STATISTICAL REPORT

EXAMPLE THREE

STATISTICAL REPORT

CHRISTOPHER ZAPPE STATISTICAL CONSULTING SERVICESS

BLOOMINGTON, INDIANA

To: Dave Clements, financial manager

Subject: Forecasting overhead

Date: November 6, 2007

Dave, here is the report you requested. (See also the attached Figure 1 that contains the details of my analysis. By the way, it was done with the help of the StatPro add-in for Excel. If you plan to do any further statistical analysis, I would strongly recommend purchasing this add-in.) As I will explain in this report, regression analysis is the best-suited statistical methodology for your situation. It fits an equation to historical data, it uses this equation to forecast future values of overhead, and it provides a measure of accuracy of these forecasts. I believe you will be able to "sell" this analysis to your colleagues. The theory behind the regression analysis is admittedly complex, but the outputs I will provide are quite intuitive, even to nonstatisticians.

Objectives and data

To ensure that we are on the same page, I will briefly summarize my task. You supplied me with Bendrix monthly data, from July 1997 through June 1999, on three variables: Overhead (total overhead expenses during the month), MachHrs (number of machines hours used during the month), and ProdRuns (number of separate production runs during the month). You suspect that Overhead is directly related to MachHrs and ProdRuns, and you want me to quantify this relationship so that you can forecast *future* overhead expenses on the basis of (estimated) future values of MachHrs and ProdRuns. Although you did not state this explicitly in your requirements, I assume that you would also like a measure of the accuracy of the forecasts.

Statistical methodology

Fortunately, there is a natural methodology for solving your problem: regression analysis. Regression analysis was developed specifically to quantify the relationship between a single *response* variable and one or more *explanatory* variables (assuming that there is a relationship to quantify). In you case, the response variable is Overhead, the explanatory variables are MachHrs and ProdRuns, and from a manufacturing perspective, there is every reason to believe that

Overhead is related to MachHrs and ProdRuns. The outcome of the regression analysis will be a regression equation that can be used to forecast future values of Overhead and provide a measure of the accuracy of these forecasts. There are a lot of calculations involved in regression analysis, but statistical software such as StatPro takes care of these calculations easily, allowing you to focus on the interpretation of the results.

Preliminary analysis of the data

Before diving in to the regression analysis itself, it is always a good idea to check graphically for relationships between the variables. The best type of chart for your problem is a scatterplot, which shows the relationship between any pair of variables. The scatterplots in Exhibits 1 and 2 illustrate how Overhead varies with MachHrs and ProdRuns. In both charts the points follow a reasonably linear pattern from bottom left to upper right. That is, Overhead tends to increase linearly with MachHrs and ProdRuns, which is probably what you suspected. The correlations in the upper right corners of these plots indicate the strength of the linear relationships. The magnitudes of these correlations, 0.632 and 0.521, are fairly large. (The maximum possible correlation is 1.0.) They provide hope that regression analysis will yield reasonably accurate forecasts of overhead expenses.

