

## An Introduction to Linear Regression Analysis (An Example)

George H Olson, Ph. D.  
Doctoral Program in Educational Leadership  
Appalachian State University  
(Fall 2010)

(Note: prior to reading this document, you should review the companion document, [Partitioning of Sums of Squares in Simple Linear Regression](#), which can be found in the [Instructional Stuff for Statistics](#) chapter in the website for this course.)

In this example, I used the data found in the [Appendix](#) (the data have been taken from Pedhauer (1997; p. 98). To make this example meaningful, let's let the variable, *Score*, be taken as a score a difficult 10-item, 4-option, multiple choice achievement test; let X1 be a measure of hours of homework completed the week prior to the test; and X2, a measure, on a 10-point scale, indicating how much students value high grades.

I used SPSS to conduct the analysis, but other statistical software packages would produce the same or similar output. (A much more detailed discussion of regression analysis can be found [HERE](#).)

*Descriptive Statistics.* My first step was to obtain the descriptive statistics (Ns, Means, and Standard Deviations for each variable in the analysis). These are shown in Table 1, where a minimal, sufficient set of descriptive statistics is given. If we wanted, say, the standard errors of the mean, we could compute these easily. For example,

$$SEM_{\text{score}} = \text{STD}_{\text{score}} / \sqrt{(N_{\text{score}} - 1)} = 2.763/4.359 = .634.$$

**Table 1**  
**Descriptive Statistics**

	Mean	Std. Deviation	N
Score	5.70	2.736	20
X1	4.35	2.323	20
X2	5.50	1.670	20

On the 10-item test, a chance score would be 2.5 items correct, and the standard deviation of a chance score is  $1.37^1$ . From Table 1 it is apparent that the mean score on the test was nearly

---

<sup>1</sup> These values are computed using the binomial distribution with  $M_{\text{chance}} = k(.25)$ , (where  $k$  = number of items, and  $.25$  is  $p$ , the probability of a correct answer due to chance); and  $SD_{\text{chance}} = \sqrt{np(1-p)}$ .

two standard deviations above a chance score. On average, the group taking the test spent a little over four hours, the previous week, doing homework. Furthermore, the desirability of high grades was not strikingly high.

Next, I had SPSS compute the correlations among all variables in the analysis. This yielded Table 2, where the first three rows in the body of the table (i.e, the values enclosed in the larger box) give the *correlation matrix* for the three variables. In the matrix, the set of lightly shaded entries is called the *diagonal* of the matrix. In a correlation matrix, these are always 1.0, since each gives the correlation of a variable with itself. I should note, also, that a correlation matrix is a *symmetric* matrix: the upper triangular half of the matrix is a mirror image of the lower triangular half (e.g., the correlation of *Score* with X1 is .771—in the first row to the right of the shaded 1.0—which is equal to the correlation of X1 with *Score* (the .771 just below the shaded 1.0 in the first column of the matrix).

**Table 2**  
**Correlations**

		Score	X1	X2
Pearson Correlation	Score	1.000	.771	.657
	X1	.771	1.000	.522
	X2	.657	.522	1.000
Sig. (1-tailed)	Score	.	.000	.001
	X1	.000	.	.009
	X2	.001	.009	.
N	Score	20	20	20
	X1	20	20	20
	X2	20	20	20

The second three rows in Table 2 give the (one-tailed) statistical significance of the corresponding correlations in the correlation matrix. All correlations are statistically significant. For instance, the significance of the correlation between *Score* and X1 is given as .000. This does not mean that the probability of a correlation of .771 due to sampling error is zero. It just means that SPSS rounded the actual probability to three decimal places. Therefore, the actual probability is less than .0005. The third set of three rows in Table 2 give the number of cases involved in each of the correlations.

From the table, we learn that the correlations between all pairs of variables are both statistically significant and appreciably large. Having strong correlations between the dependent variable (*Score*) and each of the independent variables (X1 and X2) is desirable because it means that the dependent variable shares variance with each of the independent variables.

Correlations, especially large correlations, between the independent variables, on the other hand, are not desirable. In this case the *covariance* between the dependent variable and each of the independent variables is not unique (in an ideal situation, each independent variable would have a unique and independent association with the dependent variable). I will address this later in this presentation.

*Regression Analysis.* The regression analysis is summarize in the next several tables. Table 3 gives a general summary of the analysis. The R is the *multiple correlation*: the correlation between the dependent variable (Score) and the *weighted linear composite* of the independent variables, i.e,  $r_{\text{score}, \hat{Y}}$  where  $(\hat{Y} = b_0 + b_1X_1 + b_2X_2)$ . The multiple R is interpreted in the same way as a simple *zero-order* correlation between any two variables.

The next value of interest is R-Square ( $R^2$ ). This is an important statistic for it gives the percent of variance in the dependent variable (Score) *explained* or *accounted for* by the independent variables. Another name for  $R^2$  is the *coefficient of determination*, a term used mainly when regression analysis is used for prediction. The  $R^2$  of .683 tells us that 68% of the variance in Score is associated (we would say *explained, accounted for, or predicted*) by the independent variables, X1 and X2.

The next statistic in Table 3 is the *Adjusted  $R^2$* , a statistic that is not used often. It is an adjustment for the number of independent variables *in the model*.

