Multiple Linear Regression in Excel

Dr. Richard Charnigo
Professor of Statistics and Biostatistics
RJCharn2@aol.com
08 and 11 January 2016

Motivation

Suppose we want to understand the relationship between a dependent variable (also called: outcome variable, response variable) and several independent variables (also called: explanatory variables, predictor variables).

As a running example throughout this workshop, we will consider (fictional) data on salary, age, experience, education, and urban/rural status for 100 employees. We will regard salary as a dependent variable and the other four attributes as independent variables from which salary is to be predicted. (See Sheet 1 of the accompanying Excel file.)

Motivation

Although we might calculate a correlation between (or perform a simple linear regression involving) salary and each of the four independent variables (and we do so on Sheet 1, though this may not be strictly appropriate with urban/rural status), there are at least four reasons why such calculations might be insufficient:

1. *Desire for unambiguous prediction of outcome, utilizing all independent variables*. Instead of trying to reconcile four disparate predictions based on the four independent variables individually (using the trend lines displayed in the graphs of Sheet 1), we might prefer to have a single prediction combining information from all four independent variables.

Motivation

2. *Desire for quantification of how well an outcome can be predicted from multiple independent variables collectively*. By itself age explains 56% of the variation in salary, and by itself experience explains 55% of the variation in salary. Clearly the two combined do not explain 111% of the variation in salary, so what we desire cannot be obtained by correlation (or simple linear regression).

3. *Desire to exhibit adjusted rather than unadjusted relationships*. Age is strongly positively related to salary. But age is also positively related to experience, which itself is positively related to salary. If we adjust for experience, does age still matter ? Put differently, is the relationship between age and salary due solely to their mutual relationship with experience ?

Motivation / Formulation

4. *Desire to test whether the association of one independent variable with the outcome depends on the level of another independent variable*. For example, is education level more strongly related to salary in an urban setting or in a rural one ?

To address these questions, we may fit a *multiple linear regression model*, which is expressed symbolically as follows:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k + \text{error}$$

Here, Y is the outcome, X_1 through X_k are predictors, and the error is a random quantity satisfying certain assumptions.

Formulation

More specifically, the error is assumed to follow a normal distribution with mean 0 and standard deviation σ. The standard deviation is assumed to be fixed, unrelated to X_1 through X_k. Errors for different subjects are assumed to be independent.

We interpret b_1 as the amount by which the outcome is predicted (but not guaranteed !) to change when X_1 increases by one unit, with X_2 through X_k fixed. If a one unit change is not meaningful, we may also note that $c\, b_1$ is the amount by which the outcome is predicted to change when X_1 increases by c units, again with X_2 through X_k fixed.

Formulation

If X_1 is a logarithm of some underlying quantity, then $LN(2)\ b_1$ is the amount by which the outcome is predicted to change when the underlying quantity doubles, with X_2 through X_k fixed.

If Y is a logarithm of some underlying quantity, then the predicted value of the underlying quantity is *multiplied* by $\exp(c\ b_1)$ when X_1 increases by c units, with X_2 through X_k fixed.

If both Y and X_1 are logarithms of underlying quantities, then the predicted value of the quantity underlying Y is multiplied by $\exp(\ LN(2)\ b_1\ ) = 2^{b1}$ when the quantity underlying X_1 is doubled, with X_2 through X_k fixed.

Formulation

Although $b_0, b_1, \ldots, b_k$ are unknown, we may estimate them using the principle of least squares. The estimates, which we may call $b_0^*, b_1^*, \ldots, b_k^*$, satisfy the following inequality:

$$\sum (Y - b_0^* - b_1^* X_1 - \ldots - b_k^* X_k)^2 \leq \sum (Y - a_0 - a_1 X_1 - \ldots - a_k X_k)^2,$$

where $a_0, a_1, \ldots, a_k$ are any numbers.

The least squares estimates are, in a very specific mathematical sense (on which I shall unfortunately not be able to elaborate here), values for $b_0, b_1, \ldots, b_k$ which would have been most likely to generate data similar to what we actually observed.

