The Regression Control Chart

B. J. MANDEL

Post Office Department, Washington, D. C.

Linear regression and control chart theory are combined to yield an effective technique for controlling manhours in a situation in which the workload is variable but highly correlated with the manhours. Worked examples are discussed in detail. Several managerial applications are suggested.

Introduction

In a previous article [9], the statistical programs of the Post Office Department were described and exemplified by a large variety of postal activities to which statistical science is being, or will be, applied. Some of the applications involved the use of conventional control chart techniques. This paper discusses in detail the combination of the conventional control chart and regression analysis. This combination, referred to as the regression control chart, was the subject of a dissertation [10] and was recently reported on in Quality Progress [1].

The regression control chart has proved useful for a variety of postal management problems and offers possibilities for more widespread application in government, business, and industry. Several years of experience with this technique are described, and detailed instructions for setting up this type of control chart are provided.

Origin of the Regression Control Chart

The earliest reference in the quality control literature to the regression control chart appeared in a paper by DiPaola [4]. However, the chart described was for the simultaneous control of two variables and might better have been labeled a correlation control chart. This type of control problem was later treated more extensively by Jackson [8] and illustrated by Weis [12]. A description of a regression control chart in the sense intended in the present article is given by Wallis and Roberts [11]. These authors exemplify the use of confidence limits about a regression line for control purposes, but point out the desirability of a tolerance band. This topic, and the standard error of the cumulative sum are taken up in the Appendix.

A study of the manhours required to handle and sort mail in the 74 major post offices of the United States revealed a clear-cut linear relation between the number of manhours expended and the volume of mail handled. By fitting regression lines to historical data and establishing limits around the regression line, a regression control chart was formed for each of the 74 post offices involved. One of these charts has been used to exemplify this procedure in the section entitled "Establishing the Regression Control Chart".

By posting current data on manhours used to process specified volumes of mail and observing the location of these points on the regression control chart, it became possible to detect "out-of-control" points for which it was worth management effort to seek assignable causes. These studies were made in 1965, 1966, and 1967 without the participation of operating staff, merely as an experiment at Headquarters in Washington to determine the feasibility of adopting this approach to control the use of manhours in mail processing activities on a total post office basis. The results were encouraging.

In 1967, four large post offices were selected to participate in a study of the value of the regression control chart under actual operating conditions, but at the lowest supervisory levels. The results were sufficiently conclusive to cause Departmental management to adopt this technique on a more widespread basis starting July 1968. The success of this technique to date warrants (1) description of the basic elements of the regression control chart, (2) a review of statistical and technical analyses required to establish and use this chart, and (3) giving illustrative applications to a variety of postal management problems.

Dr. Mandel is the Director, Office of Statistical Programs, Bureau of Finance and Administration, Post Office Department. He is a Senior Member of ASQC. The opinions expressed here are those of the author and do not necessarily represent those of the Department.

Vol. 1, No. 1, January 1969

1

Journal of Quality Technology
Elements of the Regression Control Chart

The conventional control chart uses a line of average performance with upper and lower control limits paralleling this center line. For example, the $\bar{X}$-chart has the average of averages $\bar{X}$ as the center line and $\bar{X} \pm 3\bar{X}$ as the upper and lower control lines. The probability of obtaining an out-of-control point when the process is actually in control is about 1 in 360. These three lines are all parallel to the $x$-axis, thus indicating that a single average is being controlled.

The regression control chart differs from the conventional control chart in several respects:

- It is designed to control a varying (rather than a constant) average; in post offices it is the man-hours $(y)$ expended which are to be controlled, and this variable is dependent upon the magnitude of the independent variable, mail volume $(x)$. The center line is the regression line, $Y = a + bx$, where $Y$ is used to represent the number of man-hours predicted from the regression equation, while $y$ is the number of man-hours observed in a given situation.