Finally, the last statistic in Table 3 is the *Standard Error of Estimate* (SEE). This is the standard deviation of the residuals,  $e$ , ( $= y - \hat{y}$ ) and, as such, gives a measure of the accuracy of the model to predict the Scores (a more detailed, yet tractable, description of the SEE can be found in [Online Statistics](#)).

**Table 3**  
**Model Summary<sup>(b)</sup>**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.826 <sup>(a)</sup>	.683	.646	1.628

a Predictors: (Constant), X2, X1

b Dependent Variable: Score

The next table, Table 4, is an Analysis of Variance table for the regression analysis. Most of the statistics given in the table should already be familiar. The Sums of Squares terms are  $SS_{\text{reg}}$  and  $SS_{\text{res}}$ , which are used for computing  $MS_{\text{reg}}$  and  $MS_{\text{res}}$  (by dividing each SS term by its corresponding degrees of freedom). The  $F$  statistic is then computed by dividing  $MS_{\text{reg}}$  by  $MS_{\text{res}}$ , yielding 18.319 which is significant at  $p < .0005$ . Therefore, we conclude that we do have a linear model that predicts (or accounts for) variance in the dependent variable.

**Table 4**  
ANOVA<sup>(b)</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	97.132	2	48.566	18.319	.000 <sup>(a)</sup>
	Residual	45.068	17	2.651		
	Total	142.200	19			

a Predictors: (Constant), X2, X1

b Dependent Variable: Score

In Table 5 the regression coefficients statistics of the model are given. Here, since we are testing only one model (we could have tested more) only one model is given. The variables in the model are given first (the constant is equal to  $\bar{Y} - b_1X_1 - b_2X_2$ ). Then the computed values of the *unstandardized* regression coefficients, the  $b_i$ 's, (B in the table) are given, along with their standard errors. The Std. Error's are used to test the null hypotheses that the unstandardized regression coefficients equals zero. This is done, for each  $b$ , using a  $t$  test with one degree of freedom:

$$t = \frac{b}{StdErr}$$

For X1, we have  $t_{(1)} = 3.676$ ;  $p = .002$ . We conclude that X1 accounts for a statistically significant percent of the variance in Scores ( $Y$ ).

The standard errors are used, also, to construct confidence intervals around the  $b$ s. If you've read the section on Confidence Intervals and Significance Tests in Hyperstat then you may recall that the 95% confidence interval around a population regression coefficient ( $\beta$ ) is given by,

$$CI95 = \{b \pm t_{crit} (StdErr_b)\}.$$

Where  $t_{crit}$  is the tabled value of  $t$  with N-1 degrees of freedom at the .05 level.

Hence, the 95% confidence interval for  $b_1$  is:

$$\begin{aligned} CI95: \{.693 - (2.1)(.189) \leq \beta \leq .693 + (2.1)(.189)\} \\ = .296 \leq \beta \leq 1.090, \text{ which DOES NOT include zero.} \end{aligned}$$

The unstandardized coefficients, the  $b$ 's, can be used to construct the actual, sample-estimated, regression equation:

$$\hat{Y}_i = -.467 + .693(X_{i1}) + .572(X_{i2}).$$

Hence, for the first individual in the sample, the estimated (or predicted) score is:

$$1.948 = -.461 + .693(1) + .572(3),$$

and this individual's *residual* score ( $e_i$ ) is

$$(Y - \hat{Y}) = 2 - 1.948 = .052.$$

The *standardized* coefficients (the Betas) are used to make inferences about the relative importance (or strength) of the independent variables in predicting the dependent variable. Hence, from the table we see that X1 has a stronger, independent, association with  $Y$  since its Beta coefficient is larger.

**Table 5**  
**Coefficients(a)**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-.461	1.285		-.359	.724		
	X1	.693	.189	.589	3.676	.002	.727	1.375
	X2	.572	.262	.349	2.181	.044	.727	1.375

a Dependent Variable: Score

The collinearity statistics are a more advanced topic and can be dealt with, here, only briefly. *Tolerance* gives an indication of the percent of variance in an independent variable that cannot be accounted for by the other predictors. Very small values (e.g., values less than .1) indicate that a predictor is redundant (i.e., that it carries about the same predictive information as other independent variables in the model.) The VIF stands for *variance inflation factor* and gives a measure of the extent to which predictive information in a particular independent variable is already contained in the other independent variables. As such, it is a measure of how much a regression coefficient is “inflated” by including other, correlated independent variables in the model. Independent variables with VIFs greater than 10 should be investigated further. A VIF of 1.0 means that there is no inflation.

The statistics shown in Table 5 suggest that we do not have a problem with collinearity in the model. The tolerance for X1, for example, tells us that about 70% of the variance in X1 is NOT predicted by X2 (i.e., is not strongly associated with X2). Furthermore, X1's VIF is only 1.375. This tells us that the coefficient for X1,  $b_1$ , is inflated by a factor of, about, only 1.4.

# APPENDIX

## Example Data set

Student	Score	X1	X2
1	2	1	3
2	4	2	5
3	4	1	3
4	1	1	4
5	5	3	6
6	4	4	5
7	4	5	6
8	9	5	7
9	7	7	8
10	8	6	4
11	5	4	3
12	2	3	4
13	8	6	6
14	6	6	7
15	10	8	7
16	9	9	6
17	3	2	6
18	6	6	5
19	7	4	6
20	10	4	9