Multiple Linear Regression: Two Quantitative IVs

The extension of simple linear regression to regression with two quantitative variables is straight-forward and requires the learning of only a few new concepts. As with simple regression, the role of multiple regression is twofold: explanation and prediction. The discussion of these two topics presented earlier continue to hold here.

The Regression Equation

Suppose a researcher is interested in determining whether academic achievement is related to students' time spent studying and their academic ability. Hypothetical data for these variables are presented in Table 1. In the corresponding regression equation for this model, achievement is denoted Y, time spent studying X_1, and academic ability X_2. The population regression model is

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon_i, \quad (1)$$

where

Y_i signifies the i^{th} student's achievement score;

β_1 is the population partial regression coefficient expressing the relationship between X_1 and Y, controlling for X_2;

β_2 is the population partial regression coefficient expressing the relationship between X_2 and Y, controlling for X_1;

β_0 is the population intercept for the equation; and
ε_i is, supposedly, a random error.

Note the change in the interpretation of the regression coefficients. Now each β_i represents the partial effect of X_i on Y, controlling (or partially out) the effects of the other Xs. What is a partial effect? Often researchers wish to investigate models in which more than one IV can be used to explain Y. When this occurs, typically the Xs will be inter-related; that is, the Xs will be correlated. With the current example, both ability and time spent studying will likely be related to achievement. The question of interest is whether, for instance, time spent studying will affect achievement once ability is taken into account—i.e., controlled or partial out. Thus, multiple regression allows one to examine the effects of a given X upon Y while simultaneously taking into account the effects of other Xs. And this ability to partial-out the effects of Xs is the strength of multiple regression.

The multiple regression equation can easily be extended to any number of Xs. For example, one may want to model achievement using not only time spent studying and academic ability, but also learning style, prior exposure to the topic, parental support, instructional strategy used, etc. The regression equation is simply extended with additional regression coefficients and their respective Xs. For example,

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + ... + \beta_k X_k + \varepsilon_i, \quad (1)$$

where

β_i indicates the k^{th} coefficient for the k^{th} independent variable, X_k.

To keep things simple, the discussion of multiple regression will focus upon the two independent variable situation. The sample regression equation for the hypothetical example of achievement is:

$$Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + e_i, \quad (2)$$

where b_0 is the sample intercept; b_1 is the sample regression coefficient for X_1 controlling for the effect of X_2; b_2 is the sample regression coefficient for X_2 controlling for the effect of X_1; and e_i is the sample error term. Sample data for this example is given in Table 1.

The goal of regression is to find a mathematical solution for b_0, b_1, and b_2 that will best fit the data reflected in Y, X_1, and X_2. As before, OLS estimates will minimize the sum of the squared residuals, $\sum e_i^2$, and will therefore provide the best mathematical fit of the data.

Table 1
Achievement, Time Spent Studying, and Academic Ability

Achievement	Time (in hours)	Ability
88	8	6
75	6	2
64	0	2
99	9	9
95	5	9
93	8	7
85	7	5
82	5	4
79	1	5
78	1	3
91	4	7
85	4	9

Note. Higher scores indicate higher levels of each variable.

The Prediction Equation and Residuals

The prediction equation for the example data is:

$$Y' = b_0 + b_1X_1 + b_1X_2. \quad (3)$$

Actual OLS estimates for this model are:

$$Y' = 63.90 + 1.30(X_1) + 2.52(X_2).$$

Residuals are obtained in a manner identical to that described earlier. Namely, one obtains the predicted value Y' and subtracts this value from the observed Y, i.e.,

$$e_i = Y - Y'.$$

Consider, for instance, the residual for the first data entry in Table 1. The predicted value of achievement is:

$$\begin{aligned} Y' &= 63.90 + 1.30(8) + 2.52(6), \\ &= 63.90 + 1.30(8) + 2.52(6), \\ &= 63.90 + 10.4 + 15.12, \\ &= 89.42. \end{aligned}$$

The observed value of achievement is 88, so the residual is:

e_1 = Y - Y',
= 88 - 89.42,
= -1.42.

This negative value indicates that this person's score was over-predicted; that is, this person's observed score was less than the score predicted for this person given this individual's level of ability and time spent studying.

As previously noted, the goal of OLS is to obtain estimates for the regression coefficients that will provide the smallest possible residuals for all observations in the sample, and when certain assumptions are met, OLS estimates do provide the smallest possible residuals (for proof, see the Pedhazur text). In short, OLS attempts to find the regression line that passes through all observations and that provides the smallest set of squared residuals, $\sum e_i^2$.

