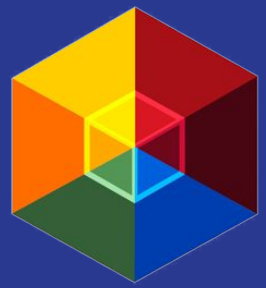
Linear Regression

Tim Analisis
Oct 20 2016

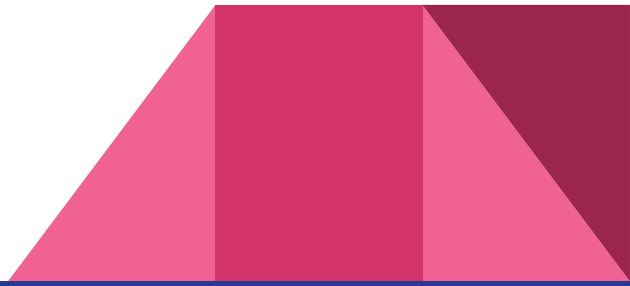
KMKLabs - Senayan City



1. Regression VS Classification

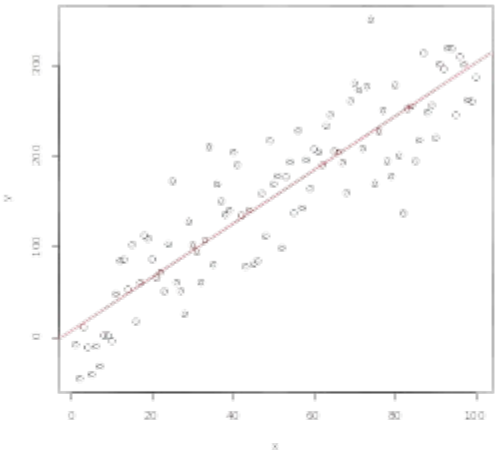
- **Regression** is used to predict continuous values.

- **Classification** is used to predict which class a data point is part of (discrete value)

