Six Sigma – Part 4: DMAIC: Improving and Controlling

An Online Continuing Education Course for Engineers

Course Number: P-4013

Credit: 4 Hours / 4 PDH / 4 CPD

Module 1

Forecasting Future Performance

In This Module

- Quantifying relationships between variables
- Calculating and interpreting the correlation coefficient
- Fitting predictive $Y = f(X) + \varepsilon$ curves to data
- Checking the validity of fitted prediction curves

Part II of this book shows you how to launch an improvement project and identify the potential variables that influence a critical outcome. Part III covers tools and methods for analyzing the situation and determining which *X*s are actually critical and which are trivial.

But in this part, you're turning over a whole new leaf. With the critical few factors known, you're ready to begin devising improvements. In this module, we show you how to quantify potential improvement effects so you know how much improvement in the output you'll get for any given adjustment to the critical inputs. We also help you understand the relationships between the input variables and the critical outputs. Knowing the outcome of a potential improvement without spending the resources to test it out is the essence of Six Sigma improvement power!

Seeing the Correlation

Scatter plots (explained in a previous course) are a great way to visually discover and explore relationships between variables — both between *Y*s and *X*s and between *X*s and *X*s. In a scatter plot, you graph the values of one variable against paired values of another variable. As an example, Table 1-1 is a list of paired data for the curb weight (in pounds) of some common automobiles and their corresponding fuel economies (in miles per gallon).

Table 1-1 Automobile Curb Weight versus Fuel Economy

Make/Model	*Curb Weight (lbs.)*	*Fuel Economy (mpg)*
Toyota Camry	3,140	29
Toyota Sequoia	4,875	17
Honda Civic	2,449	35
Land Rover Discovery	4,742	16
Mercedes-Benz S500	4,170	20
VW Jetta Wagon	3,078	27
Chrysler 300	3,715	22
Chevrolet Venture	3,838	23
Hyundai Tiburon	2,940	27
Dodge Ram 2500 Quad	6,039	11

A data point for each automobile in the study is plotted in Figure 1-1. The scatter plot shows a negative relationship between the curb weight of the vehicle and its fuel economy — the heavier the car, the lower its fuel economy. The plot also shows that the relationship between the two variables is approximately linear, meaning that its shape approximately follows a straight line. Finally, you can see from the figure that the relationship between the variables is fairly strong, as evidenced by the tight clustering of the plotted data points around the drawn line approximating the relationship. But how do you put numbers to this relationship? You use *correlation,* which shows how closely two variables' relationship follows a linear pattern.

Correlation tells you only how linear the relationship between the variables is. It may miss more-complicated relationships where the variables follow a nonlinear pattern. Always include a graphical scatter plot when doing a correlation analysis. That way, you can visually check to make sure the variable relationship really is linear.

To quantify how linear the relationship is between two variables, you use the following formula to calculate the *correlation coefficient* (*r*):

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma_x}\right)\left(\frac{y_i - \bar{y}}{\sigma_y}\right)$$

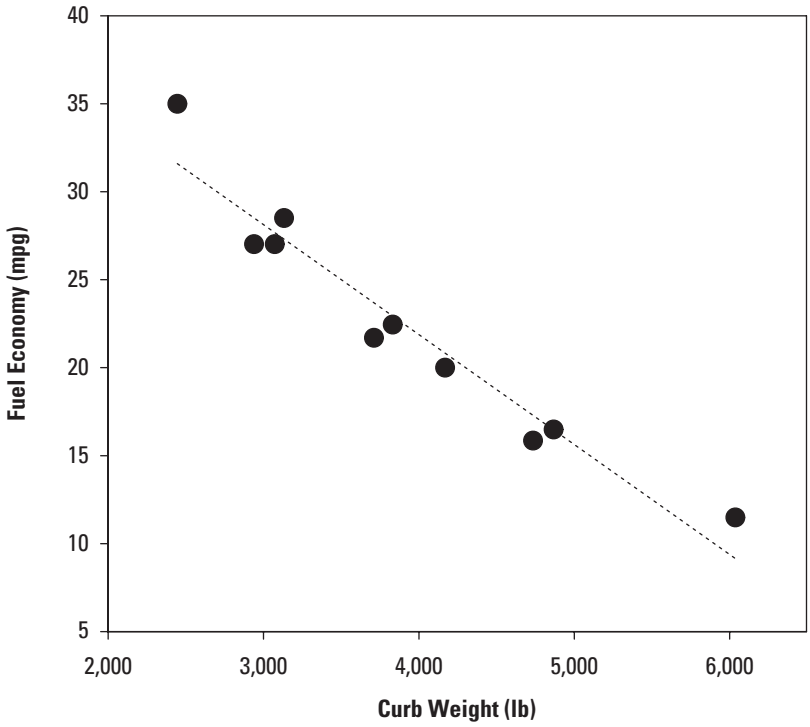

Figure 1-1: Scatter plot of vehicle curb weight versus fuel economy.

