Mathematics for Data Analysis
– Linear Regression –

Matteo Gorgone

University of Messina, Department MIFT

email: mgorgone@unime.it

What is Linear Regression?

Linear regression is a linear approach for modelling the relationship between a response and one or more explanatory variables (also known as dependent and independent variables). The case of one explanatory variable is called simple linear regression; for more than one explanatory variable, the process is called multiple linear regression. The relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data.

Models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters!

Be careful

Multiple linear regression is different from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single dependent variable.

Linear regression: applications

Applications fall into one of the following two broad categories:

- if the aim is prediction, forecasting, or error reduction, linear regression can be used to fit a predictive model to an observed data set of values of the response and explanatory variables. After developing such a model, if additional values of the explanatory variables are collected without an associated response value, the fitted model can be used to make a prediction of the response;
- if the aim is to explain variation in the response variable that can be attributed to variation in the explanatory variables, linear regression analysis can be applied to quantify the strength of the relationship between the response and the explanatory variables, and in particular to determine whether some explanatory variables may have no linear relationship with the response at all, or to identify which subsets of explanatory variables may contain redundant information about the response.

The problem

The following table shows, for 31 cherry trees, the trunk diameter and the volume of wood obtained from the felling of trees.

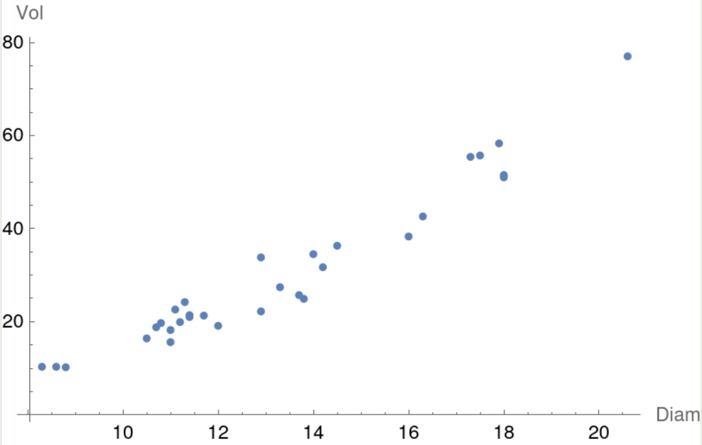
Diameter	Volume	Diameter	Volume	Diameter	Volume
8.3	10.3	11.3	24.2	14.0	34.5
8.6	10.3	11.4	21.0	14.2	31.7
8.8	10.2	11.4	21.4	14.5	36.3
10.5	16.4	11.7	21.3	16.0	38.3
10.7	18.8	12.0	19.1	16.3	42.6
10.8	19.7	12.9	22.2	17.3	55.4
11.0	15.6	12.9	33.8	17.5	55.7
11.0	18.2	13.3	27.4	17.9	58.3
11.1	22.6	13.7	25.7	18.0	51.5
11.2	19.9	13.8	24.9	18.0	51.0
20.6	77.0				

We want to use the data to get one equation that allows to predict the volume (obtainable only after the tree has been cut down) from the diameter (easily measurable)!

Data Analysis

```
data = {{"Diameter", "Volume"}, {8.3, 10.3}, {8.6, 10.3}, {8.8, 10.2}, {10.5, 16.4},
 {10.7, 18.8}, {10.8, 19.7}, {11.0, 15.6}, {11.0, 18.2}, {11.1, 22.6}, {11.2, 19.9},
 {20.6, 77.0}, {11.3, 24.2}, {11.4, 21.0}, {11.4, 21.4}, {11.7, 21.3}, {12.0, 19.1},
 {12.9, 22.2}, {12.9, 33.8}, {13.3, 27.4}, {13.7, 25.7}, {13.8, 24.9}, {14.0, 34.5},
 {14.2,31.7}, {14.5, 36.3}, {16.0, 38.3}, {16.3, 42.6}, {17.3, 55.4}, {17.5, 55.7},
 {17.9, 58.3}, {18.0, 51.5}, {18.0, 51.0}};
```

Diameter vs. Volume: Plot of data



A strong relationship is evident. Basically linear with perhaps some problems at the extremes.

Model for cherry trees

Suppose there is a linear relationship between the data. Then, we can consider a model of the type

$$\text{volume} = \alpha(\text{diameter}) + \beta + \text{error},$$

where error expresses the part of volume fluctuations not related to diameter.
It seems reasonable to try to compute α and β in order to obtain good "predictions" on the observed data set. To this purpose, let us denote by $N = 31$ the number of observations, y_i the i-th tree wood volume and x_i the i-th tree trunk diameter.
We would like to find values for parameters such that

$$\begin{aligned} y_1 &\approx \alpha x_1 + \beta, \\ y_2 &\approx \alpha x_2 + \beta, \\ &\vdots \\ y_N &\approx \alpha x_N + \beta. \end{aligned}$$

Linear regression

Suppose we have two N-uples of numerical values measuring some characters of a population, say

$$\mathbf{x} = \{x_1, x_2, \ldots, x_N\}, \qquad \mathbf{y} = \{y_1, y_2, \ldots, y_N\}.$$

When there are arguments supporting the assumption that these two characters are in some way related, we may conjecture that there is a model describing their mutual variance, *i.e.*, we want to predict the dependent variable y using only the independent variable x.