- The control limit lines are parallel to the regression line rather than the horizontal axis. This property of parallelism is discussed later. By retaining the form of the regression plot, the relationship between the variables is not obscured. An example introduced later will show the additional usefulness of constructing a conventional control chart using the deviations of actual man-hours used from those predicted by the regression line.

- The computations for the regression control chart are somewhat more complex and time-consuming than those for the conventional control chart. The standard deviation for the regression chart is taken as the so-called "standard error of estimate" of the regression line. This is the standard deviation estimate based on the deviations of the observed values about the regression line. It should not be confused with the standard error of a predicted value of the dependent variable.

- The regression control chart is appropriate for a number of purposes which the conventional control chart does not readily serve. For example, it provides a basis for measuring the gains (or losses) in man-hours; for forecasting manpower needs and scheduling manpower resources; for formulating an improved annual budget; for evaluating deviations from budgeted targets; for establishing flexible standards of performance; and for analyzing the efficiency of the organizational units covered.

The $\bar{X}$-control chart and the regression control chart are shown schematically in Figures 1 and 2. Some of the basic assumptions of the $\bar{X}$-chart and the regression chart are similar, but there are also some major differences, as the following comparison shows:

The $\bar{X}$-Chart

It is assumed that a constant-cause system exists such that random-sample averages of the quality characteristic being measured (based on equal-sized samples) are normally distributed, with a mean estimated by the average of averages $\bar{X}$, and with standard deviation $\sigma = \sigma/\sqrt{n}$ ($\sigma$ being the standard deviation of items in the universe and $n$ the number of items in the sample).

The Regression Chart

It is assumed that the $y$ values (the dependent variable) are linearly related (causally) to the $x$ values (the independent variable). For each specific $x$ value, it is assumed that the $y$ values are normally and independently distributed with a mean value estimated from the regression line, and with a standard error which is independent of the value of $x$ and is estimated from the deviations of the actual observations from the $Y$ values estimated from the regression line.

Establishing the Regression Control Chart

To illustrate the steps required to establish a regression control chart, data from one of the 74...
offices included in the 1965–1967 study conducted by the Post Office Department have been used. The records of this office for the fiscal years 1962–1963 yielded the results shown in Table 1.

As an initial step, the data from Table 1 are plotted in the form of a scattergram, which is shown in Figure 3. The purpose of the scattergram is primarily to check on the linearity of the relationship and to detect atypical points.

It is recognized that the productivity of an office, measured in terms of average number of pieces of mail handled per manhour is affected by numerous factors, some of which remain relatively constant while others may vary substantially. Some of the factors which remain fairly stable in a given post office over a period of time are:

- The composition of the supervisory staff and local operating policies.
- The skill, industry, and natural ability of the operating staff.
- The working conditions and working quarters.
- The extent to which mechanized equipment is used, and its comparative efficiency.
- The methods of processing used.

Points which depart from the linear pattern displayed by the graph are investigated. If they are found to be due to assignable causes not normally associated with the constant-cause system, they are excluded from the computations used to establish a preliminary control chart.

A scattergram of the data in Table 1 clearly showed that three points deviated enough from the general linear pattern to require investigation. Two of these points were found to be associated with the four-week periods which included Christmas (Period 7 of each fiscal year). An explanation for these is that processing of the high volume of Christmas mail involves the use of temporary help, frequently working in crowded quarters. Thus, data including the Christmas rush represent an unusual or abnormal condition and, therefore, are not part of the constant-cause system for the rest of the year. They have been excluded from the computations leading to the preliminary regression control chart.

Another extreme point was that for Period 6 of fiscal year 1963. Investigation showed that it was caused by an abnormal amount of machine downtime and was, therefore, also excluded from the computations. Figure 3 shows the scattergram for the remaining points.

