Chapter 06.01
Statistics Background of Regression Analysis

After reading this chapter, you should be able to:

1. *review the statistics background needed for learning regression, and*
2. *know a brief history of regression.*

Review of Statistical Terminologies

Although the language of statistics may be used at an elementary and descriptive level in this chapter, it makes an integral part of our every day discussions. When two friends talk about the weather (whether it will rain or not - probability), or the time it takes to drive from point A to point B (speed - mean or average), or baseball facts (all time career RBI or home runs of a sportsman -sorting, range), or about class grades (lowest and highest score - range and sorting), they are invariably using statistical tools. From the foregoing, it is imperative then that we review some of the statistical terminologies that we may encounter in studying the topic of regression. Some key terms we need to review are sample, arithmetic mean (average), error or deviation, standard deviation, variance, coefficient of variation, probability, Gaussian or normal distribution, degrees of freedom, and hypothesis.

Elementary Statistics

A statistical sample is a fraction or a portion of the whole (population) that is studied. This is a concept that may be confusing to many and is best illustrated with examples. Consider that a chemical engineer is interested in understanding the relationship between the rate of a reaction and temperature. It is impractical for the engineer to test all possible and measurable temperatures. Apart from the fact that the instrument for temperature measurement have limited temperature ranges for which they can function, the sheer number of hours required to measure every possible temperature makes it impractical. What the engineer does is choose a temperature range (based on his/her knowledge of the chemistry of the system) in which to study. Within the chosen temperature range, the engineer further chooses specific temperatures that span the range within which to conduct the experiments. These chosen temperatures for study constitute the sample while all possible temperatures are the population. In statistics, the sample is the fraction of the population chosen for study.

The location of the center of a distribution - the mean or average - is an item of interest in our every day lives. We use the concept when we talk about the average income, the class average for a test, the average height of some persons or about one being overweight (based on the average weight expected of an individual with similar characteristics) or not. The arithmetic mean of a sample is a measure of its central tendency and is evaluated by dividing the sum of individual data points by the number of points.

Consider Table 1 which 14 measurements of the concentration of sodium chlorate produced in a chemical reactor operated at a pH of 7.0.

12.0	15.0	14.1	15.9	11.5	14.8	11.2	13.7	15.9	12.6	14.3	12.6	12.1	14.8

Table 1 Chlorate ion concentration in mmol/cm^3

The arithmetic mean $\bar{y}$ is mathematically defined as

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \tag{1}$$

which is the sum of the individual data points y_i divided by the number of data points n.

One of the measures of the spread of the data is the range of the data. The range R is defined as the difference between the maximum and minimum value of the data as

$$R = y_{\max} - y_{\min} \tag{2}$$

where

$y_{\max}$ is the maximum of the values of y_i, $i = 1,2,...,n$,

$y_{\min}$ is the minimum of the values of y_i, $i = 1,2,...,n.$.

However, range may not give a good idea of the spread of the data as some data points may be far away from most other data points (such data points are called outliers). That is why the deviation from the average or arithmetic mean is looked as a better way to measure the spread. The residual between the data point and the mean is defined as

$$e_i = y_i - \bar{y} \tag{3}$$

The difference of each data point from the mean can be negative or positive depending on which side of the mean the data point lies (recall the mean is centrally located) and hence if one calculates the sum of such differences to find the overall spread, the differences may simply cancel each other. That is why the sum of the square of the differences is considered a better measure. The sum of the squares of the differences, also called summed squared error (SSE), S_t, is given by

$$S_t = \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2 \tag{4}$$

Since the magnitude of the summed squared error is dependent on the number of data points, an average value of the summed squared error is defined as the variance, σ^2

$$\sigma^2 = \frac{S_t}{n-1} = \frac{\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2}{n-1} \tag{5}$$

The variance, σ^2 is sometimes written in two different convenient formulas as

$$\sigma^2 = \frac{\sum_{i=1}^{n} y_i^{\,2} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}}{n-1} \tag{6}$$

or

$$\sigma^2 = \frac{\sum_{i=1}^{n} y_i^{\,2} - n\bar{y}^2}{n-1} \tag{7}$$

However, why is the variance divided by $(n-1)$ and not n as we have n data points? This is because with the use of the mean in calculating the variance, we lose the independence of one of the data points. That is, if you know the mean of n data points, then the value of one of the n data points can be calculated by knowing the other $(n-1)$ data points.

To bring the variation back to the same level of units as the original data, a new term called standard deviation, σ, is defined as

$$\sigma = \sqrt{\frac{S_t}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}} \tag{8}$$

Furthermore, the ratio of the standard deviation to the mean, known as the coefficient variation $c.v$ is also used to normalize the spread of a sample.

$$c.v = \frac{\sigma}{\bar{y}} \times 100 \tag{9}$$

Example 1

Use the data in Table 1 to calculate the

a) mean chlorate concentration,
b) range of data,
c) residual of each data point,
d) sum of the square of the residuals.
e) sample standard deviation,
f) variance, and
g) coefficient of variation.

Solution

Set up a table (see Table 2) containing the data, the residual for each data point and the square of the residuals.

