An Introduction to Linear Regression Analysis

George H Olson, Ph. D.

Doctoral Program in Educational Leadership
Appalachian State University
(Fall 2010, Expanded Summer 2012)

(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 A (the data have been taken from Pedhauser,1997; p. 98). To make this example meaningful, let's let the variable, *Score*, be taken as a score on 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[1]. (A much more detailed discussion of regression analysis can be found HERE.)

Descriptive Statistics. My first step was to obtain the descriptive statistics (N, Mean, and Standard Deviation) 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,

$SEM_{score} = STD_{score} / \surd(N_{score} -1) = 2.763/4.359 = .634.$

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[2]. From Table 1 it is apparent that the mean score on the test was nearly 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*

[1] An Excel version of this example is given in Appendix B.

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

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 summarized 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-Squared (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(b)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.826(a)	.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_{reg} and SS_{res}, which are used for computing MS_{reg} and MS_{res} (by dividing each SS term by its corresponding degrees of freedom). The *F* statistic is then computed by dividing MS_{reg} by MS_{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(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	97.132	2	48.566	18.319	.000(a)
	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 given. The variables in the model are given first (the constant is equal to $\bar{Y} - b_1 X_1 - b_2 X_2$). 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 equal zero. This is done, for each *b*, using a *t* test with one degree of freedom:

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

Table 5. Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	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

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 (β) 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:

CI95: {.693-(2.1)(.189) <= β <= .693-(2.1)(.189)}

= .296 <= β <= 1.090, which DOES NOT include zero.

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.

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.

Dummy, Effect, and Orthogonal Coding

When we test a null hypothesis that, say, the means of two populations are equal, e.g., H_0: $\mu_1 = \mu_2$, this is tantamount to hypothesizing that the *knowledge of group membership* provides no information to help us predict differences among the group outcomes. On the other hand, if we reject the null hypothesis, then we, essentially, are saying that knowledge of group membership does predict group outcomes. So, if we had a way to code group membership in such a way that we could regress the outcome measure on the group membership code then we could analyze the data using regression analysis. In other words, we could set up a regression model such as

$$\hat{Y}_i = b_0 + b_1 X_{i1}, \quad \text{(Eq. 1)}$$

where the X's carry the codes for group membership. In this case, testing the difference between means is equivalent to testing the significance of b_1. In what follows, we will see how this is accomplished.

There are three types of coding schemes that are widely used in regression analysis to test differences among group means: *dummy* coding, *effect* coding, and *orthogonal* coding. Each of these is considered below, first in the simple two-group case, then for the cases analogous to a one-way ANOVA, and finally for a factorial ANOVA design.

Two-group Design

We'll begin with the simple, two-group design illustrated in Table 6. There, the two groups represent, say, two treatment conditions, and the outcome measure, Y, is the dependent variable of interest.

Table 6

Outcome Measure, *Y*	
Group 1	Group 2
1	3
2	3
2	4
3	4
2	2

A *t* test of the difference between means, using Excel, yields Table 7, where it is shown that both the one-tailed and two-tailed *t* tests of the difference between means are statistically significant, $t(1) = -2.449$; $p = .020$ (one-tail); $p = .040$ (two-tail). To replicate the test using regression we use either effect or orthogonal coding since, in the two-group case, dummy coding reduces to orthogonal coding.

Table 7: t-Test: Assuming Equal Variances

	Group 1	*Group 2*
Mean	2	3.2
Variance	0.5	0.7
Observations	5	5
df	8	
t Stat	-2.449	
P(T<=t) one-tail	0.020	
P(T<=t) two-tail	0.040	

Two-Group Dummy Variable Coding

Dummy coding is, perhaps, the easiest system of coding treatment effects (i.e., group membership effects) in ANOVA. It uses columns of 1's and 0's to identify group membership. For instance, in Table 8, there are two "indicator" columns, or variables, i.e., the X's. For X1, and individual is coded 1 if the individual belongs to Group 1, and zero otherwise. For X2, an individual is coded 1 if the individual belongs to Group 2; zero otherwise. Thus, the group to which an individual belongs is uniquely identified. An individual in Group 2, for instance, has the dummy code: 0 1.