The following values were computed using the data in Table 1, excluding the three abnormal points:

\[ N \text{ (number of pairs of values) = 23} \]
\[ r \text{ (correlation coefficient) = +0.91} \]
\[ r^2 \text{ (explained variation) = 0.82} \]
\[ s_r \text{ (standard error of estimate of manhours) = 15.92 (thousands)} \]
\[ \bar{y} \text{ (average manhours) = 634.0 (thousands)} \]
period of fiscal year 1964, the ABCD office handled 162 million pieces of mail and used 628 thousand hours to process this mail, is there a gain or loss in productivity and is this acceptable performance? When the volume (x = 162 million) is located on the center line of the chart in Figure 4, it is determined that Y = 592 thousand manhours should be used for this size workload, based on history-period performance. In other words, 31 thousand more hours (623 thousand - 592 thousand) than would be expected were actually used during this four-week period. Management thus knows that performance was below normal, and a management decision whether or not to investigate the cause is essential.

In-Control Points

Without the guides offered by the control chart limits, management might be concerned with the use of 31 thousand hours more than expected to process a given volume of mail; it could take action to review the situation and determine the underlying causes. However, with the control chart as a guide, this “excess” usage of hours may be accepted as normal in relation to history-period performance, since the actual hours used for the stated volume fell within the control limits shown in Figure 4.

Thus, the regression control chart, like the conventional control chart, can serve as a performance measurement and control tool, utilizing a “management by exception” principle that when variations occur which are likely to be due to common or usual causes, accept them, since investigation is not likely to yield worthwhile findings.

Out-of-Control Points

During another four-week period, when the volume of mail processed was 171 million pieces, the

Using the Control Chart

The usefulness of the regression control chart is indicated by the variety of applications to which it can be put. The use of two-sigma control limits is a matter of management decision based on economics and experience. For our applications the narrower limits, with a higher risk of false alarms, have worked quite well. See the Appendix for technical comments.

Performance Control

The first use is in maintaining control over performance (or manhour usage) on a continuing basis. Thus, for example, if during the first accounting

Journal of Quality Technology

Vol. 1, No. 1, January 1969
When a point falls above the upper control limit, it indicates that productivity was probably significantly lower than the level of performance established in the recent past periods. This could be caused by poor performance, poor scheduling of manpower, a machine breakdown, an unusual mail mix (such as a high proportion of parcel post), a reporting error, etc.

When a point falls below the lower control limit, it could be due to a real improvement in productivity (either temporary or permanent). Or it could be due to errors of over-reporting of volume or under-reporting of man-hours. Note that a point falling outside the upper control limit, even though it is an indication of exceptionally good performance, is cause for investigation just as much as a point falling outside the upper control limit, since it may benefit management to find the reason and adopt practices which lead to efficient performance.

The example illustrates the use of a rather long time-interval of four weeks, covering an entire post office. However, the experiment in the four large post offices dealt with the individual operating units of these post offices and the time interval used was a workday of eight hours. With this reduced time interval, quicker detection of assignable causes and remedial action can be taken, where necessary, to help improve performance or prevent slippage even more effectively.

**Measuring Progress**

A simplification of the regression control chart is achieved by the use of the difference between expected and actual man-hours used, and additional information can be obtained from the regression analysis. The difference between predicted and measured manhours $T - y$ can be plotted against time in the form of a conventional control chart. In this way, the difficulty of analyzing results from a cluttered scattergram is overcome and trends and trends in time can be easily observed. Also, the cumulative number of hours in excess of (or less than) expectation as the end of a given period can be easily determined and tested for significance. The data shown in Table 2 were calculated for these purposes as follows:

<table>
<thead>
<tr>
<th>Accounting Period</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gain $+$ or Loss $-$</td>
<td>$16$</td>
<td>$21$</td>
<td>$-14$</td>
<td>$11$</td>
<td>$15$</td>
<td>$16$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cumulative Gain $+$ or Loss $-$</td>
<td>$16$</td>
<td>$41$</td>
<td>$84$</td>
<td>$89$</td>
<td>$16$</td>
<td>$20$</td>
<td>$16$</td>
<td></td>
</tr>
<tr>
<td>Mail Volume $x_i$</td>
<td>$162$</td>
<td>$171$</td>
<td>$182$</td>
<td>$175$</td>
<td>$169$</td>
<td>$190$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cumulative Devia-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tion, $\Sigma(x_i - \bar{x})$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*This value was out of control on the regression control chart.