Table 2 Data and data summations for statistical calculations.

i	y_i	$y_i^{\,2}$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
1	12	144	-1.6071	2.5829
2	15	225	1.3929	1.9401
3	14.1	198.81	0.4929	0.24291
4	15.9	252.81	2.2929	5.2572
5	11.5	132.25	-2.1071	4.4401
6	14.8	219.04	1.1929	1.4229
7	11.2	125.44	-2.4071	5.7943
8	13.7	187.69	0.0929	0.0086224

9	15.9	252.81	2.2929	5.2572
10	12.6	158.76	-1.0071	1.0143
11	14.3	204.49	0.6929	0.48005
12	12.6	158.76	-1.0071	1.0143
13	12.1	146.41	-1.5071	2.2715
14	14.8	219.04	1.1929	1.4229
$\sum_{i=1}^{14}$	190.50	2625.3	0.0000	33.149

a) Mean chlorate concentration as from Equation (1)

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{190.5}{14} = 13.607$$

b) The range of data as per Equation (2) is

$$\begin{aligned} R &= y_{\max} - y_{\min} \\ &= 15.9 - 11.2 \\ &= 4.7 \end{aligned}$$

c) Residual at each point is shown in Table 2. For example, at the first data point as per Equation (3)

$$\begin{aligned} e_1 &= y_1 - \bar{y} \\ &= 12.0 - 13.607 \\ &= -1.6071 \end{aligned}$$

d) The sum of the square of the residuals as from Equation (4) is

$$\begin{aligned} S_t &= \sum_{i=1}^{n} (y_i - \bar{y})^2 \\ &= 33.149 \text{ (See Table 2)} \end{aligned}$$

e) The standard deviation as per Equation (8) is

$$\begin{aligned} \sigma &= \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}} \\ &= \sqrt{\frac{33.149}{14-1}} \\ &= 1.5969 \end{aligned}$$

f) The variance is calculated as from Equation (5)

$$\begin{aligned} \sigma^2 &= (1.597)^2 \\ &= 2.5499 \end{aligned}$$

The variance can be calculated using Equation (6)

$$\sigma^2 = \frac{\sum_{i=1}^{n} y_i^{\,2} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}}{n-1}$$
$$= \frac{2625.31 - \frac{(190.5)^2}{14}}{14-1}$$
$$= 2.5499$$

or by using Equation (7)

$$\sigma^2 = \frac{\sum_{i=1}^{n} y_i^{\,2} - n\bar{y}^2}{n-1}$$
$$= \frac{2625.3 - 14 \times 13.607^2}{14-1}$$
$$= 2.5499$$

g) The coefficient of variation, $c.v$ as from Equation (9) is

$$c.v = \frac{\sigma}{\bar{y}} \times 100$$
$$= \frac{1.5969}{13.607} \times 100$$
$$= 11.735\%$$

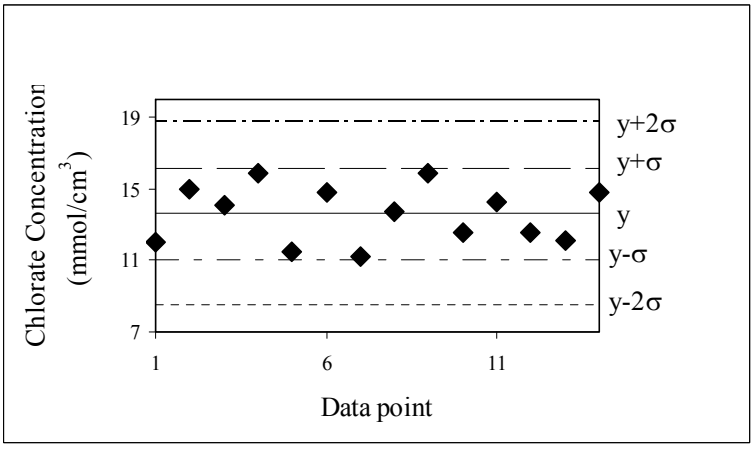

Figure 1 Chlorate concentration data points.

A Brief History of Regression

Anyone who is familiar with the Pearson Product Moment Correlation (PPMC) will no doubt associate regression principles with the name of Pearson. Although this association may be right, the concept of linear regression was largely due to the work of Galton, a cousin of Charles Darwin of the evolution theory fame. Sir Galton's work on inherited characteristics of sweet peas led to the initial conception of linear regression. His treatment

of regression was not mathematically rigorous. The mathematical rigor and subsequent development of multiple regression were due largely to the contributions of his assistant and co-worker - Karl Pearson.

It is however instructive to note for historical accuracy that the development of regression could be attributed to the attempt at answering the question of hereditary - how and what characteristics offspring acquire from their progenitor. Sweet peas were used by Galton in his observations of characteristics of next generations of a given species. Despite his poor choice of descriptive statistics and limited mathematical rigor, Galton was able to generalize his work over a variety of hereditary problems. He further arrived at the idea that the differences in regression slopes were due to differences in variability between different sets of measurements. In today's appreciation of this, one can say that Galton recognized the ratio of variability of two measures was a key factor in determining the slope of the regression line.

The first rigorous treatment of correlation and regression was the work of Pearson in 1896. In the paper in the Philosophical Transactions of the Royal Society of London, Pearson showed that the optimum values of both the regression slope and the correlation coefficient for a straight line could be evaluated from the product-moment,

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

where $\bar{x}$ and $\bar{y}$ are the means of observed x and y values, respectively. In the 1896 paper, Pearson had attributed the initial mathematical formula for correlation to Auguste Bravais' work fifty years earlier. Pearson stated that although Bravais did demonstrate the use of product-moment for calculating the correlation coefficient, he did not show that it provided the best fit for the data.

REGRESSION	
Topic	Statistics Background for Regression
Summary	Textbook notes for the background of regression
Major	All engineering majors
Authors	Egwu Kalu, Autar Kaw
Date	October 11, 2008
Web Site	http://numericalmethods.eng.usf.edu