Table 8: Effect Coding

	Effect Codes		
Group	$X1$	$X2$	Y
1	1	0	1
1	1	0	2
1	1	0	2
1	1	0	3
1	1	0	2
2	0	1	3
2	0	1	3
2	0	1	4
2	0	1	4
2	0	1	2

However, there is a problem with the coding shown in Table 8. Only one of the two X columns are required to identify the groups to which individuals belong. We can ignore either one of the X columns and still be able to uniquely identify group membership. For instance, if we ignore $X2$ then,

Individuals in Group 1 are identified by the dummy code: 1 for $X1$, and

Individuals in Group 2 are identified by the dummy code: 0 for $X1$.

Setting-up the regression analysis to regress Y on X involves testing the model given above in Eq. 1, where, here, $X1$ is Effect Code, $X1$ given in Table 3. Regressing Y on $X1$, using Excel, yields the following regression statistics (Tables: 9, 10, & 11).

Table 9: Regression Statistics

Multiple R	0.655
R Square	0.429
Standard Error	0.775
Observations	10

Table 10: ANOVA (Effect coding)

	df	*SS*	*MS*	*F*	*Sig*
Regression	1	3.600	3.600	6.000	0.040
Residual	8	4.800	0.600		
Total	9	8.400			

In dummy coding, the group coded 0 is known as the reference (or control) group. In dummy coding, the intercept is equal to the mean of the reference group. Since, in our example, Group 2 is the reference group, the mean for Group 2, given by the intercept in Table 11, is 3.2. Furthermore, in the two-group case, the coefficient associated with $X1$ is the difference between the mean of the group with the dummy code of 1 and the reference group. Hence, -1.2 is the difference between the mean of Group 1 and the reference group (Group 2).

Table 11: Regression Coefficients (Effect Coding)

	Coef.	*Standard Error*	*t Stat*	*P-value*
Intercept	3.200	0.346	9.238	0.000
$X1$	-1.200	0.490	-2.449	0.040

How do these results compare to the *t* test results given earlier in Table 7 First, recall that for a one-degree of freedom test, $F = t^2$. From Table 2, t =-2.499. t^2 = 6.000, which is equal to *F* in Table 10. Note, also, that -2.449 is the value of the *t* test for the $X1$ regression coefficient (Table 6). Furthermore, from Table 7 we note that the mean difference, $\left(X1 - X2\right) = -1.2$, which is identical to the coefficient for $X1$, given in Table 11. In other words, the *effect* of the treatment is -1.2 units—whatever the treatment Group 2 received, it resulted in Group 2 having a mean score 1.2 units below Group 1. Both the *t* test and the regression analysis provide sufficient evidence for *rejecting* the null hypothesis. The regression analysis does tell us, however, (Table 9) that the "treatment" effect accounted for 43% of the variance in Y (R^2 = .429). (Incidentally, coefficients are more commonly referred to as *regression weights* (or more accurately *unstandardized* regression weights).

In the *t* test analysis we are given the mean Y for each group. The means are not given, at least not directly, by the regression analysis. However, the means can be computed easily from the regression analysis. Recall from the readings, in HyperStat, that the intercept,

$$b_0 = \overline{Y} - b_1\overline{X}, \qquad \text{(Eq. 2)}$$

so that, ignoring group membership, the overall mean of Y is,

$$\begin{aligned}\overline{Y} &= b_0 + b_1\overline{X} \\ &= 3.2 + (-1.2)(.5) \\ &= 2.6.\end{aligned}$$

