Probability and Statistics
for Final Year Engineering Students

By Yoni Nazarathy, Last Updated: May 22, 2011.

Lecture 6s:
Least Squares / Linear Regression

Introduction:

Suppose that we wish to find a relationship between two (perhaps physical) variables, X and Y. Estimating the correlation coefficient ($\rho(X,Y)$) gives us one way to see if these variables are statistically related. This method is good for cases where we think of both X and Y as being random and we want to summarize their "dependence" using a single number in the range [-1,1].

An alternative (and often more popular) method is to use **regression analysis**. This method assumes that the values of x are non-random (so we denote them with lower case x) and the values of Y are random and depend on x in the following way:

$$Y(x) = \beta_0 + \beta_1 x + \varepsilon \quad \text{(simple linear regression).}$$

We are thus assuming that there exist constants β_0 (intercept) and β_1 (slope) such that values of Y follow the linear line $\beta_0 + \beta_1 x$ yet are subject to some random errors (noise) denoted by ε (which is assumed to have a mean of zero).

We can now use sample data $((x_1, y_1), \ldots, (x_n, y_n))$ to find $\widehat{Y(x)}$ or alternatively find $\hat{\beta}_0$ and $\hat{\beta}_1$, so that,

$$\widehat{Y(x)} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{(simple linear regression).}$$

We will shortly show that a common and useful way to carry out the estimation is using the following formulas (called the **least squares solution**):

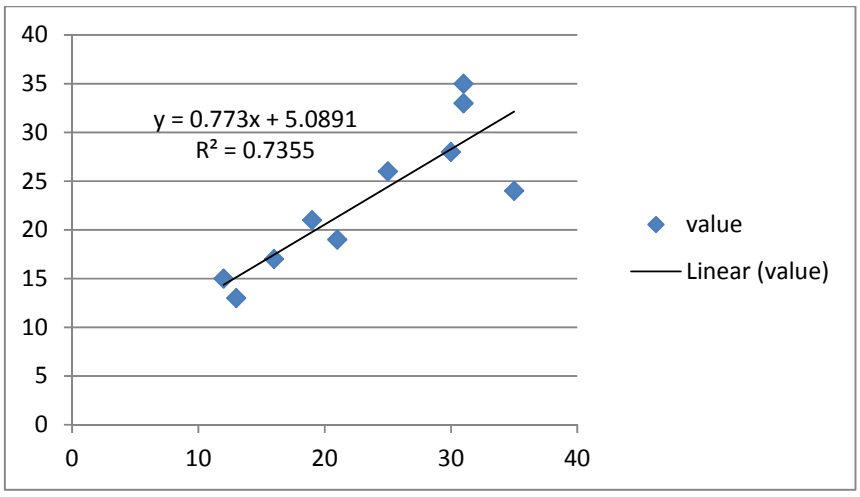
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Let us begin with an example: In the robotic arm suppose we are trying to implement a new algorithm which will "selectively give priority to picking up of items that give higher value to the warehouse". The idea is that eventually all items are picked up, but items which give "higher value" are picked up first. To do this, our arm needs a mechanism to identify the value in items. We are trying to do this solely based on the dimensions of the item, specifically based on the volume of items (which our computer vision system can estimate with very high precision).

We now collect data in the form $((x_1, y_1), \dots, (x_n, y_n))$ where x_i is the volume of item i and y_i is the value as indicated by the warehouse, given in dollars.

item	volume	Value
1	12	15
2	13	13
3	16	17
4	19	21
5	21	19
6	25	26
7	30	28
8	31	35
9	31	33
10	35	24



Least Squares Solution:

The estimators $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ and $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ are obtained by minimization of the cost: $Q(\hat{y}, y) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

This implies minimization of $Q = \sum_{i=1}^{n}\left(y_i - (\beta_0 + \beta_1 x_i)\right)^2$.

Simple calculus can be used:

$$\frac{\partial Q}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial Q}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i = 0$$

(It can be shown by use of second derivatives that the above equations indeed define a global minimum):

The conditions can be written as:

$$\beta_0 n + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i$$

We denote these as the **normal equations**.

Their solution are $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ and $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$.

The **coefficient of determination,**

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

Is a value between 0 and 1 that represents the proportion of the sum of squares of deviations of the y values about their mean that can be attributed to a linear relationship between y and x. (This value also equals the square of the correlation coefficient studied in section 4).

Multiple Least Squares Solutions:

We can assume that the dependent variable y depends on more than one (p) independent variables. In this case the model becomes.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + + \beta_p x_{ip} + \varepsilon_i$$

For example:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$$

or,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 {x_{i1}}^2.$$

In the first example we assume there is a surface parameterized by $(\beta_0, \beta_1, \beta_2)$ that describes the dependence of y on the two coordinates x_1 and x_2. In the second example we assume that there is a parabola which describes the relation between x and y.

In both of these cases the idea of least squares (fully derived above in the simple case) generalizes, the **normal equations** are still a set of linear equations but now with p+1 unknowns. These can be computed easily (you would usually use statistical/mathematical software to do so).

The statistical analysis which we describe now also generalizes to the multiple least squares case (but we do not discuss this further).

Statistical Assumptions and Results of Linear Regression:

Assumptions:

1. Predictor/independent variable is known (not random).
2. The mean of y given x is linear.
3. Common variance for the error. (Homoscedasticity as opposed to a model that is heteroscedastic).
4. Error follows a normal distribution with mean 0 and variance σ^2.
5. Independence of errors.

Results arising from these assumptions:

The distribution of β_0 and β_1.

$$\hat{\beta}_1 \sim Normal\left(\beta_1, \frac{\sigma^2}{s_x^2}\right)$$

$$\hat{\beta}_0 \sim Normal(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{s_x^2}\right))$$

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{x}\sigma^2}{s_x^2}$$

Where the independent variables, x's are known (with average $\bar{x}$) and,

$$s_x^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

So the least squares solution is an unbiased estimator. Since we now know the "distribution of the statistic" further results can be developed (confidence intervals and hypothesis tests).

Checking model assumptions:

- Residuals.
  - Constant variance
  - Normal Distribution.
  - Independence.
- Outliers.
  - Some times can be thrown away (from the least squares).
  - Robust regression.
- Model Selection in multiple-regression.