The simplest model is the linear one, that is we may assume that there is a linear equation,

$$y = \alpha x + \beta, \qquad \alpha,\ \beta \text{ suitable constants to be determined,}$$

that well describe the data.

Measuring errors

The purpose of this straight line is not just to be close to all of the data (for this we will have to wait for PCA and dimensionality reduction), but just prediction!

Linear regression

Of course, in real applications, it is unrealistic to suppose that the straight line in the *xy* plane, with equation

$$y = \alpha x + \beta$$

is such that all points (x_i, y_i) $(i = 1, \dots, N)$ lie on the straight line.

The aim is looking for coefficients α and β such that the errors, that is the quantities (linear deviations of data from the straight line or residuals),

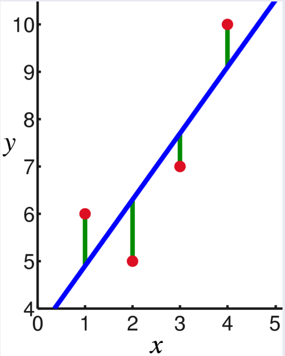
$$r_i = y_i - \alpha x_i - \beta, \qquad i = 1, \dots, N$$

i.e., the differences among the observed values and the "predicted" values, are the smallest possible.
If the deviations for each point are small, then also their sum will be small. Nevertheless, to avoid that positive deviations balance negative deviations, we consider the sum of the squared deviations, and look for α and β such that this sum attains a minimum.

Remark

Note that the errors are not the orthogonal distance (or Euclidean distance) from the points (x_i, y_i) to the straight line, but the (signed) distance from the points (x_i, y_i) to the corresponding point on the straight line with the same x_i-coordinate.

Errors



The observations (red) are assumed to be the result of random deviations (green) from an underlying relationship (blue) between the dependent variable y and the independent variable x.

Method of squared minima (or Least Square Method)

The method of least squares is a standard approach in regression analysis to approximate the solution of overdetermined systems by minimizing the sum of the squares of the residuals made in the results of each individual equation.

Least squares problems fall into two categories:

- linear or ordinary least squares;
- nonlinear least squares.

This distinction depends on whether or not the residuals are linear in all parameters to be estimated.

Remark

The linear least squares problem has a closed–form solution that is unique, provided that the number of data points used for fitting equals or exceeds the number of unknown parameters.

The nonlinear problem is usually solved by iterative procedures; at each iteration the system is approximated by a linear one, and thus the core computation is similar in both cases.

We will investigate only the linear case!

Least Square Method

The goal is to find the parameter values of a model function to best fit a data set.
Let us consider a data set made of N data points (x_i, y_i), $i = 1, \ldots, N$, where x_i is an independent variable and y_i is a dependent variable whose value is found by observation. The model function has the form

$$\mathbf{y} = f(\mathbf{x}, \boldsymbol{\alpha}) + \mathbf{r},$$

where f is a function depending on p parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)$, with $p \leq N$, $\mathbf{x} = (x_1, \ldots, x_N)$ and $\mathbf{r} = (r_1, \ldots, r_N)$ is the error.
The fit of a model to a data point is measured by its residual, defined as the difference between the observed value of the dependent variable and the value predicted by the model:

$$r_i = y_i - f(x_i, \boldsymbol{\alpha}), \qquad i = 1, \ldots, N.$$

The Least Square Method finds the optimal parameter values by minimizing the Sum of Squared Residuals (or, Sum of Squared Errors):

$$E(\boldsymbol{\alpha}) = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} \left(y_i - f(x_i, \boldsymbol{\alpha})\right)^2.$$

. . . Least Square Method

The minimum of the sum of squares is found by setting the gradient to zero, *i.e.*, $\nabla_{\boldsymbol{\alpha}} E(\boldsymbol{\alpha}) = 0$. Since the model contains p parameters, there are p gradient equations:

$$\frac{\partial E}{\partial \alpha_j} = 2 \sum_{i=1}^{N} r_i \frac{\partial r_i}{\partial \alpha_j} = 0, \qquad j = 1, \dots, p,$$

and, since $r_i = y_i - f(x_i, \boldsymbol{\alpha})$, the gradient equations become

$$\frac{\partial E}{\partial \alpha_j} = -2 \sum_{i=1}^{N} (y_i - f(x_i, \boldsymbol{\alpha})) \frac{\partial f(x_i, \boldsymbol{\alpha})}{\partial \alpha_j} = 0, \qquad j = 1, \dots, p.$$

The best fit can be found by solving the gradient (called also normal) equations. Each particular problem requires particular expressions for the model and its partial derivatives.

