On The Use of Indicator Variables in Regression Analysis

By Keith M. Bower, M.S.

Abstract

Frequently, practitioners seek to use categorical data in the course of model building using simple and multiple linear regression analysis. This may involve investigating variables such as location, color, etc. As will be shown, generally speaking, it is incorrect to recode such variables using numeric values to be included in regression analysis. However, it has come to the attention of the author that this procedure has, in fact, been implemented in several quality improvement projects. The aim of this paper is to address this issue by the use of discussion and an example, and to exhibit the correct methodology to be used; namely incorporating the use of indicator (or "dummy") variables, using the statistical software package MINITAB™.

Ordinary Least Squares Regression

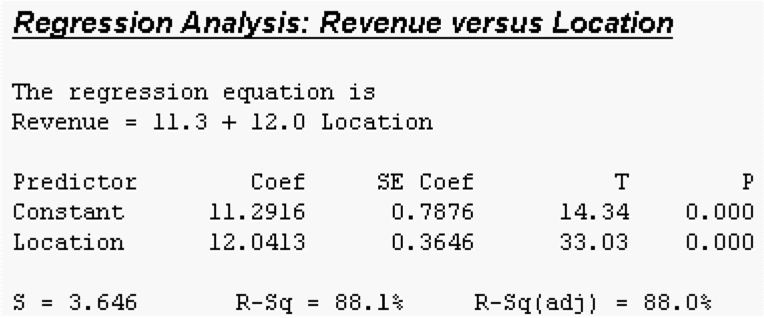
One way in which processes may be modeled is to make use of simple and multiple linear regression analysis, whereby a continuous response variable is explained in terms of various continuous and/or categorical input factors. The method used to fit this relationship is typically by way of minimizing the sum of the squared errors between the observed values and the value that would be fitted under the assumed relationship. This procedure is known as Ordinary Least Squares (OLS) regression, and, as is discussed by Box[1], was developed by Gauss in the late 18th Century.

When categorical data is being used, it is not appropriate to recode such data using numeric values such as 1, 2, 3, etc., as these values will be regarded as continuous data for the analysis. The fitted values, once a model has been developed, will therefore be dependent upon which numerical values are used. Instead, one should make use of indicator variables, which indicate whether or not that factor level is to be included in the model. This is illustrated in the following example.

Example

Consider a hypothetical situation whereby revenue information from sites in the states of Arizona, Florida and Texas have been collected, and we seek to build a model incorporating these locations. If the information were coded as AZ = 1, FL = 2, TX = 3, we would obtain results as shown in Figure 1.

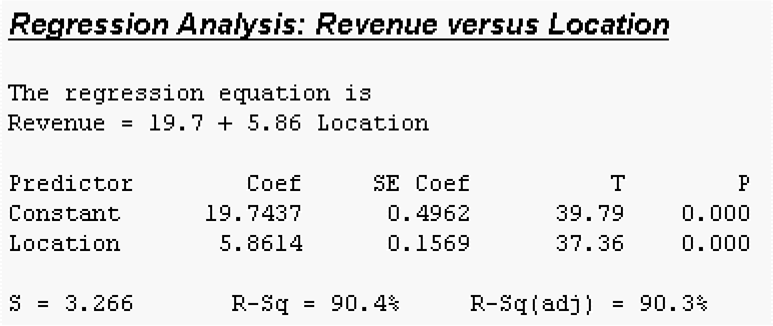
Figure 1



```
Regression Analysis: Revenue versus Location

The regression equation is
Revenue = 11.3 + 12.0 Location

Predictor        Coef     SE Coef          T        P
Constant      11.2916      0.7876      14.34    0.000
Location      12.0413      0.3646      33.03    0.000

S = 3.646       R-Sq = 88.1%     R-Sq(adj) = 88.0%
```

As was discussed previously, however, this is inappropriate for the analysis to be valid. For example, using the coding system as AZ = 1, FL = 2, TX = 5, we would obtain different results, as exhibited in Figure 2.