(a) Substitute into the estimating equation the volume $x$ actually handled during the period and calculate an estimate of the expected hours $Y$.

(b) Subtract the number of actual hours used $y$ from the number of hours expected to be used $Y$. This gives the number of hours gained (+) or lost (−) during the period.

(c) Cumulate the hours gained (or lost) so that a year-end (or other period) review can determine whether the regression line and control chart limits need revision.

At the end of the year, or earlier if necessary, the level of performance is checked to determine if it has changed significantly. This can be accomplished by a Student's $t$ test as follows:

\[
t = \frac{\sum (Y_i - y_i)}{s} \times n + \frac{\sum (x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}
\]

where $y_i$ = mail volume observations made after the regression control was established $x_i$ = manpower hours corresponding to the mail volume $y_i$

Vol. 1, No. 1, January 1969

Journal of Quality Technology
\[ Y' \text{ = predicted manpower hours corresponding to mail volume } x' \]

\[ x = \text{ mail volume used to establish the regression control chart} \]

\[ x = \text{ mean of the } x_i \]

\[ n = \text{ number of } x_i \text{ values} \]

\[ n = \text{ standard error of estimate of the regression based on history-period data} \]

This is the general formula. However, when \( N \) is large relative to \( n \), as is the case of the four post offices in the experiment, the last term in the denominator has no practical effect on the standard error of the cumulative sum. Therefore, for administrative feasibility it has been considered reasonable to use the following as an approximation for the standard error of the cumulative sum:

\[ n = \frac{x}{\sqrt{n + n}} \]

For example, to test whether the cumulative total mailhours for Post Office ABCD is significantly different from zero at the end of the eighth accounting period (see Table 2), we need the following quantities:

\[ \sum_{i=1}^{n} (y_i' - y_i^2) = (31 + 32 + 21 + 14 - 34^2 + 16) = 50 \]

\[ y = 18.92 \]

\[ n = 6 \] (data for the third and seventh accounting periods were not used)

\[ N = 23 \]

\[ \overline{y} = 174.4^8 \]

\[ \sum_{i=1}^{n} (x_i - 174.4)^8 = 3008 \]

Thus

\[ t = \frac{50}{52} = 1.54 \]

and, since this calculated value of \( t \) is less than 2, the critical value for significance at about the 5 percent level, it is concluded that the cumulative net loss is not statistically significantly different from zero. Hence, for the moment no action is indicated. Appropriate critical values for this test can be obtained from a \( t \) table using \( N - 2 \) degrees of freedom. For the 5 percent significance level for this example, the exact critical \( t \) value is 2.08. See the Appendix for technical details.

Runs and trends of data considered in time sequence can also indicate changes in the level of performance. If eight or nine, or eleven or twelve points constitute either a run (signs all positive or all negative) or a trend (steadily increasing or decreasing), a change would be indicated at about the 1 percent significance level for each type of test. A significant change in the performance level is the basis for establishing a new control chart.

The following policy is being proposed for making decisions regarding continuation of the original control chart:

(a) If the post office (or unit) has gained significantly in productivity, a new chart would be established on the basis of the current year's data, showing the new performance levels.

(b) If the post office (or unit) has lost significantly in productivity, a new chart would be established only if the loss were caused by a change in the constant-causing system, such as moving to new quarters, undertaking additional major mail processing workloads, etc. Otherwise, the original control chart would be maintained.

(c) If the post office (or unit) has lost a significant number of mailhours, but not significantly changed its conditions and operations, the original chart would be retained, and further effort to improve productivity in the current year would remain the objective of management.