Least Square Method in simple linear regression

Using the data (x_i, y_i) $(i = 1, \dots, N)$, we construct the function, depending on α and β,

$$E(\alpha, \beta) = \sum_{i=1}^{N} (\alpha x_i + \beta - y_i)^2.$$

This function may attain a minimum for values of α and β such that the partial derivatives of $E(\alpha, \beta)$ are vanishing:

$$\begin{aligned} \frac{\partial E}{\partial \alpha} &= \sum_{i=1}^{N} 2(\alpha x_i + \beta - y_i)x_i = 0, \\ \frac{\partial E}{\partial \beta} &= \sum_{i=1}^{N} 2(\alpha x_i + \beta - y_i) = 0. \end{aligned}$$

Method of squared minima

By simple computations, the previous conditions may be written as follows:

$$\begin{aligned} \alpha\overline{x^2} + \beta\overline{x} - \overline{xy} &= 0, \\ \alpha\overline{x} + \beta - \overline{y} &= 0, \end{aligned}$$

where $\overline{x} = \mathrm{mean}(\{x_1, \dots, x_N\})$ and $\overline{y} = \mathrm{mean}(\{y_1, \dots, y_N\})$, $\overline{x^2} = \mathrm{mean}(\{x_1^2, \dots, x_N^2\})$, and $\overline{xy} = \mathrm{mean}(\{x_1y_1, \dots, x_Ny_N\})$.
These equations represent a nonhomogeneous linear system for the unknowns α and β, whose solution is:

$$\begin{aligned} \alpha &= \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{\mathrm{cov}(\mathbf{x}, \mathbf{y})}{\sigma^2(\mathbf{x})}, \\ \beta &= \overline{y} - \alpha\overline{x} = \overline{y} - \frac{\mathrm{cov}(\mathbf{x}, \mathbf{y})}{\sigma^2(\mathbf{x})}\overline{x}, \end{aligned}$$

where $\mathrm{cov}(\mathbf{x}, \mathbf{y}) = \frac{1}{N-1}\sum_{k=1}^{N}(x_k - \overline{x})(y_k - \overline{y}) = \frac{N}{N-1}(\overline{xy} - \overline{x}\,\overline{y})$.

Remark

By inserting $\beta = \overline{y} - \alpha\overline{x}$ in the regression straight line $y = \alpha x + \beta$, we have

$$y = \alpha x + \beta = \alpha x + \overline{y} - \alpha\overline{x} \Longrightarrow y - \overline{y} = \alpha(x - \overline{x}),$$

i.e., the straight line goes through the point $(\overline{x}, \overline{y})$.

Mean of residuals

The sum of the residuals is zero, *i.e.*,

$$\sum_{i=1}^{N} r_i = \sum_{i=1}^{N} (y_i - \alpha x_i - \beta) = 0.$$

In fact:

$$\sum_{i=1}^{N} r_i = \sum_{i=1}^{N} (y_i - \alpha x_i - \beta) = \sum_{i=1}^{N} y_i - \alpha \sum_{i=1}^{N} x_i - \beta N = \overline{y}N - \alpha\overline{x}N - (\overline{y} - \alpha\overline{x})N = 0,$$

where the relation $\beta = \overline{y} - \alpha\overline{x}$ has been used.

Variance of residuals

Since the mean of residuals is zero, the variance of residuals is the mean of the squared residuals. Denoting with $\mathbf{r} = \{r_1, \ldots, r_N\}$, we have

$$\sigma^2(\mathbf{r}) = \frac{1}{N}\sum_{i=1}^{N} r_i^2 - \bar{r}^2 = \frac{1}{N}\sum_{i=1}^{N} r_i^2.$$

The variance of the residuals $\sigma^2(\mathbf{r})$ can be used to get a “numerical idea” of the goodness of fit of the model to the data. In fact, the more the variance of the residuals will be small, the better regression line “explains” the response variations.

Variance of residuals

The variance of residuals can be computed as

$$\sigma^2(\mathbf{r}) = \sigma^2(\mathbf{y})\left(1 - \frac{\mathrm{cov}^2(\mathbf{x},\mathbf{y})}{\sigma^2(\mathbf{x})\sigma^2(\mathbf{y})}\right).$$

In fact, using the relations $\alpha = \frac{\mathrm{cov}(\mathbf{x},\mathbf{y})}{\sigma^2(\mathbf{x})}$ and $\beta = \overline{y} - \alpha\overline{x}$, we have:

$$\begin{aligned}
\sigma^2(\mathbf{r}) &= \frac{1}{N}\sum_{i=1}^{N} r_i^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \alpha x_i - \beta)^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \alpha x_i - \overline{y} + \alpha\overline{x})^2 = \frac{1}{N}\sum_{i=1}^{N}\left((y_i - \overline{y}) - \alpha(x_i - \overline{x})\right)^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{y})^2 + \frac{\alpha^2}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2 - \frac{2\alpha}{N}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y}) = \\
&= \sigma^2(\mathbf{y}) + \alpha^2\sigma^2(\mathbf{x}) - 2\alpha\mathrm{cov}(\mathbf{x},\mathbf{y}) = \\
&= \sigma^2(\mathbf{y}) + \frac{\mathrm{cov}^2(\mathbf{x},\mathbf{y})}{\sigma^4(\mathbf{x})}\sigma^2(\mathbf{x}) - 2\frac{\mathrm{cov}(\mathbf{x},\mathbf{y})}{\sigma^2(\mathbf{x})}\mathrm{cov}(\mathbf{x},\mathbf{y}) = \\
&= \sigma^2(\mathbf{y}) + \frac{\mathrm{cov}^2(\mathbf{x},\mathbf{y})}{\sigma^2(\mathbf{x})} - 2\frac{\mathrm{cov}^2(\mathbf{x},\mathbf{y})}{\sigma^2(\mathbf{x})} = \sigma^2(\mathbf{y}) - \frac{\mathrm{cov}^2(\mathbf{x},\mathbf{y})}{\sigma^2(\mathbf{x})} = \sigma^2(\mathbf{y})\left(1 - \frac{\mathrm{cov}^2(\mathbf{x},\mathbf{y})}{\sigma^2(\mathbf{x})\sigma^2(\mathbf{y})}\right).
\end{aligned}$$

About linear regression...

Since

$$\sigma^2(\mathbf{r}) = \sigma^2(\mathbf{y}) \left(1 - \frac{\mathrm{cov}^2(\mathbf{x}, \mathbf{y})}{\sigma^2(\mathbf{x})\sigma^2(\mathbf{y})}\right) = \sigma^2(\mathbf{y})(1 - \rho^2),$$

it follows that the variance of the residuals will be small and then better is the fitting of data with linear model, if the absolute value of the correlation coefficient

$$\rho = \frac{\mathrm{cov}(\mathbf{x}, \mathbf{y})}{\sqrt{\sigma^2(\mathbf{x})\sigma^2(\mathbf{y})}}$$

is close as much possible to 1.

When $\rho \pm 1$, the data are aligned along a straight line, in particular:

- $\rho > 0$ if the data **x** and **y** grow together;
- $\rho < 0$ if **y** decreases when **x** grows.

For values of ρ distant from 1, linear regression loses its meaning, even if this does not exclude that between the data may still exist a non-linear relation (when $\rho = 0$ there is no any linear correlation among the data).

Computation of parameters in the case of cherry trees

Let x_i be the i-th trunk diameter and y_i be the i-th wood volume, with $i = 1, \dots, 31$ ($N = 31$). It is:

$$\overline{x} = \frac{1}{31}\sum_{i=1}^{31} x_i = 13.25, \qquad \overline{y} = \frac{1}{31}\sum_{i=1}^{31} y_i = 30.17,$$

$$\sigma^2(\mathbf{x}) = \frac{1}{30}\sum_{i=1}^{31}(x_i - \overline{x})^2 = 9.85, \qquad \sigma^2(\mathbf{y}) = \frac{1}{30}\sum_{i=1}^{31}(y_i - \overline{y})^2 = 270.2,$$

$$\operatorname{cov}(\mathbf{x}, \mathbf{y}) = \frac{1}{30}\sum_{i=1}^{31}(x_i - \overline{x})(y_i - \overline{y}) = 49.89, \qquad \sigma^2(\mathbf{r}) = \sigma^2(\mathbf{y})\left(1 - \frac{\operatorname{cov}^2(\mathbf{x}, \mathbf{y})}{\sigma^2(\mathbf{x})\sigma^2(\mathbf{y})}\right) = 17.48.$$

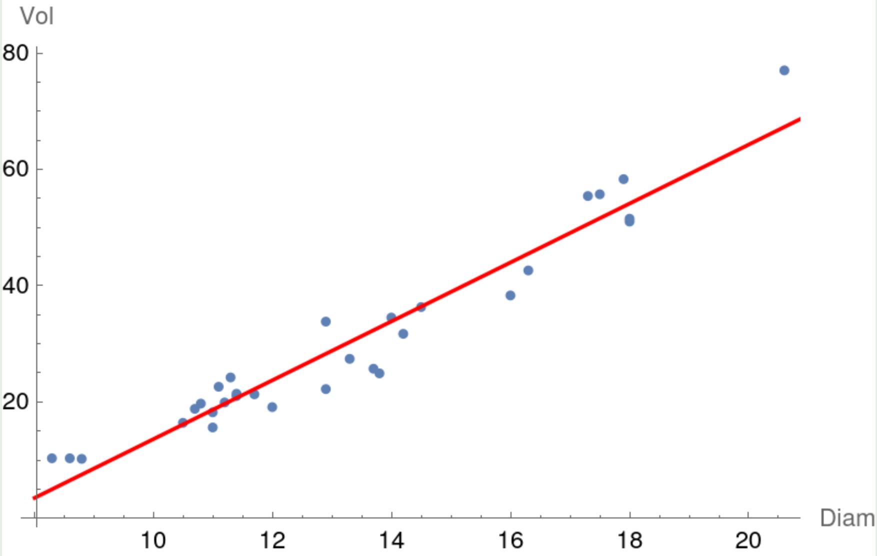
Then:

$$\alpha = \frac{\operatorname{cov}(\mathbf{x}, \mathbf{y})}{\sigma^2(\mathbf{x})} = \frac{49.88}{9.85} = 5.07, \qquad \beta = \overline{y} - \frac{\operatorname{cov}(\mathbf{x}, \mathbf{y})}{\sigma^2(\mathbf{x})}\overline{x} = 30.17 - \frac{49.88}{9.85}13.25 = -36.94.$$