Regression Coefficient Interpretation

The model intercept, b_0, is simply the point at which the regression line passes through the Y axis when *both* X_1 and X_2 equal zero. The other regression coefficients indicate the nature (direction and degree) of the *partial* relationship between an independent and dependent variable while controlling for other independent variables in the model. For example, b_1 in the achievement model equals 1.30. The partial effect indicates that a one unit increase in X_1 changes Y by b_1 (or 1.30) units, controlling for the other Xs. In terms of the example data, a one hour increase in the time spent studying is expected to increase achievement by 1.3 points, controlling for the effects of ability. Similar, since b_2 = 2.52, one could state that a one unit increase in ability results in an average increase of 2.52 points in achievement, controlling for the effects of study time.

Emphasis has been placed on the notion of partial effect. This implies that if one did not control for, say, ability, then the relationship found between time spent studying and achievement might be different. This fact can easily be illustrated. Suppose one estimated the relationship between time spent studying and achievement as follows:

$$Y' = b_0 + b_1X_1. \quad (4)$$

The sample estimates for this model are:

$$Y' = 73.17 + 2.34\ X_1.$$

Note that when one does not take into account the impact upon achievement of academic ability, time spent studying appears to have almost twice the estimated effect as found in the multiple regression model. That is, with the single regression model, a one hour increase in time spent studying is estimated to increase achievement by 2.34 points, but when ability is included in the model, the effect of one additional hour of study time on achievement is only an increase of 1.30 points. As this example illustrates, when multiple Xs are theoretically linked to the modeled Y, it is important that these Xs be included in the model in order to obtain the best estimates of the true relationships among the variables.

Another, and perhaps more informative, way to present the modeled effects of a given X is to indicate the change in Y associated with a given amount of change in X. For example, suppose one is most interested in learning by how much Y is likely to change if X changes by, say, five units. With the example data, one may be curious to know the amount of change in achievement that would be associated with an increase of five more hours spent studying during the week. To find this estimated change in Y, simply multiple the partial regression coefficient for X by the number of units of change in X. So if one

were to increase the number of hours spent studying during the week by five, then achievement would be anticipated to increase by (1.30 * 5 = 6.5) 6.5 points, or perhaps over half a letter grade.

Overall Model Fit and Statistical Inference

As previously explained, model fit refers to the degree to which Y' approximates Y. Model fit, or the lack thereof, can be measured by the amount of variation in the residuals. The smaller this variation, the better the model reproduces the observed data, i.e., the better Y' estimates Y. As noted earlier, three measures of model fit include Multiple R, R^2, and adjusted R^2. Multiple R is simply the Pearson correlation between Y and Y', and when this value is squared, the resultant index, R^2, indicates the proportion of variance in Y that can be explained or accounted for by the combination of Xs in the multiple regression model. Since OLS tends to maximize the estimated relationships among the Xs and Y, R^2 tends to overestimate the fit of the model to the data, so a better index of fit is adjusted R^2. For a more thorough treatment of these indices of model fit, see the earlier discussion presented in "Simple Linear Regression: One Quantitative IV."

In multiple regression, one should initially test the tenability of H_0: $R^2 = 0.00$ before proceeding to examine the individual regression coefficients. Should H_0 not be rejected, then one may tentatively conclude that estimated model does not adequately reproduce or explain the observed data, and therefore examination of the individual regression coefficients might be potentially misleading. As with simple linear regression, H_0 is tested via the overall F test. The F ratio is defined as

$$F = \frac{SSR/df_r}{SSE/df_e} = \frac{SSR/k}{SSE/(n - k - 1)} = \frac{MS_{reg}}{MSE} = \frac{R^2/k}{(1-R^2)/(n-k-1)}$$

where
SSR = regression sums of squares;
SSE = residual sums of squares;
df_r = regression degrees of freedom (also denoted df_1);
df_e = residual degrees of freedom (also denoted df_2);
k = number of independent variables (or vectors) in the model;
n = sample size (or number of observations in sample);
MS_{reg} = mean square (same as ANOVA) due to regression (e.g., between);
MSE = mean square error (same as ANOVA mean square within).

The observed R^2 for the achievement data is .874. Using the R^2 formula, the calculated overall F ratio is

$$F = \frac{R^2/k}{(1-R^2)/(n-k-1)} = \frac{.8742/2}{(1 - .8742)/(12 - 2 - 1)} = 31.27,$$

or

$$F = \frac{SSR/df_r}{SSE/df_e} = \frac{906.559/2}{130.441/9} = 31.27.$$

With $df_1 = k = 2$ and $df_2 = n - k - 1 = 12 - 2 - 1 = 9$, the .05 level critical F value is

$$_{.05}F_{2,9} = 4.26.$$

Since 31.27 is larger than 4.26, H_0 is rejected and one may conclude that the model of Y using X_1 and X_2 explains statistically more variability than would be expected by chance alone.