Figure 2



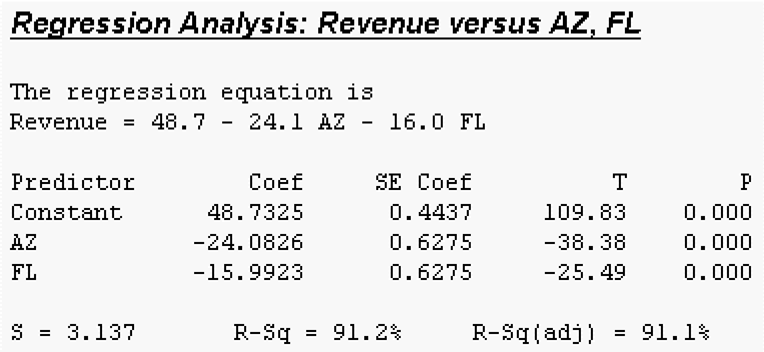
```
Regression Analysis: Revenue versus Location

The regression equation is
Revenue = 19.7 + 5.86 Location

Predictor        Coef     SE Coef          T        P
Constant      19.7437      0.4962      39.79    0.000
Location       5.8614      0.1569      37.36    0.000

S = 3.266       R-Sq = 90.4%     R-Sq(adj) = 90.3%
```

Instead, one would use indicator variables for the location effect to conduct this type of analysis.

It is possible to use the Calc>Make Indicator Variables functionality in MINITAB to create a series of ones and zeros for these three levels of the location factor. Note that MINITAB will generate the series by assessing the column from which the indicator variables will result, and generate the new indicator columns in alphabetical order (unless another value order had been declared.) The first column generated would, therefore, be for AZ, followed by FL, then TX (since A<F<T.)

Importantly, only two indicator variables need to be included in our new model (consider – if we are told that the location is not AZ or FL, then it **has** to be TX for this example.) In general, with r levels of a categorical factor, only r-1 indicator variables need be included in the regression model. With this type of model, results such as those in Figure 3 would therefore be valid.

Figure 3



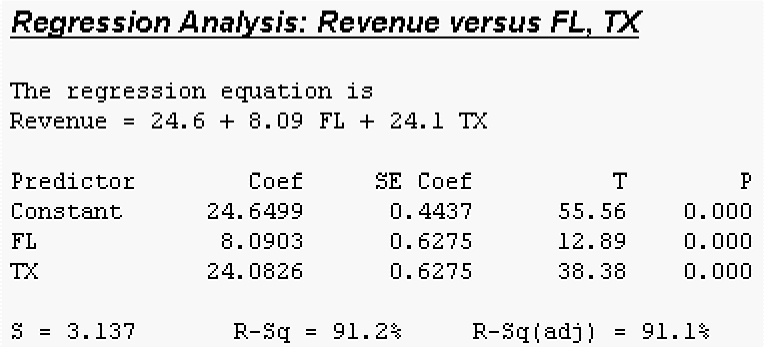
Regression Analysis: Revenue versus AZ, FL

```
The regression equation is
Revenue = 48.7 - 24.1 AZ - 16.0 FL

Predictor        Coef     SE Coef          T        P
Constant      48.7325      0.4437     109.83    0.000
AZ           -24.0826      0.6275     -38.38    0.000
FL           -15.9923      0.6275     -25.49    0.000

S = 3.137       R-Sq = 91.2%     R-Sq(adj) = 91.1%
```

For example, under the model with indicator variables, we would estimate revenue for AZ as roughly 48.7 - 24.1 = 24.6, revenue for FL as 48.7 - 16.0 = 32.7, and revenue for TX as 48.7 (note that we multiply the indicator variables by 1 or 0 as required.)

Importantly, it is irrelevant which indicator variable is left out of the model (we could choose AZ, FL or TX to exclude.) To illustrate, as shown in Figure 3, with FL and TX only being included in the model, one would obtain identical results for the predicted revenues.

Figure 4



Regression Analysis: Revenue versus FL, TX

```
The regression equation is
Revenue = 24.6 + 8.09 FL + 24.1 TX

Predictor        Coef     SE Coef          T        P
Constant      24.6499      0.4437      55.56    0.000
FL             8.0903      0.6275      12.89    0.000
TX            24.0826      0.6275      38.38    0.000

S = 3.137       R-Sq = 91.2%     R-Sq(adj) = 91.1%
```

Conclusion

Though the use of numerical levels is acceptable when using the Analysis of Variance (ANOVA) procedure (e.g. see Bower[2]), such usage is inappropriate when making use of simple and multiple regression analysis with more than two levels for a factor(s). The use of indicator variables is a valid way of incorporating categorical variables in regression analysis.

Keith M. Bower has an M.S. in Quality Management and Productivity from the University of Iowa, and is a Technical Training Specialist with Minitab, Inc.

Reference:

1. Box, G.E.P. (1984), "The Importance of Practice in the Development of Statistics" *Technometrics*, 26(1)
2. Bower, K.M. (February 2000), "Analysis of Variance (ANOVA) Using MINITAB" Sc*ientific Computing & Instrumentation*.