Furthermore, the correlation coefficient is positive and very close to 1! In fact:

$$\rho = \frac{\operatorname{cov}(\mathbf{x}, \mathbf{y})}{\sqrt{\sigma^2(\mathbf{x})\sigma^2(\mathbf{y})}} = \frac{49.88}{\sqrt{9.85 \cdot 270.20}} = 0.97$$

Volume vs. Diameter: Regression straight line



$$Vol = 5.07\,\mathrm{diam} - 36.94, \qquad \rho = 0.97, \qquad \sigma^2(\mathbf{r}) = 17.48$$

Italian Football League, 2014-2015

Team	Matches	Victories	Draws	Defeats	GS	GC	GD	Score
Juventus	38	26	9	3	72	24	48	87
Roma	38	19	13	6	54	31	23	70
Lazio	38	21	6	11	71	38	33	69
Fiorentina	38	18	10	10	61	46	15	64
Napoli	38	18	9	11	70	54	16	63
Genoa	38	16	11	11	62	47	15	59
Sampdoria	38	13	17	8	48	42	6	56
Inter	38	14	13	11	59	48	11	55
Torino	38	14	12	12	48	45	3	54
Milan	38	13	13	12	56	50	6	52
Palermo	38	12	13	13	53	55	-2	49
Sassuolo	38	12	13	13	49	57	-8	49
Verona	38	11	13	14	49	65	-16	46
Chievo	38	10	13	15	28	41	-13	43
Empoli	38	8	18	12	46	52	-6	42
Udinese	38	10	11	17	43	56	-13	41
Atalanta	38	7	16	15	38	57	-19	37
Cagliari	38	8	10	20	48	68	-20	34
Cesena	38	4	12	22	36	73	-37	24
Parma	38	6	8	24	33	75	-42	19

Data Analysis

```
data =
  {{"Team", "Goals scored", "Goals conc.", "Goal diff.", "Score"},
   {"Juventus", 72, 24, 48, 87}, {"Roma", 54, 31, 23, 70},
   {"Lazio", 71, 38, 33, 69}, {"Fiorentina", 61, 46, 15, 64},
   {"Napoli", 70, 54, 16, 63}, {"Genoa", 62, 47, 15, 59},
   {"Sampdoria", 48, 42, 6, 56}, {"Inter", 59, 48, 11, 55},
   {"Torino", 48, 45, 3, 54}, {"Milan", 56, 50, 6, 52},
   {"Palermo", 53, 55, -2, 49}, {"Sassuolo", 49, 57, -8, 49},
   {"Verona", 49, 65, -16, 46}, {"Chievo", 28, 41, -13, 43},
   {"Empoli", 46, 52, -6, 42}, {"Udinese", 43, 56, -13, 41},
   {"Atalanta", 38, 57, -19, 37}, {"Cagliari", 48, 68, -20, 34},
   {"Cesena", 36, 73, -37, 24}, {"Parma", 33, 75, -42, 19}};
```

Correlation Matrix for Italian Football League

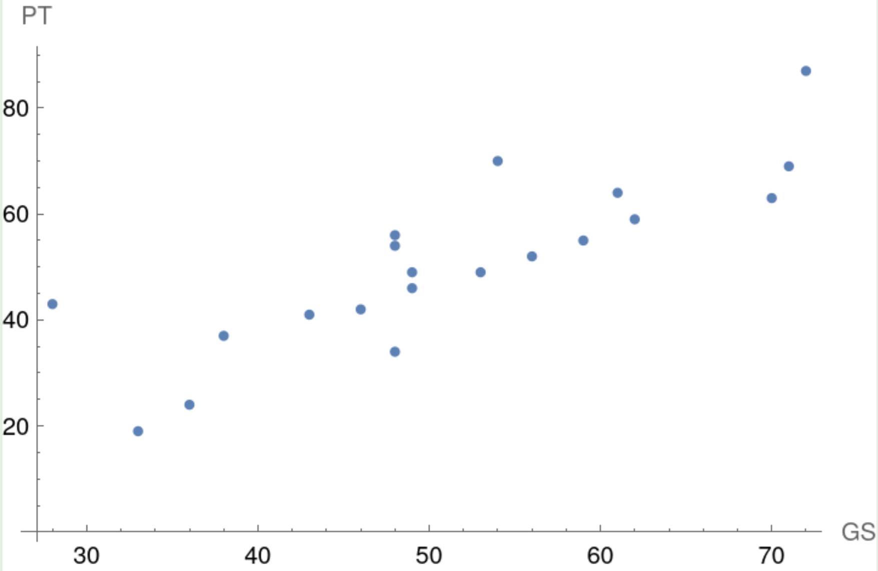
Taking the scores, the number of scored goals, the number of conceded goals, and the differences between the last two, we have the following correlation matrix:

$$C = \begin{bmatrix} 1. & 0.838748 & -0.883734 & 0.982628 \\ 0.838748 & 1. & -0.538252 & 0.869634 \\ -0.883734 & -0.538252 & 1. & -0.884162 \\ 0.982628 & 0.869634 & -0.884162 & 1. \end{bmatrix}.$$

Comment

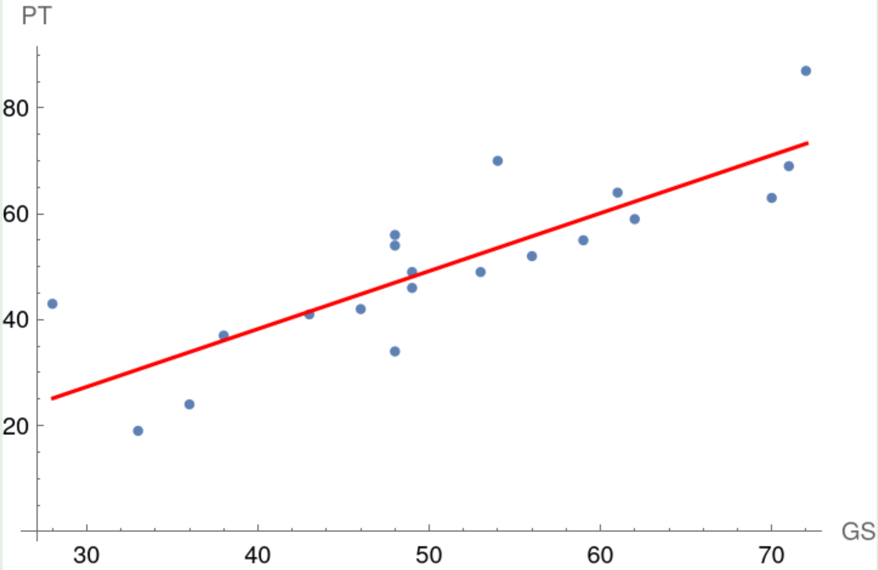
Correlation matrix is a symmetric matrix: on the main diagonal we have entries equal to 1 (each variable is trivially correlated to itself); the score is positively correlated to the scored goals and to goal differences, and negatively correlated to the conceded goals. Moreover, the best value of correlation (a number very close to 1) is between the score and the goal difference!

Gol scored vs. Score: Plot of data



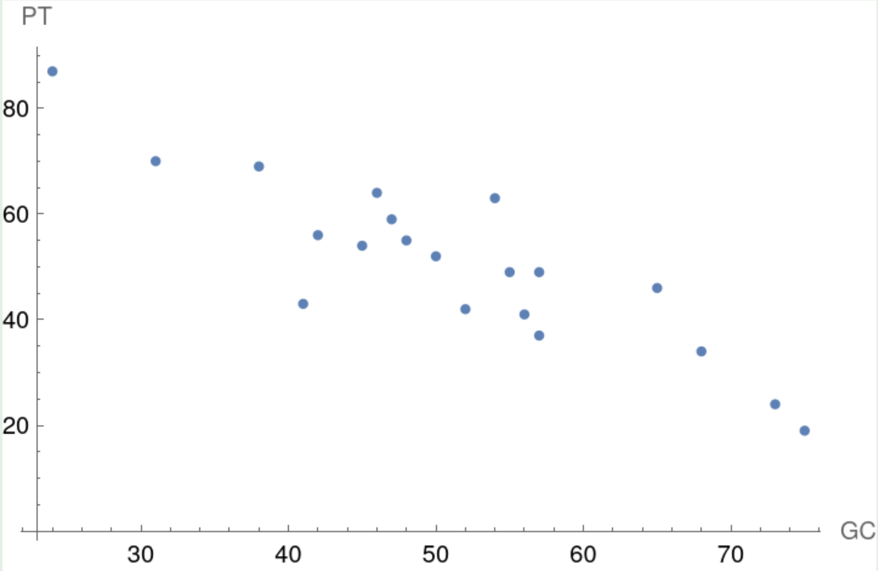
Positive Correlation!