Of course p-values may also be provided for the F test, with the usual decision rule. The p-value for the obtained F ratio is .000, which is clearly less than .05, so the same conclude regarding H_0: $R^2 = 0.00$ is research.

Inferential Procedures for Regression Coefficients

If the overall null hypothesis, H_0: $R^2 = 0.00$, is rejected, the next step in multiple regression (and ANOVA) is to examine the individual variables for statistical significance. As noted above, in multiple regression one is interested in testing whether there is a partial relationship between the k^{th} X and Y, controlling for the other Xs in the model. The null hypothesis states that their is no partial relationship between X_k and Y, that is,

H_0: $\beta_k = 0.00$.

If the null is rejected, then one may conclude that a partial relationship does exist between X_k and Y. A non-directional alternative hypothesis states that a partial relationship does exist, thus:

H_1: $\beta_k \neq 0.00$.

As with simple regression, each regression coefficient has a corresponding standard error (SE_{bk}). The ratio of the partial regression coefficient to its SE provides a t-ratio. As with the two-group t-test and simple regression, the calculated t-ratio may be compared against a critical t ratio to determine statistical significance. Or, one may simply use p-values obtained for each coefficient to determine statistical significance.

For example, in the achievement data the estimated partial effect of time spent studying on achievement is $b_1 = 1.30$, and the SE for b_1 is $SE_{b1} = .437$, so the t ratio for b_1 is:

$t = b_1 / SE_{b1} = 1.30 / .437 = 2.975$.

The critical t value, using an α of .05, and df = n - k - 1, is $_{.05}t_9 = \pm 2.262$, so the null hypothesis is rejected:

If $|t| \geq t_{crit}$ reject H_0, otherwise fail to reject H_0.

If $2.98 \geq 2.262$ *reject* H_0, otherwise fail to reject H_0.

In terms of p-values, the p-value for the obtained t ratio of 2.98 is $p = .015$. The usual decision rule, for non-directional tests, applies:

If $p \leq \alpha$, reject H_0, otherwise fail to reject H_0,

so, since the obtained *p* is less than alpha, H_0 is rejected:

If $.015 \leq .05$, reject H_0, otherwise fail to reject H_0,

and one may conclude that a positive relationship exists between time spent studying and achievement, holding constant the effects of ability. Both b_0 and b_2 may be tested in a similar fashion.

Interval Estimation: Confidence Intervals (CI)

Recall that the CI represents the upper and lower bound to the point estimate of the regression coefficients. Thus, the CI represents, with a set level of precision, a range of possible values for b_k, and therefore the CI provides some indication of the exactness or precision with which sample estimates are derived from the sample data.

As with simple regression, the CI for b_1 may be formed as:

$$b_1 \pm t_{(\alpha/2,df)}SE_{b1}$$

where t is the critical t value obtained from a table of t values representing a two-tailed alpha (α) level (such as .05) with degrees of freedom equal to n-k-1, and SE_{b1} is the standard error of b_1 described above.

For the current example, the 95% confidence interval (.95CI) for b_1 is

$$.95CI: b_1 \pm t_{(\alpha/2,df)}SE_{b1}$$

$$.95CI: 1.302 \pm (2.262)(0.437)$$

$$.95CI: 1.302 \pm 0.988$$

$$.95CI: (2.29, 0.314).$$

CIs for b_0 and b_2 are constructed in a similar fashion.

Such a CI enables the researcher to state that one may be 95% confident that the true population coefficient may be as high as 2.29 or as low as 0.314. As this illustrates, CIs provide a sense of precision that point estimates do not.

Obtaining and Reporting Multiple Regression Results

Regression results can be obtained from SPSS using the same command as used for simple regression. One simply adds additional Xs into the independent variables command section. Sample output from the achievement data is provided below.

```
Multiple R           .93499
R Square             .87421
Adjusted R Square    .84626
Standard Error      3.80703

Analysis of Variance
                    DF      Sum of Squares      Mean Square
Regression           2           906.55875        453.27937
Residual             9           130.44125         14.49347

F =      31.27472       Signif F =  .0001

---------------------- Variables in the Equation ----------------------

Variable              B        SE B     95% Confdnce Intrvl B       Beta
TIME           1.302289     .437043      .313629     2.290948    .399660
ABILITY        2.524210     .499843     1.393487     3.654934    .677329
(Constant)    63.901747    2.835634    57.487101    70.316394

----------- in ------------

Variable           T  Sig T
TIME           2.980  .0155
ABILITY        5.050  .0007
(Constant)    22.535  .0000
```