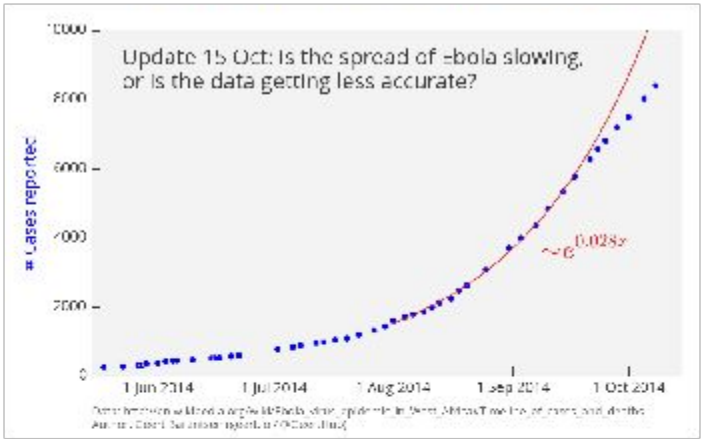


2. Regression

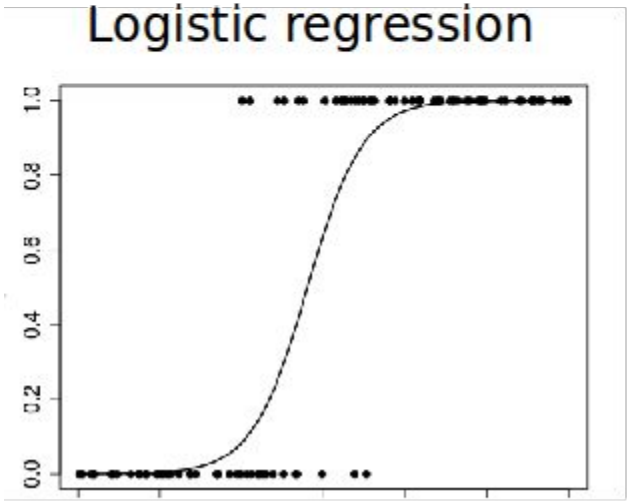
Linear regression



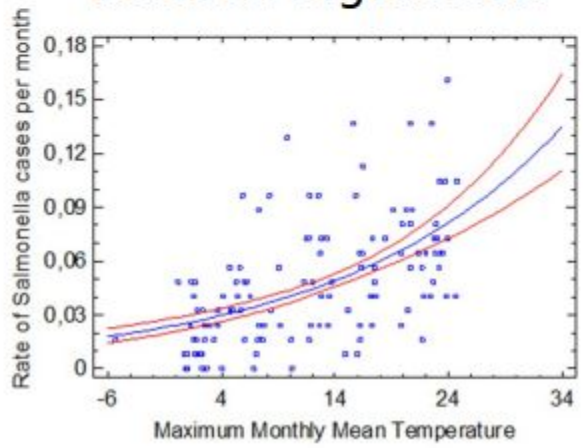
Non-linear regression



Logistic regression

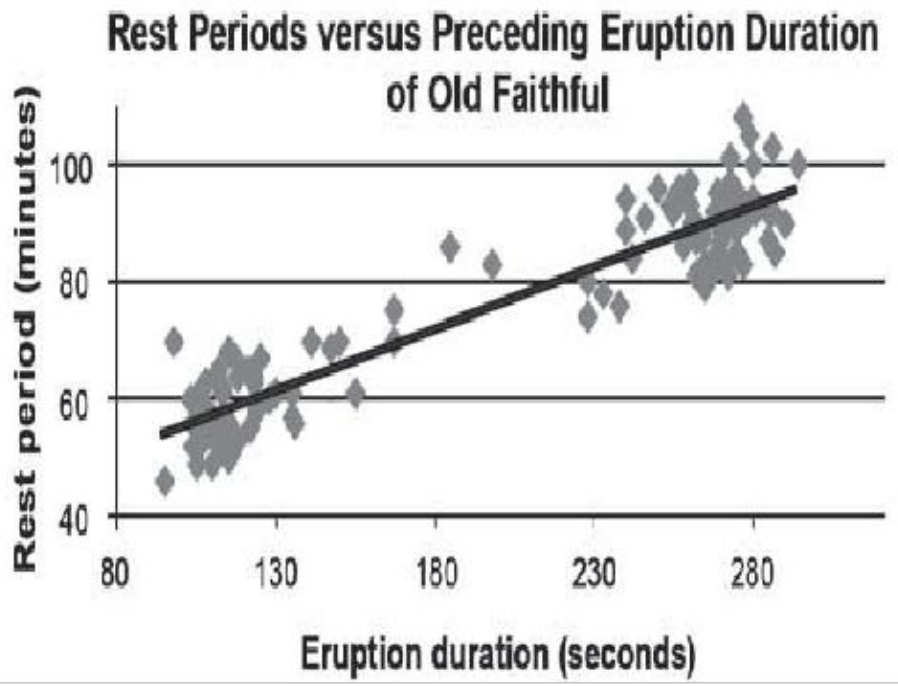


Poisson regression

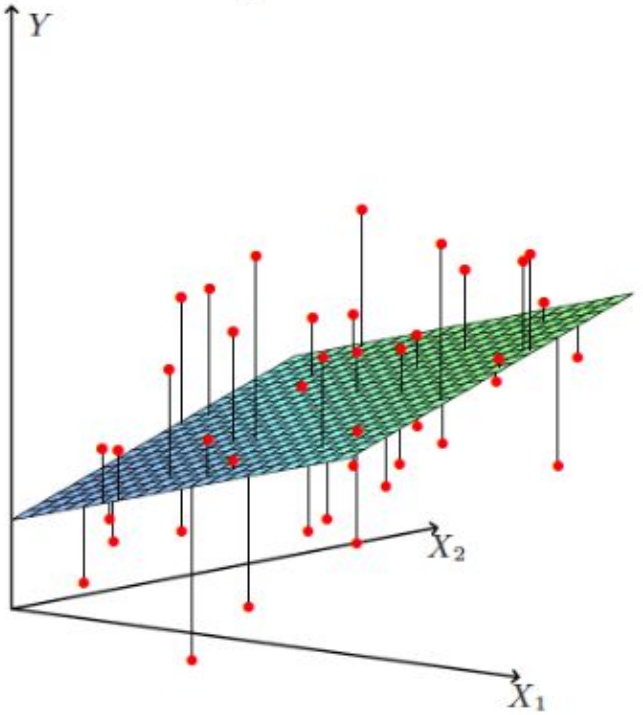


3. Simple VS Multivariate

Linear regression

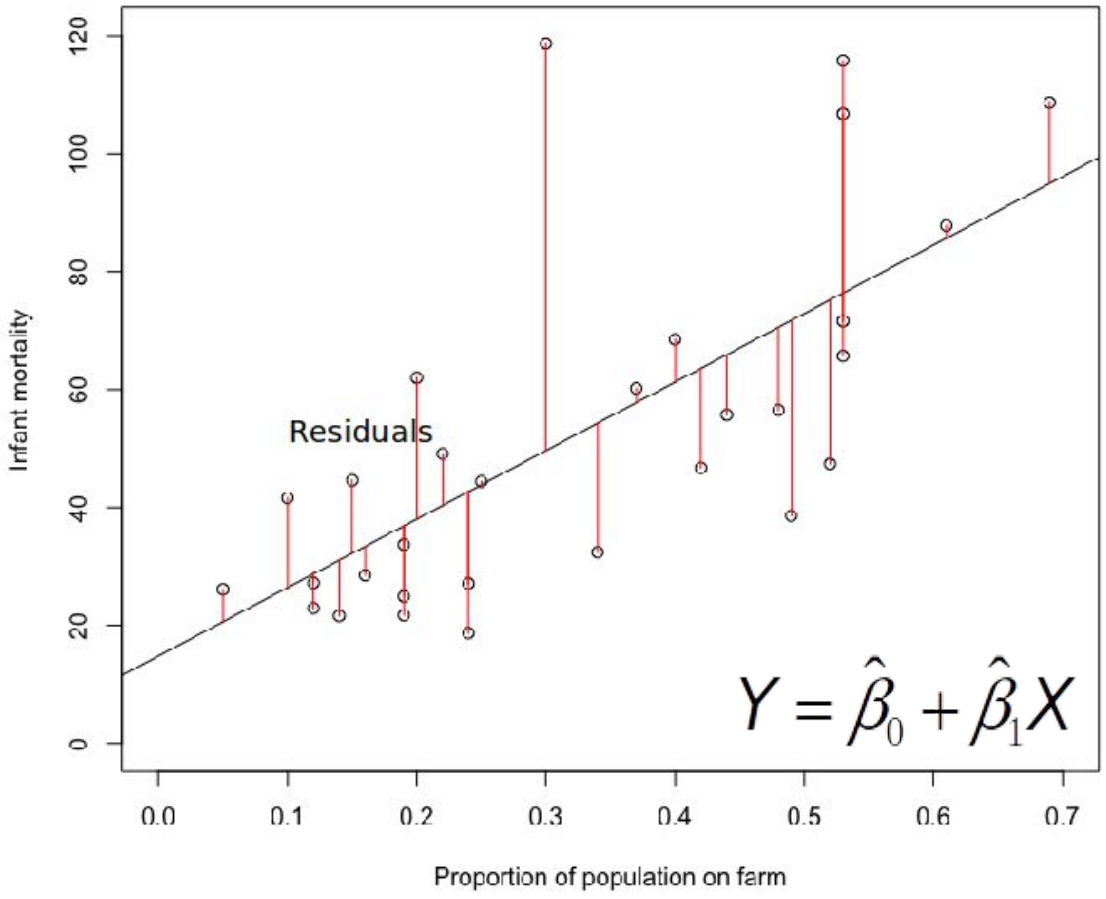


Multivariate regression



Target or dependent variable
Explanatory or independent variable

4. Model Representation



The best line is the one that minimizes the residuals on aggregate

$$\sum_{i=1}^{n} \left(Y_i - \left(\beta_0 + \beta_1 X_i \right) \right)^2$$

5. How to Determine β_0 and β_1 ?

Simple Matrix Multiplication

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

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{x}_1^{\mathrm{T}} \\ \mathbf{x}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{x}_n^{\mathrm{T}} \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix},$$

$$\beta = X^{-1}.y$$

5. How to Determine β_0 and β_1 ?

Stochastic Gradien Descent

$$J(\beta_0, \beta_1) = \sum_{(x_i, y_i) \in X \times Y} (y_i - \hat{y}(x_i))^2 = \sum_{(x_i, y_i) \in X \times Y} (y_i - (\beta_0 + \beta_1 x_i))^2$$

Gradient Descent

There is an incredibly simple way to minimize a multivariable function iteratively: gradient descent. As you may remember from your calculus class, the gradient of a function $g(x, y)$ is

$$\nabla g(x, y) = \begin{bmatrix} \frac{\partial g}{\partial x} \\ \frac{\partial g}{\partial y} \end{bmatrix}.$$

$$\beta_t = \beta_{t-1} - \alpha e$$

6. Linear Regression On Sklearn

sklearn.linear_model.LinearRegression

class `sklearn.linear_model.` **`LinearRegression`** (*fit_intercept=True*, *normalize=False*, *copy_X=True*, *n_jobs=1*) [source]

Attributes:

coef_ : array, shape (n_features,) or (n_targets, n_features)

Estimated coefficients for the linear regression problem. If multiple targets are passed during the fit (y 2D), this is a 2D array of shape (n_targets, n_features), while if only one target is passed, this is a 1D array of length n_features.

residues_ : array, shape (n_targets,) or (1,) or empty

Sum of residuals. Squared Euclidean 2-norm for each target passed during the fit. If the linear regression problem is under-determined (the number of linearly independent rows of the training matrix is less than its number of linearly independent columns), this is an empty array. If the target vector passed during the fit is 1-dimensional, this is a (1,) shape array.