Goals scored vs. Score: Regression straight line



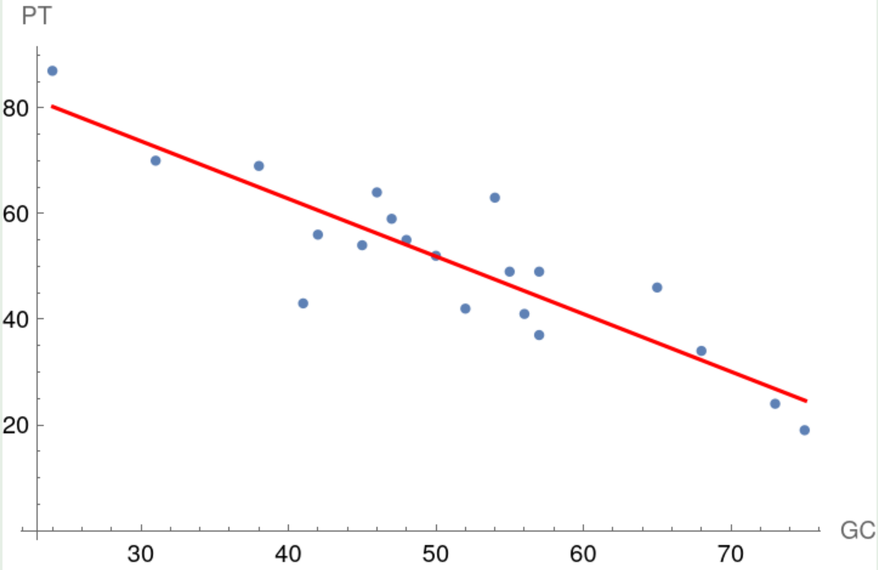
$$Score = 1.09367\,GS - 5.34602, \qquad \rho = 0.838748, \qquad \sigma^2(\mathbf{r}) = 76.8182$$

Goals conceded vs. Score: Plot of data



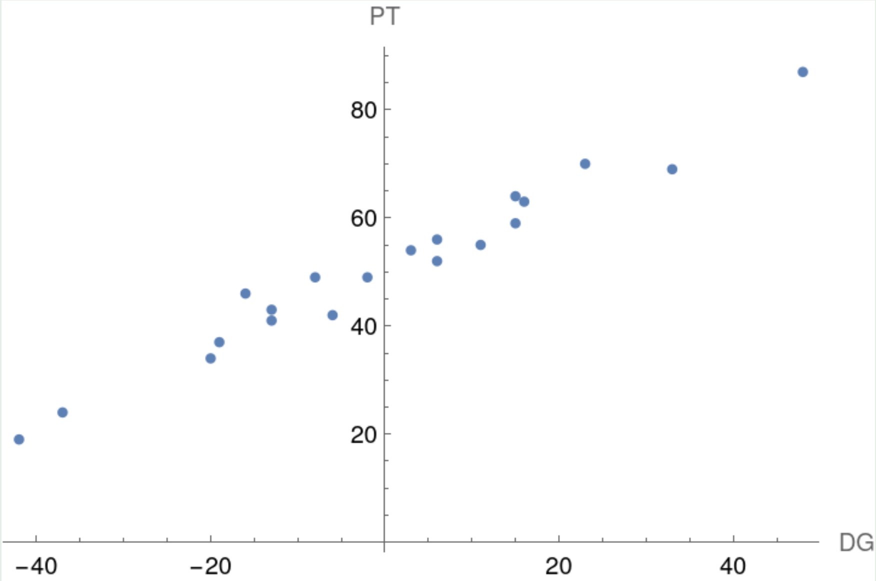
Negative correlation!

Goals conceded vs. Score: Regression straight line



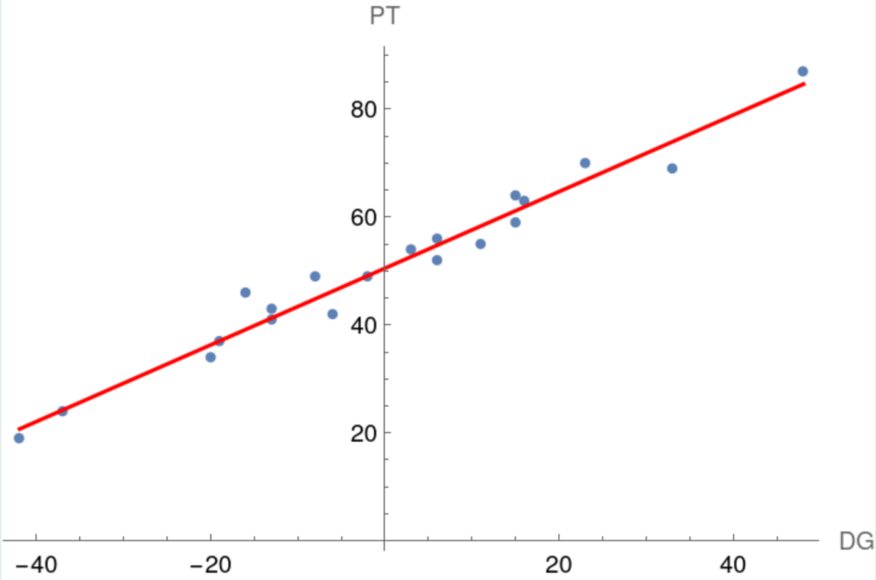
$$Score = -1.09044\,GC + 106.48, \qquad \rho = -0.883734, \qquad \sigma^2(\mathbf{r}) = 56.7424$$