Formulation

Alternatively and perhaps more intuitively, we may regard

$b_o^* + b_1^* X_1 + ... + b_k^* X_k$ as the best available "prediction" for Y.

Thus, $\sum (Y - b_o^* - b_1^* X_1 - ... - b_k^* X_k)^2$ is called a *residual sum of squares*. I also note that $\sum (b_o^* + b_1^* X_1 - ... + b_k^* X_k - \text{mean}(Y))^2$ is called a *regression sum of squares* and that $\sum (Y - \text{mean}(Y))^2$ is called a *total sum of squares*.

Exploring the data

Graphs of the type shown on Sheet 1, though not sufficient in and of themselves, are useful for preliminary assessment of whether a proposed multiple linear regression model makes sense.

In essence, we may ask whether we believe that the expected value of Y should be expressed as $b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k$. While this appears a bit complicated because of the multiple independent variables, this is actually about the simplest possibility we might consider. In particular, there are no nonlinear terms. (If you have had calculus, the mathematical rationale for the aforementioned expression is that it represents a first-order Taylor approximation.)

Exploring the data

The graphs on Sheet 1 actually look pretty decent. However, we notice that the trend lines for the graphs involving age and education do not quite fit the patterns of data points. Although not markedly so, these patterns appear nonlinear.

Another concern, which is not evident from the graphs nor from simple diagnostic changes on individual variables (like examining minimum and maximum), is with the first subject, who is alleged to have 31 years of experience at age 39. Since I made up the (fictional) data, I know that the 31 "should" be 13.

A first attempt at modeling

These two issues noted, let us still proceed with a multiple linear regression analysis and see what happens. Using the Data Analysis add-in to Excel, one can obtain the results on Sheet 1R; I will demonstrate now. I have annotated these results in detail, so let's discuss them…

In summary, about 63% of variability in salary is explained by the four independent variables, and a typical random error is about $16,000. At fixed levels of education and experience, and in a fixed geographic setting, each year of age adds about $800 to the predicted salary. This is statistically significant (p-value = 0.020). Experience and education are also statistically significant, but geographic setting is not.

A second attempt at modeling

On Sheet 2 I have corrected the 31 to a 13 and used Excel to display exponential rather than linear trend lines in the graphs. (This is essentially equivalent to asserting linear relationships with Y re-defined to be the natural logarithm of salary rather than salary itself; that is why I have created a new column "LogSalary".)

The results appear on Sheet 2R. About 65% of variability in log salary is explained by the four independent variables, and a typical random error is about 20%-25% of the expected salary. The 65% can't be directly compared to the 63% obtained earlier; however, when I express predictions on the original scale and calculate residuals on the original scale, I obtain that 64% of variability in salary (not log salary) is explained.

A third attempt at modeling

On Sheet 3 I have defined a new variable "UrbanEdu" as the product of the urban and education variables. I have also created scatterplots depicting the relationships between salary and each of the three continuous variables, in rural and urban settings.

Note that the curve relating salary to education appears steeper in the rural setting than in the urban setting, suggesting an *interaction*: education appears to be more determinative of salary in the rural setting.

A third attempt at modeling

If I include "UrbanEdu" in the regression model along with the four independent variables, I obtain the results shown on Sheet3R. Now about 67% of variability in log salary is explained.

Note that the coefficient estimates for education, urban, and UrbanEdu are statistically significant with respective p-values 0.001, 0.024, and 0.036. What does this represent ?

We must consider how education can increase by one unit without any other variable increasing; clearly this is possible only when urban = 0 (since otherwise UrbanEdu would change with education)…

A third attempt at modeling

Hence, the p-value of 0.001 indicates that education is a significant predictor of (log) salary in the rural setting. This p-value does not say anything about whether education is (or is not) a significant predictor of salary in the urban setting.

The p-value of 0.024 indicates that urban is a significant predictor of salary when there is no education. This is meaningless, however, because everyone has at least 12 years of education.

