10.001: Correlation and Regression

R. Sureshkumar

January 14, 1997

In the last two classes, we learned to compute statistical measures from a sample of observations on a statistical variable. Moreover, we discussed how the sample statistics can be used to arrive at a model for the entire population. Today, we discuss how we can measure the *correlation*, i.e., the degree of linear association, between two statistical variables. Once again, let's try to learn the key concepts through an example.

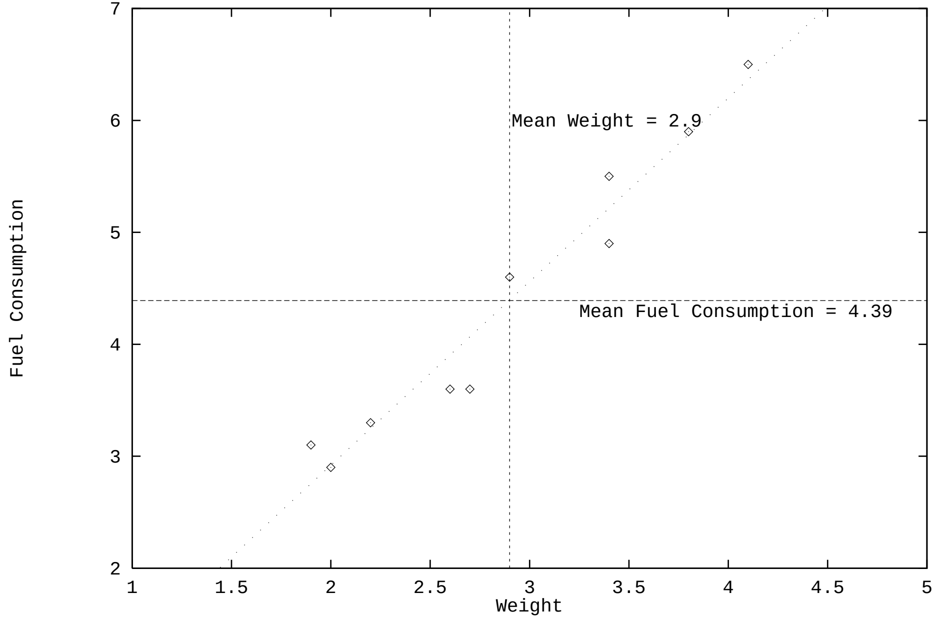
1 Example

The following data consist of observations for the weights of 10 different automobiles (in 1000 pounds) and the corresponding fuel consumptions (gallons per 100 miles).

Weight (x)	Fuel Consumption (y)
3.4	5.5
3.8	5.9
4.1	6.5
2.2	3.3
2.6	3.6
2.9	4.6
2.0	2.9
2.7	3.6
1.9	3.1
3.4	4.9

We would like to find out how y is correlated to x and whether we could represent that correlation in a functional form valid within the range of the data.

Figure 1: Fuel Consumption vs. Weight and the Best Fit Line



2 Correlation Analysis

The simplest way to find out qualitatively the correlation is to plot the data. In the case of our example, as seen from Figure 1, a strong *positive* correlation between y and x is evident, i.e., the plot reveals that as the weight increases, the fuel consumption increases as well. How can we quantify the degree of correlation? This is usually done by specifying the correlation coefficient R, defined as

$$R = \frac{1}{n-1} \sum_{i=1}^{n} \frac{x_i - \mu_x}{\sigma_x} \; \frac{y_i - \mu_y}{\sigma_y}, \tag{1}$$

where μ_x and σ_x denote the sample mean and the sample standard deviation respectively for the variable x and μ_y and σ_y denote the sample mean and the sample standard deviation respectively for the variable y.

Now, let's assume that a perfect linear relationship exists between the variables x and y. i.e., $y_i = ax_i + b$ for $i = 1, 2, \cdots, n$ with $a \neq 0$. Now verify using the definitions of the mean and the variance that $\mu_y = a\mu_x + b$ and $\sigma_y = |a|\sigma_x$. This implies from Eq. 1 that $R = a/|a|$. Or in other words, $R = 1$ if $a > 0$ and $R = -1$ if $a < 0$. The case $R = 1$ corresponds to the maximum possible linear positive association between x and y, meaning that all the data points will lie exactly on a straight line of positive slope.

Similarly, $R = -1$ corresponds to the maximum possible negative association between the statistical variables x and y. In general, $-1 \leq R \leq 1$ with the magnitude and the sign of R representing the *strength* and *direction* respectively of the association between the two variables. For the data given in Figure 1, $R = 0.977$ implying a strong positive correlation between the fuel consumption and the weight of the automobile.