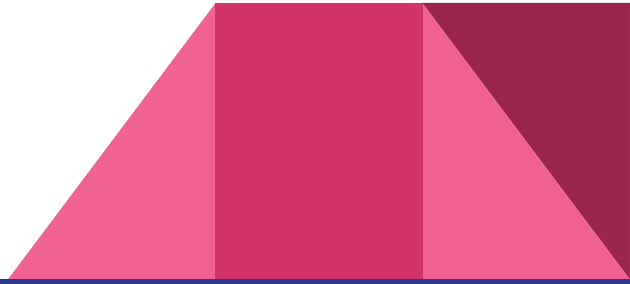
New in version 0.18.

intercept_ : array

Independent term in the linear model.

6. Linear Regression On Sklearn

```
>>> data = [[1,2,3],[3,2,1],[2,3,1]]

>>> label = [50, 65, 70]

>>> from sklearn.linear_model import LinearRegression

>>> lr = LinearRegression()

>>> lr.fit(data, label)
<<< LinearRegression(copy_X=True, fit_intercept=True, n_jobs=1,
normalize=False)

>>> lr.coef_
<<< array([ 0.83333333,  5.83333333, -6.66666667])

>>> lr.intercept_
<<< 57.499999999999993
```



7. Evaluating Linear Rgeression

R^2 (Coefficient of Determination)

The **coefficient of determination** (denoted by R^2) is a key output of regression analysis. It is interpreted as the proportion of the variance in the dependent variable that is predictable from the independent variable.

- The coefficient of determination is the square of the correlation (r) between predicted y scores and actual y scores; thus, it ranges from 0 to 1.
- With linear regression, the coefficient of determination is also equal to the square of the correlation between x and y scores.
- An R^2 of 0 means that the dependent variable cannot be predicted from the independent variable.
- An R^2 of 1 means the dependent variable can be predicted without error from the independent variable.
- An R^2 between 0 and 1 indicates the extent to which the dependent variable is predictable. An R^2 of 0.10 means that 10 percent of the variance in *Y* is predictable from *X*; an R^2 of 0.20 means that 20 percent is predictable; and so on.

The formula for computing the coefficient of determination for a linear regression model with one independent variable is given below.

Coefficient of determination. The coefficient of determination (R^2) for a linear regression model with one independent variable is:

$$R^2 = \{ (1 / N) * \Sigma [(x_i - \bar{x}) * (y_i - \bar{y})] / (\sigma_x * \sigma_y) \}^2$$

where N is the number of observations used to fit the model, Σ is the summation symbol, x_i is the x value for observation i, $\bar{x}$ is the mean x value, y_i is the y value for observation i, $\bar{y}$ is the mean y value, σ_x is the standard deviation of x, and σ_y is the standard deviation of y.

7. Evaluating Linear Rgeression

R^2 (Coefficient of Determination)

The **coefficient of determination** (denoted by R^2) is a key output of regression analysis. It is interpreted as the proportion of the variance in the dependent variable that is predictable from the independent variable.

- The coefficient of determination is the square of the correlation (r) between predicted y scores and actual y scores; thus, it ranges from 0 to 1.
- With linear regression, the coefficient of determination is also equal to the square of the correlation between x and y scores.
- An R^2 of 0 means that the dependent variable cannot be predicted from the independent variable.
- An R^2 of 1 means the dependent variable can be predicted without error from the independent variable.
- An R^2 between 0 and 1 indicates the extent to which the dependent variable is predictable. An R^2 of 0.10 means that 10 percent of the variance in *Y* is predictable from *X*; an R^2 of 0.20 means that 20 percent is predictable; and so on.

The formula for computing the coefficient of determination for a linear regression model with one independent variable is given below.

Coefficient of determination. The coefficient of determination (R^2) for a linear regression model with one independent variable is:

$$R^2 = \{ (1 / N) * \Sigma [(x_i - \bar{x}) * (y_i - \bar{y})] / (\sigma_x * \sigma_y) \}^2$$

where N is the number of observations used to fit the model, Σ is the summation symbol, x_i is the x value for observation i, $\bar{x}$ is the mean x value, y_i is the y value for observation i, $\bar{y}$ is the mean y value, σ_x is the standard deviation of x, and σ_y is the standard deviation of y.