The independent variables are time spent studying, TIME, and academic ability, ABILITY. The outcome variable is achievement, ACH. Note all components previously discussed, such as R, R^2, and adj. R^2; the regression coefficients b_0 (denoted as constant), b_1 (B for TIME), and b_2 (B for ABILITY); the ANOVA summary table with SSR, SSE, and MSE; and the inferential tests—overall F test (F = 31.27472, p = .0001), and t tests for the partial coefficients.

Reporting results may take several forms. Most common is a tabular display, although for models with few IVs, reporting results within the text of your manuscript may be feasible. The tabular format will be illustrated.

Table 1
Descriptive Statistics and Correlations among Achievement, Time, and Ability

Variable	Correlations		
	Achievement	Time	Ability
Achievement	---		
Time	.720*	---	
Ability	.866*	.472	---
Mean	84.500	4.833	5.667
SD	9.709	2.980	2.605

Note. n = 12
* p < .05

Table 2
Regression of Achievement on Time Spent Studying and Academic Ability

Variable	b	se	95%CI	t
Time	1.30	0.437	0.31, 2.29	2.98*
Ability	2.52	0.500	1.39, 3.65	5.05*
Intercept	63.90	2.836	57.49, 70.32	22.54*

Note. R^2 = .874, adj. R^2 = .846, F = 31.27*, df = 1,9; n = 12.
*p < .05.

(or, the F ratio and df can be reported like this: $F_{1,9}$ = 31.27*)

Both the correlations and regression results show that achievement is positively, strongly, and significantly related at the .05 level to both time spent studying and academic ability. In summary, the more time spent studying and the higher one's academic ability, the greater one's achievement.

Exercises

(1) A researcher wishes to know whether number of hours studied at home is related to general achievement amongst high school students. Student ability should be controlled to assess better the effects of studying.

Student	High School GPA	IQ	Time Spent Studying Per Week (in Hours)
Bill	3.33	117	3
Bob	1.79	90	5
Stewart	2.21	101	12
Linda	3.54	121	9
Lisa	2.89	105	11
Ann	2.54	110	1
Fred	2.66	112	0
Carter	1.10	85	3
Kathy	3.67	128	2

(2) Does SAT adequately predict college success, once rank is controlled?

Student	Freshmen Collegiate GPA	HS Rank*	SAT Scores
Bill	3.33	52	1000
Bob	1.79	233	750
Stewart	2.21	150	890
Linda	3.54	43	1100
Lisa	2.89	95	900
Ann	2.54	43	860
Fred	2.66	120	1010
Carter	1.10	280	640
Kathy	3.67	33	1240

*Out of 300 students.

(3) A teacher is convinced that frequency of testing within her classroom increases student achievement. She runs an experiment for several years in her algebra class. The frequency in which she presents tests to the class varies across quarters. For example, one quarter students are tested only once during the term, while in another quarter students are tested once every week. Is there evidence that testing frequency is related to average achievement?

Quarter	Testing Frequency During Quarter	Average IQ in Class	Overall Class Ach. on Final Exam
Fall 1991	1	105	85.5
Winter 1992	2	108	86.5
Spring 1992	3	108	88.9
Summer 1992	4	109	89.1
Fall 1992	5	107	87.2
Winter 1993	6	110	90.5
Spring 1993	7	108	89.8
Summer 1993	8	114	92.5
Fall 1994	9	110	89.3
Winter 1994	10	112	90.1

(4) An administrator wishes to know whether a relationship exists between the number of tardies or absences a student records during the year and that student's end-of-year achievement as measured by GPA. The administrator randomly selects 12 students and collects the appropriate data. The principal also has standardized ITBS test scores for each student.

Student	GPA	ITBS	Tardies/Absences
Bill	3.33	65	2
Bob	1.79	40	10
Stewart	2.21	50	5
Linda	3.54	70	6
Lisa	2.89	49	3
Ann	2.54	55	4
Fred	2.66	58	6
Carter	1.10	37	12
Bill	3.10	55	3
Sue	2.10	45	8
Loser	2.31	51	6
Kathy	3.67	63	2

Exercises Answers

(1) A researcher wishes to know whether number of hours studied at home is related to general achievement amongst high school students. Student ability should be controlled to assess better the effects of studying.