Formulating an Improved Annual Budget

The regression control chart provides a basis for determining the number of hours to be allocated to each office each year for its mail processing activity. By eliminating the mail volume workload which is anticipated for each accounting period of the coming year into the estimating equation, the expected hours for each accounting period can be derived. The expected hours for each period are then added to yield an expected annual total. When appropriate, this grand total may be adjusted on the basis of judgment, to yield a more reasonable target gain (or loss) in productivity.

Evaluation of Variations from Established Targets

When mailhours have been allotted to an office on the basis of anticipated volume, it later becomes

\[ 1 \text{ This, of course, would require an equation based on a four-week time-interval, rather than a workday time-interval, as in the case of organizational units within a post office.} \]
### Table 3. Analysis of Actual Versus Target Manhours Variation for the XYZ Post Office for the Fourth Accounting Period for Fiscal Year 1964

<table>
<thead>
<tr>
<th>Workload (millions)</th>
<th>Manhours (thousands)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20.5</td>
<td>20.5</td>
</tr>
<tr>
<td>23.8</td>
<td>23.8</td>
</tr>
<tr>
<td>Difference (W-V)</td>
<td>0.7</td>
</tr>
<tr>
<td>Difference in Manhours Expected on Two Bases (W-E)</td>
<td>1.0</td>
</tr>
</tbody>
</table>

**Workload Target (W):**

- 20.5 Manhours Target (W)
- 23.8 Manhours Actual (A)
- Difference (W-A) = 3.3
- Explained by All Other Reasons (W-A) - (E-E) = 0.7

**Workload Actual (V):**

- 20.5 Manhours Preferred for Target Workload (E)*
- 23.8 Manhours Preferred for Actual Workload (E)*
- Difference in Manhours Expected on Two Bases (E-E) = 1.0

Possible to calculate deviations from allotted hours caused by actual volume experienced.

To analyze the differences between target hours and actual hours, it is necessary to break down the differences into two parts: that caused by larger (or smaller) volume, and that caused by gains (or losses) in productivity or other reasons. The balance sheet in Table 3 illustrates such an analysis. A statistical evaluation of (T-A) = (E-E) requires an estimate of its standard error, and this problem is under consideration as part of the research and improvement program.

### Measuring the Effect of Procedural or Processing Changes

The control chart can be used to measure the effect of new mail processing systems on productivity. For example, the effectiveness of productivity improvement programs on mail processing, such as mechanization, can be measured by comparing the number of expected hours with actual hours both before and after its installation, provided proper controls are established over the other variables.

### Checking the Accuracy of Data Reported at the Source

Another useful application of the regression chart is the Post Office Department deals with controlling the quality of data on postage revenue and weight of mail. These data are reported by clerk mail analysts who collect source data for one of our probability samples which provides estimates of revenue by class of mail and service. Figure 3 illustrates the high linear relationship found between the number of pieces of mail recorded in a given class with the corresponding amount of postage revenue. The establishment of suitable control limits has provided a basis for mechanically rejecting reports with a high likelihood that data have been erroneously recorded at the source by the clerk mail analysts. Investigation of the causes for such variations is, therefore, based on the knowledge that it is highly probable that an error has been made which needs to be corrected, or procedural misapplications exist which need to be remedied.

### Other Potential Uses

A number of additional uses may be made of the regression control chart data. For example, the estimating equation provides an objective basis for current manhours scheduling, once an estimate of volume is made. It may also provide a basis for detecting when an office has reached a state of overcrowding due to volume growth.

### Appendix

**Control Limits for the Regression Control Chart**

Although it is recognized that the control limits around the line of regression should widen progressively as they move away from the mean (see, for example, Draper and Smith [1]), two facts make the application of parallel control limits feasible. First, if the variation about the regression line is not very great, the control limits will not be too curved. Second, the parallel limit lines will give stricter control at distances extreme from the mean; and this is where greater importance tends to be associated with assignable causes.