7. Evaluating Linear Rgeression

R^2 (Coefficient of Determination)

```
>>> data_test = [[1,2,1],[1,1,3],[2,2,1]]

>>> label_test = [45, 60, 68]

>>> lr.predict(data_test)
<<< array([ 63.33,  44.16  64.17])

>>> lr.score(data_test, label_test)
<<< -1.2059902200489008
```

score (*X*, *y*, *sample_weight=None*) [source]

Returns the coefficient of determination R^2 of the prediction.

The coefficient R^2 is defined as (1 - u/v), where u is the regression sum of squares ((y_true - y_pred) ** 2).sum() and v is the residual sum of squares ((y_true - y_true.mean()) ** 2).sum(). Best possible score is 1.0 and it can be negative (because the model can be arbitrarily worse). A constant model that always predicts the expected value of y, disregarding the input features, would get a R^2 score of 0.0.

Parameters:	**X** : array-like, shape = (n_samples, n_features) Test samples. **y** : array-like, shape = (n_samples) or (n_samples, n_outputs) True values for X. **sample_weight** : array-like, shape = [n_samples], optional Sample weights.
Returns:	**score** : float R^2 of self.predict(X) wrt. y.

8. However …

Assumptions

As with most statistical procedures, linear regression makes certain assumptions about the data used in an analysis; if these assumptions are violated, the results of the analysis might not be valid. Key assumptions for simple linear regression include:

Data appropriateness

The outcome variable should be continuous, measured at the interval or ratio level, and be unbounded (or at least cover a wide range); the predictor variables should be continuous or dichotomous. Categorical predictors with more than two categories can be recoded into a series of dichotomous dummy variables; this is covered in Chapter 10.

Independence

Each value of the outcome variable is independent of each other value. This would be violated if there were some pattern of time dependency, for instance, or if some of the dependent variables were measured from subjects clustered into larger units (such as members of the same family or children studying in the same classroom) in some way that affected their value on the dependent variable. This assumption is checked by your knowledge of the data and how it was collected.

Linearity

The relationship between the predictor and outcome variable resembles a straight line. This assumption is checked by graphing the data; if it resembles a shape other than a straight line, you might need to transform one or both variables or choose another procedure.

Distribution

The continuous variables are approximately normally distributed and do not have extreme outliers. The distribution of continuous variables may be checked by creating a histogram (eyeballing the data) and by a statistical test for normality such as the Kolmogorov-Smirnov. An outlier is defined as a data value that is far from the other values for the same variable in a data set; sometimes it is described as a data value that doesn't seem to belong with the others. Outlier detection is partly a matter of judgment, is further discussed in Chapter 17, and can be a multistep process. (An unusual data value can be the result of an error in data entry, for instance, or it might be an apparently valid value.)

Homoscedasticity

The errors of prediction are constant over the entire data range. This means that the errors are not, for instance, smaller when the y value is small and larger when the y value is large. This assumption is checked by graphing the standardized residuals against the standardized predicted values; the data should resemble a cloud without any indication that the errors of prediction are not constant over the whole range of the data. Figure 8-3 shows homoscedastic data and Figure 8-4 heteroscedastic data.

Independence and normality of the errors

The error of prediction for each data point should be independent of the error of prediction for each other data point, and the errors should be normally distributed. The independence assumption is checked by the Durbin-Watson test (discussed later), and the normality assumption is checked by graphing the residuals (error terms).

9. Homework

Given a set of data as shown below :

No	Luas Tanah	Luas Bangunan	Jumlah Kamar	Jumlah Lantai	Harga (jt)
1	800	500	2	2	1000
2	760	400	2	1	800
3	900	600	2	1	950
4	1000	500	3	2	1200
5	1200	1100	4	1	1500
6	975	500	3	2	750
7	650	550	2	1	800
8	900	700	2	1	1300
9	560	480	3	2	450
10	720	630	2	1	975

Questions :

1. Use K-Means in sklearn to divide **ALL** data into 3 cluster (use the random_state = 14)
   a. Which data belongs to cluster 1?
   b. Which data belongs to cluster 2?
   c. Which data belongs to cluster 3?
2. Use Linear Regression to create the linear regression model by using **ONLY** the GREEN data.
   a. What is the coefition of each predictor?
   b. What is the intercept of the function?
   c. Use the RED data as the test data. What isthe predicted price of the house using the model you created?
   d. Evaluate the prediction you made. What is coefficient of determination score?

THANKYOU :*