3 Regression Analysis: Method of Least Squares

Once we have established that a strong correlation exists between x and y, we would like to find suitable coefficients a and b so that we can represent y using a best fit line $\hat{y} = ax + b$ within the range of the data. The method of least squares is a very common technique used for this purpose. The rationale used here is as follows. For each pair of observations (x_i, y_i), we define the *error* e_i as

$$e_i = (ax_i + b - y_i). \tag{2}$$

Now, we find a and b in such a way that the sum of the squared errors over all the observations is minimized. i.e., the quantity we are interested in minimizing is

$$S(a,b) = \sum_{i=1}^{n} [ax_i + b - y_i]^2. \tag{3}$$

We know from calculus that to minimize this, we need $\partial S/\partial a \equiv 0$ and $\partial S/\partial b \equiv 0$. These conditions yield

$$\begin{aligned} nb + \left(\sum_{i=1}^{n} x_i\right) a &= \sum_{i=1}^{n} y_i \\ \left(\sum_{i=1}^{n} x_i\right) b + \left(\sum_{i=1}^{n} x_i^2\right) a &= \sum_{i=1}^{n} x_i y_i. \end{aligned} \tag{4}$$

Eq. 4 gives two linear equations in a and b, which can be solved to get

$$a = \frac{n \sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}, \tag{5}$$

with b obtained through subsequent substitution of a in either of the two equations given by Eq. 4.

In the case of the data given in Figure 1, the best fit line has a slope of 1.64 and intercept of -0.36. Or in other words, $\hat{y} = 1.64x - 0.36$. Note that this is only a best fit line which can be used to compute the fuel consumption given the weight *within or very close to the range* of the measurements. Its *predictive* power is rather limited. For instance, for $x = 0$, we get $y = -0.36$, which is non-physical. A physical *model* for the fuel consumption would have predicted 0 consumption of fuel for 0 weight.

How are the slope and the intercept of the best fit line related to the correlation coefficient? To examine this, we rewrite Eq. 5 as

$$\begin{aligned} a &= \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \\ &= \frac{\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)/n}{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2/n} \\ &= \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{\sum_{i=1}^{n}\left(x - \mu_x\right)^2} \quad \text{(Verify this step)} \\ &= \frac{(n-1)R\sigma_x\sigma_y}{(n-1)\sigma_x^2} \quad \text{(See Eq. 1)} \\ &= R\frac{\sigma_y}{\sigma_x}. \end{aligned} \tag{6}$$

Similarly, from the first of Eq. 4 and the above result we get

$$b = \mu_y - R\frac{\sigma_y}{\sigma_x}\mu_x, \tag{7}$$

so that the equation of the best fit line can be represented by

$$\hat{y} = \mu_y + (R\frac{\sigma_y}{\sigma_x})(x - \mu_x). \tag{8}$$

4 Tests for the Regression Equation

Correlation analysis gives us the correlation coefficient which is a measure of the strength and the direction of the linear association between the variables. This information can be used to decide the suitability of model calibration using a linear regression analysis. The square of the correlation coefficient may be thought of as the percentage of the total variation in y that is explained by the association of y and x. Hence, for $R = 1$, all the variation is explained by the linear association between the two variables. In this case, all the observations will lie on a straight line of slope σ_y/σ_x, passing through the point (μ_x, μ_y).

Another measure used to evaluate the goodness of fit is the standard deviation of the errors σ_e, defined as

$$\sigma_e^2 = \frac{1}{\nu} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2, \tag{9}$$

where $\hat{y}_i = ax_i + b$ and ν represents the number of degrees of freedom. The number of degrees of freedom is equal to the sample size minus the number of unknowns estimated in by the regression procedure. In this case, we have the slope and the intercept as the unknowns, so $\nu = n - 2$. If σ_e is very small, we attribute a high reliability to the results of the regression analysis.

4.1 The Hypotheses behind Regression Analysis

The method of least squares explained above makes at least 4 assumptions, the adherence to which may be checked a posteriori. These assumptions concern with the error $e_i = \hat{y}_i - y_i$, i.e., the difference between the best fit prediction and the observation. The assumptions are that these errors (a). are mutually independent (b). have zero mean (c). have a constant variance across all the values of the statistical variables and (d). are *normally* distributed. Violation of these assumptions can be identified in many cases by simply examining a plot of e_i vs. x_i. Note that the first of Eq. 4 guarantees that the mean value of e_i is 0 within the precision of the computation (To see this better rewrite that equation as $\sum(b + ax_i - y_i) = 0$). However, this is not necessarily true of non-linear regression analysis.

4.2 Nonlinear Models and Linear Regression

In many cases, simple transformation of variables help to recast a non-linear model in a linear form. For instance, suppose we wish to fit certain kinetic data to the exponential model $\hat{y} = \alpha \exp(\beta x)$. There are non-linear regression programs which accomplish this task, but we can use a linear regression procedure if we try to fit $y^* \equiv \ln(\hat{y})$ vs. x. This is because, we have $\ln(\hat{y}) = \ln(\alpha) + \beta x$ and letting $y^* = \ln(\hat{y})$ and $\alpha^* = \ln(\alpha)$, we get $y^* = \alpha^* + \beta x$. The linear regression procedure will give α^* and β for the best fit.