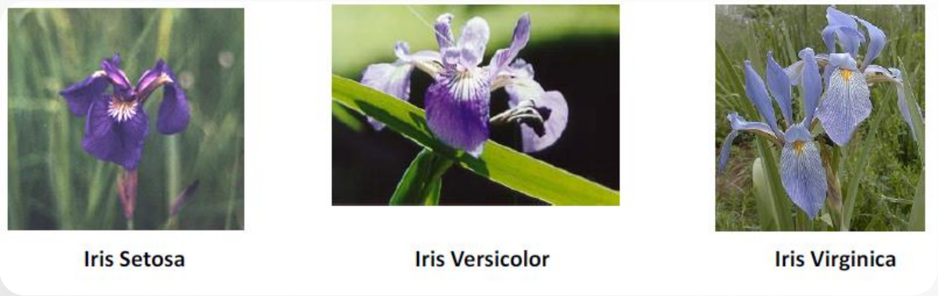
Largura & comprimento da sépala

Iris Setosa

Iris Versicolor

Iris Virginica

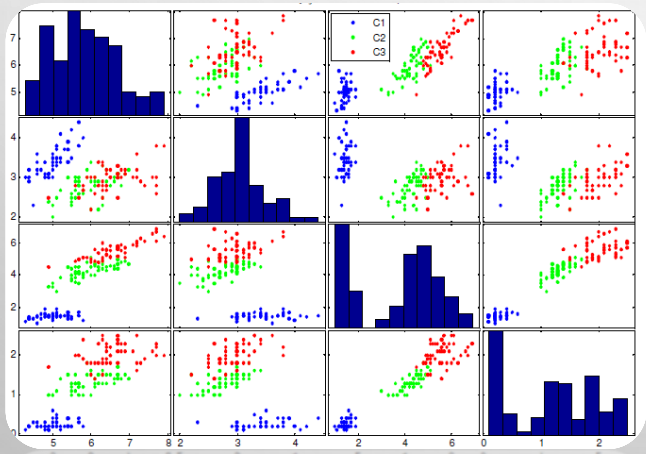
REPRESENTAÇÃO



Características das flores

Largura & comprimento da pétala

Largura & comprimento da sépala



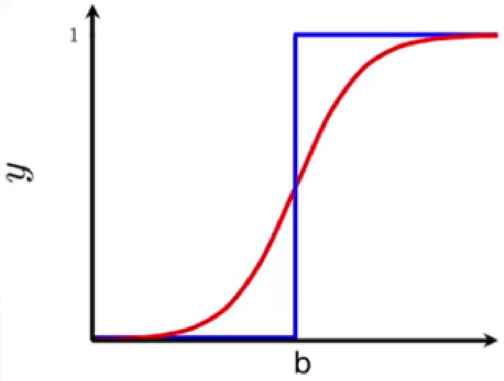
http://archive.ics.uci.edu/ml/datasets/Iris

Espaço de atributos com **4 dimensões!**

MODELAGEM

- **REDE NEURAL FEED FORWARD**
  - REPRESENTAÇÃO: 4 ATRIBUTOS
  - HIPERPARÂMETROS: GRIDSEARCH NO # DE NEURÔNIOS DA CAMADA OCULTA.
  - TREINAMENTO: BASE DE TREINO COMPLETA.
    - MSE
    - VALIDAÇÃO CRUZADA 10 FOLDS

Iris Setosa | Iris Versicolor | Iris Virginica



$$y = \frac{1}{1+e^{-(w^T x+b)}}$$

- $\sum_{i=1}^{n} w_i x_i$

REGRESSÃO IRIS

PRÓXIMA AULA: REGRESSÃO IRIS