Exhibit 1. Scatterplot of Overhead versus MachHrs

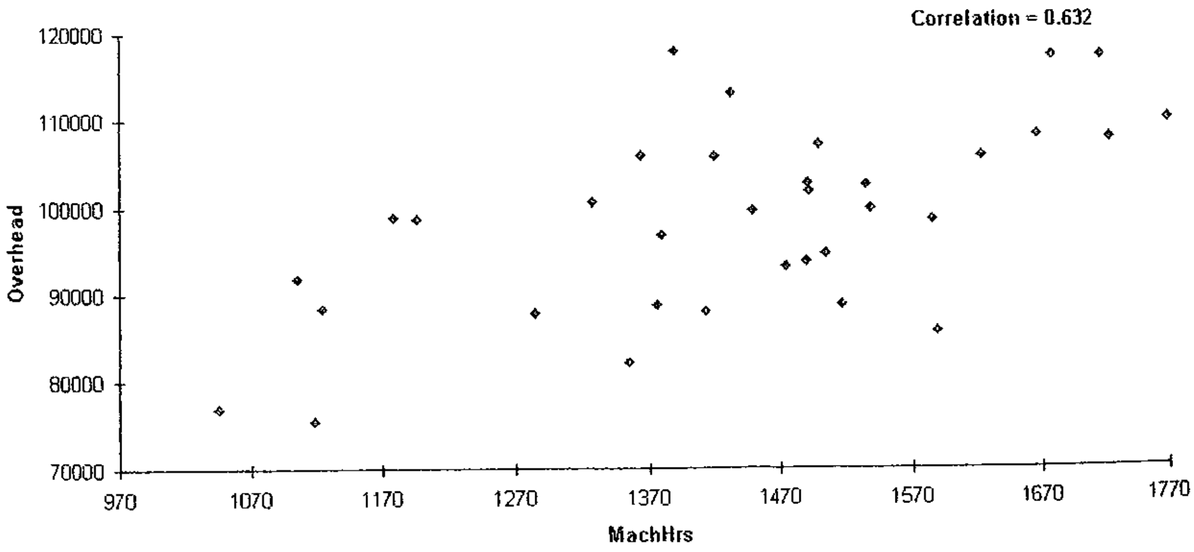
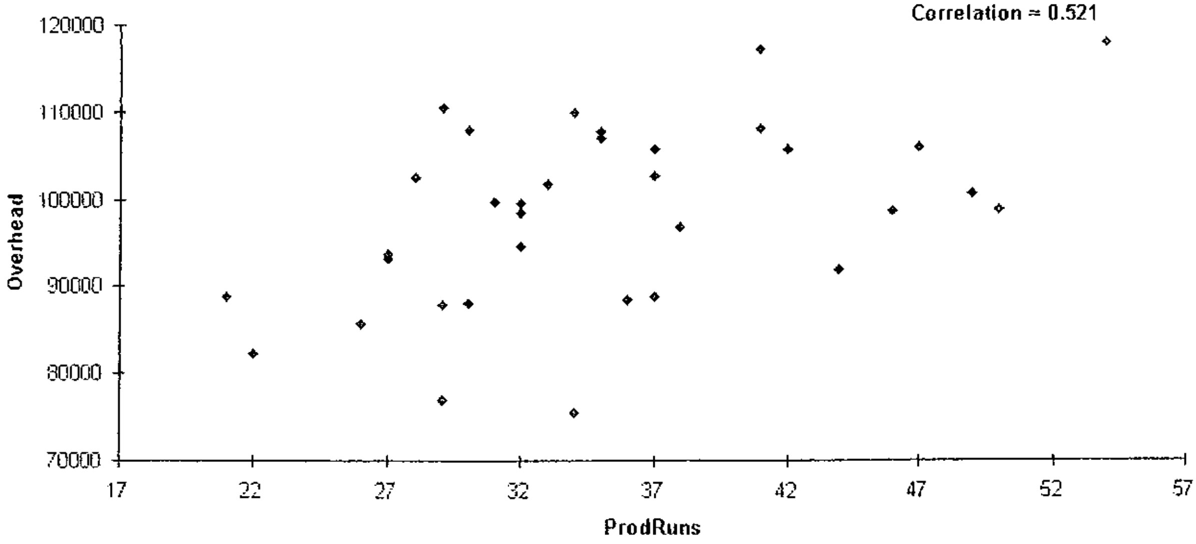
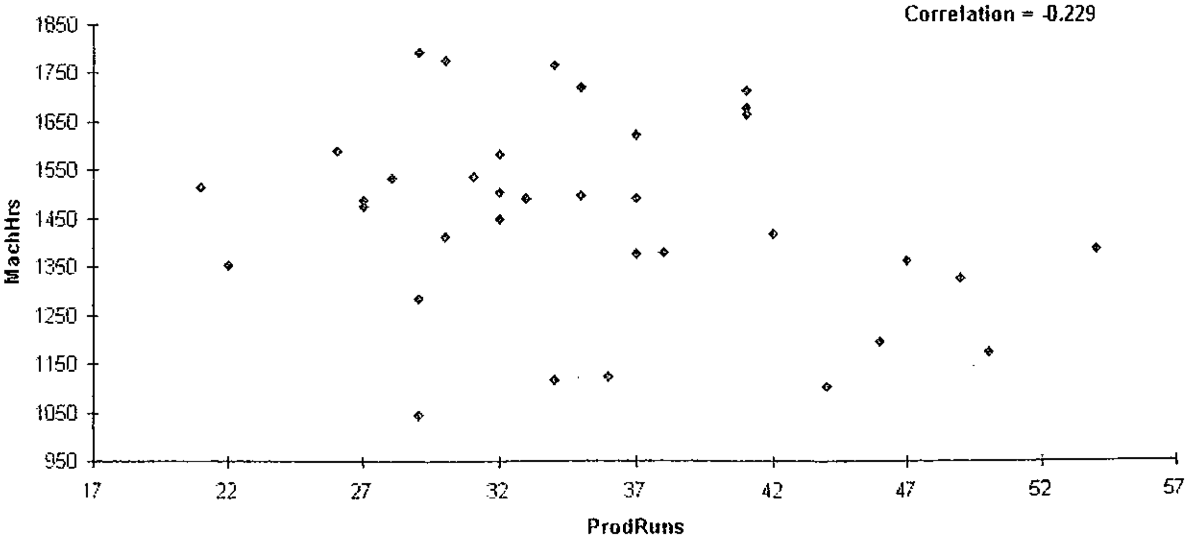