Table 1
Descriptive Statistics and Correlations among GPA, Time Spent Studying, and IQ

Variable	Correlations		
	GPA	Time	IQ
GPA	---		
Time	.033	---	
IQ	.963*	-.149	---
Mean	2.637	5.11	107.67
SD	.844	4.46	14.053

n = 9
* p < .05

Table 2
Regression of GPA on Time Spent Studying and IQ

Variable	b	se	95%CI	t
Time	.034	.015	-.005, .07	2.16
IQ	.059	.005	.047, .072	11.83*
Intercept	-3.93	.56	-5.32, -2.56	-7.00*

Note. R^2 = .96, adj. R^2 = .95, $F_{2,6}$ = 70.09, n = 9.
*p < .05.

Regression results show that time spent studying is not statistically related to GPA once students' IQ scores are taken into account. The relationship between IQ and GPA is statistically significant and positive. The greater one's IQ, the higher, on average, is one's GPA. Time spent studying does not appear to predict one's GPA.

(2) Does SAT adequately predict college success, once rank is controlled?

Table 1
Descriptive Statistics and Correlations among GPA, HS Rank, and SAT

Variable	Correlations		
	GPA	Rank	SAT
GPA	---		
Rank	-.93*	---	
SAT	.94*	-.83*	---
Mean	2.64	116.56	932.22
SD	0.84	89.33	180.33

n = 9
*$\underline{p}$ < .05.

Table 2
Regression of GPA on Rank and SAT

Variable	$\underline{b}$	$\underline{se}$	95%CI	t
Rank	-.005	.002	-.008, -.001	-3.03*
SAT	.002	.001	.001, .004	3.28*
Intercept	.864	.862	-1.24, 2.97	1.00

Note. R^2 = .95, adj. $\underline{R}^2$ = .93, $\underline{F}_{2,6}$ = 57.43*, n = 9.
*$\underline{p}$ < .05.

Both HS rank and SAT scores are statistically related to student GPA. Results show that as one's rank increases, predicted GPA declines; however, as SAT scores increase, GPA tends to increase.

(3) A teacher is convinced that frequency of testing within her classroom increases student achievement. She runs an experiment for several years in her algebra class. The frequency in which she presents tests to the class varies across quarters. For example, one quarter students are tested only once during the term, while in another quarter students are tested once every week. Is there evidence that testing frequency is related to average achievement?

Table 1
Descriptive Statistics and Correlations between Testing Frequency, Achievement, and IQ

Variable	Correlations		
	Final	Testing	IQ
Final	---		
Testing	.75*	---	
IQ	.89*	.77*	---
Mean	88.94	5.50	109.10
SD	2.06	3.03	2.56

n = 10
* p < .05

Table 2
Regression of Achievement on Testing Frequency and IQ

Variable	b	se	95%CI	t
Testing	.12	.18	-0.31, 0.54	0.66
IQ	.61	.21	0.10, 1.11	2.86*
Intercept	22.06	22.45	-31.02, 75.14	0.98

Note. R^2 = .80, adj. R^2 = .74, $F_{2,7}$ = 13.94*, n = 10.
*p < .05.

Correlations show that testing frequency and IQ are both related to student achievement, but regression results show that once IQ is taken into account, testing frequency no longer predicts student achievement. Students with higher IQs tend to obtain higher achievement scores; testing frequency does not seem to relate to achievement.

(4) An administrator wishes to know whether a relationship exists between the number of tardies or absences a student records during the year and that student's end-of-year achievement as measured by GPA. The administrator randomly selects 12 students and collects the appropriate data. The principal also has standardized ITBS test scores for each student.

Table 1
Descriptive Statistics and Correlations among GPA, ITBS, and Absences

Variable	Correlations		
	GPA	ITBS	Absences
GPA	---		
ITBS	.92*	---	
Absences	-.85*	-.72*	---
Mean	2.60	53.17	5.58
SD	.76	9.93	3.15

n = 12
*$\underline{p}$ < .05.

Table 2
Regression of GPA on Absences and ITBS scores

Variable	$\underline{b}$	$\underline{se}$	95%CI	t
ITBS	.048	.01	.03, .07	4.62*
Absences	-.096	.03	-.17, -.02	-2.92*
Intercept	.58	.70	-1.00, 2.17	0.83

Note. R^2 = .92, adj. $\underline{R}^2$ = .90, $\underline{F}_{2,9}$ = 50.47*, n = 12.
*$\underline{p}$ < .05.

Regression results show that both ITBS scores and number of absences are statistically related to student GPA. As the number of absences increase, GPA tends to decline; as ITBS scores increase, GPA tends to increase.