where

- n is the number of data pairs.
- x_i and y_i are the individual x-variable and y-variable measurements.
- $\bar{x}$ and $\bar{y}$ are the averages of the X and Y measurements, respectively.
- σ_X and σ_Y are the standard deviations of the X and Y measurements, respectively.
- Σ is a capital Greek letter telling you to add up all the $\left(\frac{x_i-\bar{x}}{\sigma_x}\right)\left(\frac{y_i-\bar{y}}{\sigma_y}\right)$ terms, from 1 to n.

The calculated correlation coefficient is always between –1 and 1. Remember:

- **The sign of *r* tells you the direction of the relationship between the variables.** If r is greater than zero (positive), that means that the variable relationship is positive; if the value of one variable increases, the other variable also increases. If r is less than zero (negative), the variable relationship is negative; if the value of the independent variable increases, the value of the dependent variable decreases, and vice versa.

- **The absolute value of *r* tells you how strong the relationship is.** The closer *r* gets to –1 or 1, the stronger the variable relationship is. An *r* equal to 1 or –1 indicates a perfect linear relationship, with all points being exactly on the line. An *r* close to 0 indicates that the data don't fit a linear model. For the automobile fuel economy example introduced earlier, the calculated correlation coefficient *r* is

$$r = \frac{1}{10-1}\sum_{i=1}^{10}\left(\frac{x_i - 3{,}899}{1{,}087}\right)\left(\frac{y_i - 23}{7}\right) = \frac{1}{9}(-8.738) = -0.971$$

An *r* of –0.971 verifies that the relationship between the two variables is indeed linear and is negative. Also, this *r* value is very close to –1, telling you that the relationship is very strong.

Correlation basically just confirms the existence of a linear relationship between two variables and quantifies how linear that relationship is. What correlation does *not* tell you is how much a given change in one variable changes a related variable. To get that kind of information, you need to become acquainted with some predictive tools.

Correlation doesn't equal causation. The fact that two variables correlate doesn't mean that one *causes* the other. For example, studies show that a person's reading comprehension ability (*Y*) is correlated with his or her height (*X*), so you may conclude that height *causes* reading comprehension. But if you think about it for a second, you realize that young children haven't yet developed cognition and reading skills. In the teenage years, physical growth continues along with maturation of mental and reading abilities. By the time you're a full-grown adult, your brain and mental abilities have fully developed. Thus, you can see that height is just an indirect indicator of overall maturation and growth (including cognitive abilities), not a direct cause of those abilities. So be very careful: Don't assume a causal link when you see correlation.

Getting a Handle on Curve Fitting

A step beyond correlation (see the preceding section) is curve fitting. In *curve fitting,* you actually determine the equation for the curve that best fits your data. Armed with this information, you know quantitatively what effect one variable has on another, which variables are significant influencers, and which ones are just in the noise. Finally, you know how much of the system behavior your equation does *not* explain.

In some rare cases, you know the exact details of the $Y = f(X) + \varepsilon$ equation relating the *X*s to the *Y* without having to do any curve fitting, either because you have a very mature understanding of the physics of your process or system or because you have some other source of knowledge. These situations are called *deterministic* because you know with certainty that setting

the input *X*s to certain values always leads to the exact same value for the output *Y,* even when the process is repeated. For the vast majority of cases, however, you don't know the exact relationship between the *X*s and the *Y* or *Y*s due to the system's complexity and a human inability to address all the factors that truly exert influence on the output. Because of this natural limitation, repeating the same input values in the system doesn't always produce the same output performance. These situations are called *statistical.*

The goal of curve fitting is to develop an approximate equation that describes the system's or process's st cal behavior as s possible. When you work to create an approx t has a single output *Y* and a sing ed *simple linear reg* s more tha

Fin ion

In sim nt Y_i can be

$$Y =$$

The first lly, the deco

Time war he $\beta_0 + \beta_1 X$ p by itself te s you the lin ter value equal

In simple lin determine values for β_0 and β_1 so that the re observed $X - Y$ data as closely as possible with the mini amount of error. Then you determine how wide the ε portion needs to be so that it accounts for all the extra *Y* variation that the line doesn't already capture. You calculate β_0 and β_1 from the following equations:

$$\beta_1 = \frac{\sum (x_i - \bar{x})\, y_i}{\sum (x_i - \bar{x})^2}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

where x_i and y_i are the paired data points and $\bar{x}$ and $\bar{y}$ are the calculated averages for all the *X* points and all the *Y* points, respectively.