Exhibit 2. Scatterplot of Overhead versus ProdRuns



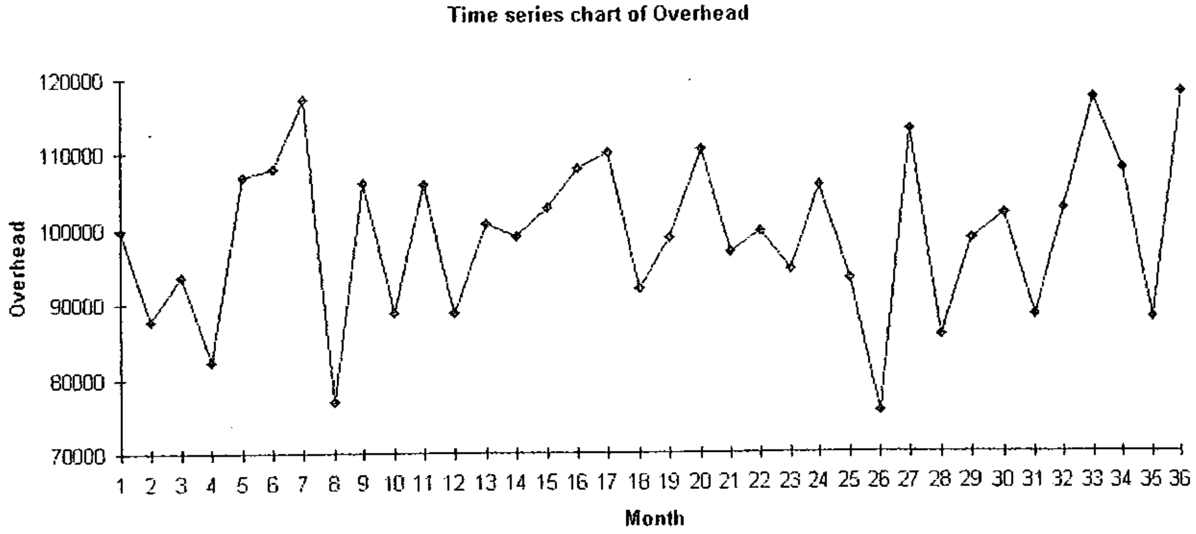
Before moving to the regression analysis, there are two other charts you should consider. First, you ought to check whether there is a relationship between the two explanatory variables, MachHrs and ProdRuns. If the correlation between these variables is high (negative or positive), then you have a phenomenon called *multicollinearity*. This is not necessarily bad, but it complicates the interpretation of the regression equation. Fortunately, as Exhibit 3 indicates, there is virtually no relationship between MachHrs and ProdRuns, so multicollinearity is not a problem for you.