Goal difference vs. Score: Plot of data



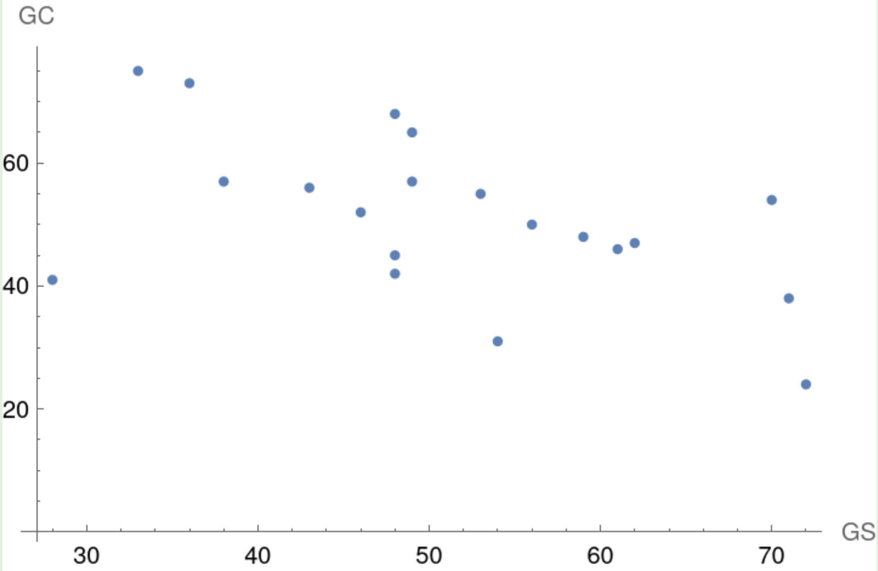
Positive correlation!

Goal difference vs. Score: Regression straight line



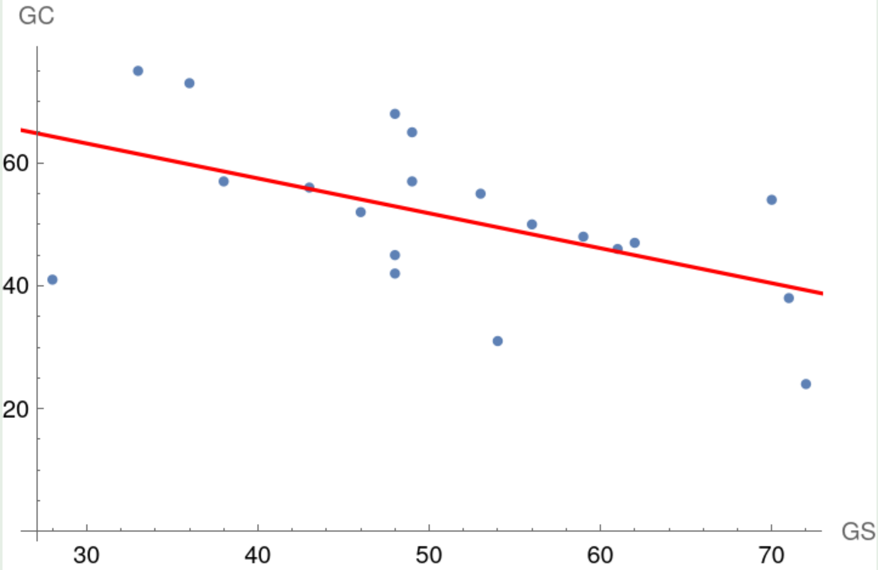
$$Score = 0.710253\,GD + 50.65, \qquad \rho = 0.982628, \qquad \sigma^2(\mathbf{r}) = 8.92314$$

Goals scored vs. Goals conceded: plot of data



Negative correlation!

Goals scored vs. Goals conceded: Regression straight line loses its meaning!



$$GS = -0.568804\,GC + 80.3227, \qquad \rho = -0.538252, \qquad \sigma^2(\mathbf{r}) = 108.233$$

Details of the computation in C

Using a C program, the data **x** and **y** can be represented as two arrays `x` and `y` of `n` elements (float). Then we need some functions to compute the mean value and the covariance (in fact, the variance is simply the covariance of a set of data with itself). And this is all you need to determine α and β.

```
float mean(float *a, int n)
// Mean of the array a
{
  int k;
  float ma=0.0;
  for (k=0;k<n;k++)
    ma=ma+a[k];
  return ma/n;
}
```

```
float cov(float *a,float *b,int n)
// Covariance of arrays a,b
{
  int k;
  float ma,mb,cov=0.0;
  ma = mean(a,n);
  mb = mean(b,n);
  for (k=0;k<n;k++)
    cov=cov+(a[k]-ma)*(b[k]-mb);
  return cov/(n-1);
}
```

```
alpha = cov(x,y,n)/cov(x,x,n);
beta = mean(y)-alpha*mean(x);
```