The control lines in Figure 4 were placed at ±2σ. More exact control lines can be based on the standard error given by the following equation:
Figure 5. Correlation Between Postage Revenue Affixed and the Number of Pieces of First Class Mail Delivered on Residential Foot Routes, February 1966 (for clusters involving 101 to 300 pieces of mail).

\[
x_{r-c} = s_{r} \sqrt{\frac{N + 1}{N} + \frac{(x_{r} - \bar{x})^2}{\sum_{i=1}^{N} (x_{i} - \bar{x})^2}}
\]

where the value of \( x_{r} \) is the particular value of \( x \) at which the standard error of \( Y - y \) is wanted. These limits are curved, moving away from the regression line at the extremes. For the example considered, however, the control lines based on plus and minus twice the standard error given by the equation are less than three percent wider at the mean value of \( x \) and eight percent wider at the extremes of the data than the parallel lines. A discussion of the use of parallel limits of finite length around a regression line has been given by Graybill and Bowden [7].

Although it leads to even further complexity, it should be mentioned that it is tolerance limits which we would really like to have. The calculation of such limits which can be expected, with a specified confidence, to include a specified proportion of future observations is discussed in detail by Acton [2] and Bowden [3].

Standard Error of Cumulative Hours Gained (or Lost)

In order to measure progress in manhours usage and test the significance of the difference between the original regression line and the one developed from current data, we are concerned with cumulating the differences between new observations of manhours \( y_{r}' \), which are associated with mail volume \( x_{r}' \), and the corresponding calculated values of \( Y_{r}' \). The cumulative sum is

\[
\sum_{r=1}^{n} (Y_{r}' - y_{r}) = (a + bx_{r}' - y_{r}) + (a + bx_{r}' - y_{r}) + \cdots
\]

\[
= [g + b(x_{r} - \bar{x}) + y_{r}'] + [g + b(x_{r} - \bar{x}) - y_{r}]
\]

\[
= (N \cdot n) + \sum_{r=1}^{n} (x_{r} - \bar{x}) \rightarrow \sum_{r=1}^{n} y_{r}.
\]

The variance of this cumulative sum is

\[
\text{Var} \left[ \sum_{r=1}^{n} (Y_{r}' - y_{r}) \right] = \frac{n \cdot \sigma_{y}^{2}}{N} + \frac{n \cdot \sigma_{y}^{2}}{N} \sum_{r=1}^{n} (x_{r} - \bar{x})^2 + \frac{n \cdot \sigma_{y}^{2}}{N} \sum_{r=1}^{n} (x_{r} - \bar{x})^2 + \frac{n \cdot \sigma_{y}^{2}}{N} \sum_{r=1}^{n} (x_{r} - \bar{x})^2
\]

Letting \( C \) stand for \( \sum_{r=1}^{n} (Y_{r}' - y_{r}) \), we find the standard error to be

\[
s_{c} = \frac{\sigma_{y}}{\sqrt{N + n}} \sqrt{\sum_{r=1}^{n} (x_{r} - \bar{x})^2 + \frac{n \cdot \sigma_{y}^{2}}{N} \sum_{r=1}^{n} (x_{r} - \bar{x})^2}
\]

Note that \( x_{r} \) and \( x_{r}' \) are different. The \( x_{r} \) are from the original data used to establish the regression line, whereas the \( x_{r}' \) are \( n \) subsequent observations. Both \( \sigma \) and \( \beta \) defined as follows,

\[
\beta = \frac{1}{N} \sum_{r=1}^{n} x_{r}^2 - \bar{x}^2 = \frac{1}{N} \sum_{r=1}^{n} x_{r}^2
\]

are from the original set of data.

An abbreviated but less refined standard error of the cumulative sum, which has worked out reasonably well in practice, is

\[
s_{c} \approx \frac{\sigma_{y}}{\sqrt{N + n}}
\]

References
4. DiPAGLIO, P. P., "Use of Correlation in Quality Con-


Key Words: Control Chart, Regression, Masquerade Scheduling

Vol. 1, No. 1, January 1969