Exhibit 3. Scatterplot of MachHrs and ProdRuns



You should also check the time series nature of your overhead data. For example, if your overhead expenses are trending upward over time, or if there is a seasonal pattern to your expenses, then MachHrs and ProdRuns, by themselves, would probably not be adequate to forecast future values of Overhead. However, as illustrated in Exhibit 4, a time series plot of Overhead indicates no obvious trends or seasonal patterns.

Exhibit 4. Time series plot of Overhead



Regression analysis

The plots in Exhibits 1-4 provide some confidence that regression analysis for Overhead, using MachHrs and ProdRuns as the explanatory variables, will yield useful results. Therefore, I used StatPro's multiple regression procedure to estimate the regression equation. As you may know, the regression output from practically any software package, including StatPro, can be a lit intimidating. For this reason, I will report only the most relevant outputs. (You can see the rest in the Overhead.xls file if you like.) The estimated regression equation is

$$\text{Forecasted Overhead} = 3997 + 43.54\text{MachHrs} + 883.62\text{ProdRuns}$$

Tow important summary measures in any regression analysis are R-square and the standard error of estimate. Their values for this analysis are 93.1% and $4109.

Now let's turn to interpretation. The two most important values in the regression equation are the coefficients of MachHrs and ProdRuns. For each extra machine hour your company uses, the regression equation predicts that an extra $43.54 I overhead will be incurred. Similarly, each extra production run is predicted to add $883.62 in overhead. Of course, these values should be considered approximate only, but they provide a sense of how much extra machine hours and extra production runs add to overhead. (Don't spend too much time trying to interpret the constant term, 3997. Its primary use is to get the forecasts to the correct "level".)

The R-square value indicates that 93.1% of the variation in overhead expenses you observed during the past 36 months can be "explained" by the values of MachHrs and ProdRuns your company used. Alternatively, 6.9% of the variation in overhead has still not been explained. To explain this remaining variation, you would probably need data on one or more *other* relevant variables. However, 93.1% is quite good. In statistical terms, you have a good fit.

For forecasting purposes, the standard error of estimate is even more important than R-square. It indicates the approximate magnitude of forecast errors you can expect when you base your forecasts on the regression equation. This standard error can be interpreted much like a standard deviation. Specifically, there is about a 68% chance that a forecast will be off by no more than one standard error, and there is about a 95% chance that a forecast will be off by no more than two standard errors.

Forecasting

Your forecasting job is now quite straightforward. Suppose, for example, that you expect 1525 machine hours and 45 production runs next month. (These values are in line with your historical data.) Then you simply plug these values into the regression equation to obtain the forecasted overhead:

$$\text{Forecasted overhead} = 3997 + 43.54\,(1525) + 883.62\,(45) = \$101{,}158$$

Given the standard error of estimate of $4109, you can be about 68% confident that this forecast will be off by no more than $4109 on either side, and you can be about 95% confident that it will be off by no more than $8218 on either side. Of course, I'm sure you know better than to take any of there values too literally, but I believe this level of forecasting accuracy should be useful to your company.

One last recommendation I have is to update the analysis as time moves on. As you observe future values of the variables, incorporate them into the data set (and remove old values if you believe they are obsolete), and return the regression analysis. You can do this easily with the same Figure 1 I have attached.

If you have any questions, feel free to call me at any time. You have my number.