The p-value of 0.036 indicates a significant interaction between education and urban. More specifically…

A third attempt at modeling

…the estimated change in average log salary in moving from a rural setting to an urban setting is $0.887 - 0.049\,X$, where X is the number of years of education. This corresponds to predicted salary being multiplied by $\exp(0.887 - 0.049\,X) = 2.43\,X^{-0.049}$ in moving from a Rural setting to an Urban setting.

A "tradeoff" occurs around $X \approx 18$. For persons with less education, salary tends to be higher in an urban setting; for persons with more education, salary tends to be higher in a rural setting.

A third attempt at modeling

Moreover, each year of education multiplies the predicted salary by $\exp(0.0599 - 0.0493\,Z)$, where Z denotes urban status. Since Z can only equal 0 or 1, this is either $\exp(0.0599) \approx 1.06$ or $\exp(0.0106) \approx 1.01$.

We may ask, is the 0.0106 significantly different from 0 ? Equivalently, is the 1.01 significantly different from 1 ? The Excel output doesn't seem to give the answer to that question, but we could obtain it by appropriately defining a new variable "RuralEdu" and using it instead of "UrbanEdu".

Potential pitfalls and limitations

A lot of things can go "wrong" with multiple linear regression. Some of them may be readily addressed; others may not. Consulting with a statistician, when in doubt, is a good idea.

1. Relationships may be nonlinear. If the nonlinearity is not too severe, you may be able to ignore it or accommodate it through a simple transformation (e.g., logarithmic). However, if the nonlinearity is pronounced, you may need a polynomial, nonparametric, or semiparametric model. While a polynomial model may seem appealing because of its availability in Excel, the resulting coefficient estimates are difficult to interpret.

Potential pitfalls and limitations

2. Assumptions regarding the error terms may not be satisfied:

   a. If independence fails (as may occur with time series data, or repeated measurements on the same persons), you may need a more general statistical model.

   b. If normality fails (and a transformation cannot fix the problem), a variant called “robust regression” – a modification of least squares – can be used.

   c. If fixed variance fails, another variant called “weighted least squares” can be used.

Potential pitfalls and limitations

3. We implicitly assume that we have the "correct" variables. This has both substantive and mathematical components:

   a. Substantively (i.e., in terms of the underlying subject matter), we may not know which variables are "correct". Even if we do, they may not have been measured (especially if we work with secondary data) or may not be measurable.

   b. Moreover, even if we have the correct variables, we may not think to look for interactions. Because the number of possible interactions generally exceeds the number of predictors (two continuous predictors can interact with each other !), checking for each possible interaction may be cumbersome.

Potential pitfalls and limitations

c. Mathematically, we cannot effectively estimate a large number of parameters from a small number of observations. So, even if we knew that 20 predictor variables were potentially relevant, trying to include all of them in a model based on a sample of size 50 would be imprudent. Collinearity may also be an issue.

d. Due to both substantive and mathematical issues, being able to select variables for a regression model is essential. Note that we *cannot* use R^2 for this purpose, if the *number* of variables has not been fixed a priori. A common variable selection technique, easy to implement in Excel though probably not optimal, is backward elimination.

Practice exercises

Here are some practice exercises, if you wish to go beyond replicating my analyses:

1. Assuming that urban/rural setting interacts with age (even if the p-value is not significant) in a regression model for log salary, estimate the amount by which predicted salary is multiplied for each year of age in a rural setting. Is this significantly different from 1 ?

2. Continuing, estimate the amount by which predicted salary is multiplied for 10 years of age in an urban setting. Is this significantly different from 1 ?

Practice exercises

3. Continuing, estimate the amount by which predicted salary is multiplied when a 45-year-old moves from a rural setting to an urban one ? Is this significantly different from 1 ? (Hint: Define a new variable which is the product of {Age – 45} × Urban.)

4. What part of variability in log salary is explained by a model in which age interacts with urban/rural status ? Following work similar to that on Sheet 2R, what part of variability in salary (not log salary) is explained by such a model ?

5. How do your answers to the preceding questions change, if we model salary directly (rather than log salary) ?