Bivariate Linear Regression

See General Linear Model for a More Rigorous Presentation, including proofs of estimates and variance of estimates.

The multiple regression model is used when we have a set of continuously varying predictors which are used to predict a continuously varying criterion. For example we may use various measures of personality (e.g., impulsiveness, sensation seeking both measured on 100 point scales) to predict severity of alcohol abuse (amount of alcohol per week) for individuals. The bivariate regression model is just a special case involving only two predictors. But this is complex enough to illustrate all the basic issues of multiple regression.

First we define the bivariate model and the variables that enter the model.

1. *Definitions*

Bivariate Regression Model:

$$y_i' = b_0 + b_1X_{1i} + b_2X_{2i}$$

Y_i = observed criterion score for row i (eg. alcohol intake of individual i in row i of a table of data)

$X1_i$ = value of first predictor variable for row i (eg. impulsiveness score for person i)

$X2_i$ = value of second predictor variable for row i (e.g. sensation seeking score for person i)

Y_i' = predicted criterion score for row i (generated by the model)

e_i =(Y - Y')= residual score for row i

b_1 = the regression coefficient representing the change in *y* produced by each unit change in *X*1. In other words, this represents the effect of *X*1 on y.

b_2 = the regression coefficient representing the change in *y* produced by each unit change in *X*2. In other words, this represents the effect of *X*2 on y.

How we determine these two coefficients is discussed later. Next we review the sample statistics that enter into the formulas used to compute the coefficients, which are then used to computer the predictions of the bivariate regression model. (click here for a review of these statistics)

2. ***Sample Statistics:***

Notation

M_{x1} = sample mean of X1

M_{x2} = sample mean of X2

M_y = sample mean of Y

s_{x1}^2 = sample variance of X1

s_{x2}^2 = sample variance of X2

s_y^2 = sample variance of Y

$r_{x1,y}$ = correlation between X1 and Y

$r_{x2,y}$ = correlation between X2 and Y

$r_{x1,x2}$ = correlation between X1 and X2

TSS = total sum of squared deviations scores = $\Sigma\,(Y_i\text{-}M_y)^2$

SSE = sum of squared error scores = $\Sigma\, e_i^2$

SSR = TSS-SSE = sum of squared predictions

R-square

R^2 = SSR / TSS

This is an important measure of the fit of the model to the data. It is interpreted as the proportion of variance in the criterion predicted by the model.

The next statistic is a sample estimate of the error variance for the model. This is important for computing standard errors and for performing statistical tests.

MSE = SSE / (N-3) = mean squared error
df = N-3

df refers to the number of free residuals. If we have N rows in the data table, then we have N residuals, but not are all free. Three constraints on are placed on the residuals due to three model parameters estimated from the data. The df is important for performing f-tests. In general, df = N – number of model parameters.

Initially we do not know what the coefficients , b_0, b_1, and b_2 should be used to generate predictions. So we need to find the coefficients that produce the best fit to the data. Fit is measured by the sum of squared residuals. Thus we need to find the coefficients that minimize the sum of squared error.

3. ***Least squares estimates of regression coefficients***

(see reference for a general proof)

Example of a SSE surface as a function of b1 and b2

For the bivariate model

$$y_i' = b_0 + b_1 X1_i + b_2 X2_i$$

the formulas for the coefficients that minimize SSE are

$$b_1 = \frac{s_y}{s_{x1}} \cdot \frac{r_{x1,y} - r_{x1,x2} \cdot r_{x2,y}}{1 - r_{x1,x2}{}^2}$$

$$b_2 = \frac{s_y}{s_{x2}} \cdot \frac{r_{x2,y} - r_{x1,x2} \cdot r_{x1,y}}{1 - r_{x1,x2}{}^2}$$

$$s_{b1}{}^2 = \frac{MSE}{N-1} \cdot \frac{1}{s_{x1}{}^2} \cdot \frac{1}{1 - r_{x1,x2}{}^2}$$

$$s_{b2}{}^2 = \frac{MSE}{N-1} \cdot \frac{1}{s_{x2}{}^2} \cdot \frac{1}{1 - r_{x1,x2}{}^2}$$

It is important to study these formulas, not because you will use them in your research to computer numerical answers, but to understand how to interpret these estimates. Note how the standard deviations and the correlations influence each estimate.

Compare these bivariate estimates to the estimate obtained from the simple linear regression model: $y' = b_0 + b_1 \cdot X1_i$, which is

$b_1 = r_{x1,y} (s_y / s_{x1})$

Note that sign and magnitude of $r_{x1,x2}$ can change the sign of the regression coefficient for b_1 when comparing the simple vs. bivariate model.

These estimates of the coefficients are usually based on a small sample of data, and so they are estimated with error. We also need to determine how precisely we have estimated the coefficients, and this is determined from the standard error of the regression coefficients. Intuitively, the standard error indicates how much we expect our estimate to deviate from the true population coefficient (which can never be known exactly). The formulas for the variance of each coefficient are given above as well.

To get the std errors we take the square roots of the above:

$$s_{b1} = \text{sqrt}(s_{b1}^2) = \text{standard error of } b_1$$

$$s_{b2} = \text{sqrt}(s_{b2}^2) = \text{standard error of } b_2$$

We use the standard errors to obtain confidence interval estimates of the coefficients.

4. **95% Confidence Interval** for one of the coefficients of the bivariate model:

Lower bound of the est for b_1 :

$LB = b_1 - (s_{b1})(t_c)$

Upper bound of the est for b_1 :

$UB = b_1 + (s_{b1})(t_c)$

Thus the interval is [LB, UB]

where t_c is the table t value obtained from row df = N-3 and using the column for the 5% error (two tail).

Using a classic statistics interpretation, we say that there is a .95 probability that this sample interval covers the true population coefficient. Using a Bayesian interpretation, we say that there is a .95 probability that the true coefficient falls within this interval.

We can perform a statistical test of a coefficient from the bivariate model as follows (Note: E(X) = population mean of X):

H_0: $E(b_1) = \beta_1 = 0$.

Reject H_0 if the confidence interval [LB,UB] does not cover zero.

Alternatively we can do the same test as follows:

T-test for a parameter.

$H_0 : E(b_1) = 0$

$T = (b_1 / s_{b1})$

reject reject H_0 if $p < .05$

What is the p-value? Suppose $b_1 > 0$, then

$p =$ 2 · (Probability of obtaining a T statistic greater than the one you observed given that the null hypothesis is true). There are no other interpretations except wrong ones.

Next we consider comparing various models. The most complex model we consider is the bivariate model, and so it is the complete model. This produces the smallest sum of squared errors. A restricted model is formed by restricting $b_2 = 0$ reducing it to a simple linear regression model $y' = b_0 + b_1X1$. The simplest model is formed by restricting $b_1 = b_2 = 0$ to form the null model $y' = b_0$ which has the largest sum of squared error (which is called TSS).

There are various methods for comparing models. One is based on R-square. The complete model will always have the highest R-square, but we can evaluate how much improvement we gain going from a simple to a complex model in terms of R-square. Another method is to use an f-test to statistically test the difference between models. R-square is more meaningful for large samples, and the f-test is only meaningful for small samples.)Almost anything will be significant with a large sample.)

5. *Model Comparisons*

First we list 4 models and their sum of squared errors:

Null model: $Y' = b_o$

which produces a sum of squared error = TSS

(this just predicts the mean)

X1 alone model: $Y' = b_0 + b_1X1$

which produces a sum of squared error = SSE(X1)

(this of course is a simple linear regression model)

X2 alone model: $Y' = b_0 + b_2X2$

which produces a sum of squared error = SSE(X2)

(this of course is a simple linear model)

X1 and X2 model: $Y' = b_0 + b_1X1 + b_2X2$

which produces a sum of squared error =SSE(X1,X2)

(this of course is the bivariate model)

Now we do compare various models. The first step in any comparison is to compute the difference in sum of squared errors between two models, which is denoted SSR = SSE1 – SSE2. The f – test is based on F* = MSR/MSE, where MSR = SSR/q, and q is the difference in number of parameters used by each model.

1. Suppose we wish to compare the bivariate model to the null model and test H0: $E[b_1] = E[b_2] = 0$

SSR(X1,X2) = TSS - SSE(X1,X2)

MSR = SSR(X1,X2) / 2

MSE = SSE(X1,X2)/(N-3)

F* = MSR/MSE , reject if F* > f(2,N-3)

R^2 = SSR(X1,X2)/TSS

2. Suppose we wish to compare the bivariate model to the simple linear model using only X2, H0: E[b_1] = 0 (unique effect of X1)

SSR(X1|X2) = SSE(X2) - SSE(X1,X2)

MSR = SSR(X1|X2) / 1

MSE = SSE(X1,X2) / (N-3)

F* = MSR/MSE, reject if F* > f(1,N-3)

R^2 change = SSR(X1|X2)/TSS

3. Suppose we wish to compare the bivariate model to the simpler linear model using only X1, H0: E[b_2] = 0 (unique effect of X2)

SSR(X2|X1) = SSE(X1) - SSE(X1,X2)

MSR = SSR(X2|X1) / 1

MSE = SSE(X1,X2) / (N-3)

F* = MSR/MSE, reject if F* > f(1,N-3)

R^2 change = SSR(X2|X1)/TSS

4. Suppose we wish to compare the simple linear model using only X1 to the null model, H0: $E[b_1] = 0$ (effect of X1 ignoring X2)

SSR(X1) = TSS - SSE(X1)

MSR = SSR(X1) / 1

MSE = SSE(X1,X2) / (N-3)

F* = MSR/MSE, reject if F* > f(1,N-3)

R^2 = SSR(X1)/TSS

Note that this comparison tests the b_1 coefficient in a different way than comparison 2. This ignores the effect of X2, whereas comparison 2 examines the contribution of X1 that cannot be explained by X2. The results are different whenever $r_{x1,x2} \neq 0$. (Look back at the bivariate vs simple linear regression coefficient formulas).

Reference for an example multiple regression application

Ganzach, Y. (1995) Nonlinear models of clinical judgment: Meehl's data revisited.

Psychological Bulletin, 118, 422-429.