Considering the groups individually, for Group 1, $\overline{X}$ = 1, so that $\overline{Y}$ for Group 1 is given by,

$$b_0 = 3.2 + (-1.2)(1)$$
$$= 2.0.$$

For Group 2, $\bar{X}$ =0, so that $\bar{Y}$ for Group 2 is given by,

$$b_0 = 3.2 + (0)(1)$$
$$= 3.2.$$

Two-Group Effect and Orthogonal Coding

In the two-group case, Effect Coding and orthogonal coding are identical. Both types of coding use 1's and -1's to denote group membership (later we will see that, in orthogonal coding, 0's and other values are used, depending upon the number of groups in the design). For the current example, effect and orthogonal coding would look like the coding in Table 12. In the table, you can see that the individuals in each group are uniquely identified by their $X1$ value.

Table 12: Orthogonal Coding

Group	Effect Code (X)	Y
1	1	1
1	1	2
1	1	2
1	1	3
1	1	2
2	-1	3
2	-1	3
2	-1	4
2	-1	4
2	-1	2

Using the **Regression** procedure in Excel yields the results shown in Tables 13, 14 and 15. An advantage of effect coding is that, when the group sizes are equal, the intercept, i.e., the coefficient, b_0, is equal to the overall mean; furthermore, the coefficient, b_1, is proportional to the difference between the means. In this two-group case, b_1 is equal to half the difference between the means of Group 1 and Group 2:

$$b_1 = \frac{\bar{Y}_1}{2} - \frac{\bar{Y}_2}{2}$$
$$= \frac{2}{2} - \frac{3.2}{2}$$
$$= 1 - 1.6$$
$$= -.6,$$

which, is the value given for the coefficient, $X1$ in Table 15.

The percent of variance in Y accounted for (or explained by) treatment (R^2) is .43—the same as that given in the earlier tables. In fact, the values in Table 13 (and Table 14) are identical to the values given in Table 9 (and 10)—an indication that, regardless of the type of coding used, the percentage of variance explained by group identification is the same.

Table 13: Regression Statistics

Multiple R	0.655
R Square	0.429
Standard Error	0.775
Observations	10

Table 14: ANOVA

	df	*SS*	*MS*	*F*	*Sig*
Regression	1	3.600	3.600	6.000	0.040
Residual	8	4.800	0.600		
Total	9	8.400			

Table 15: Regression Coefficients

	Coef.	*Std Err*	*t Stat*	*Sig*
Intercept	2.600	0.245	10.614	0.000
OrthCode (X)	-0.600	0.245	-2.449	0.040

What does change, however, are the regression coefficients (compare Table 15 with Table 11). Depending upon the type of coding used, the coefficients represent different estimates. In dummy coding, the intercept, b_0, is the mean of the group having X coded as zero. With orthogonal coding, on the other hand, b_0 represents the overall grand mean. Furthermore, with dummy coding the coefficient, b_1, is the difference between the mean of the group coded 1 and the mean of the group coded 0. With orthogonal coding, b_1, is proportional to the differences between the groups.

Note, also, that in Table 15 (and earlier, in Table 11), *t* statistics are provided (*t* Stat). These *t* statistics are used to test the null hypothesis that the corresponding coefficient is zero. For, the intercept, *t* is 10.614 with $p < .001$. For the coding variable, $t = -2.449$ with $p = .040$. Both coefficients are statistically significant. The coefficient for the intercept will virtually always be statistically significant and, hence, is of little or no interest. The significance of the coded variable (X) tells us that the variable represented by X has an effect, i.e. that the difference between Group 1 and Group 2 is statistically significant.

One-way ANOVA Design

We will now consider a four-group, one-way ANOVA design. The design is shown in Table 16

Table 16: One-way ANOVA Design

Group 1	Group 2	Group 3	Group 4
2	3	3	5
3	1	2	3
2	2	4	4
1	3	4	2
2	2	3	5

An analysis of variance, using Excel, yielded the results shown in Tables 17 & 18.

Table 17: Descriptive Statistics

Group	*N*	*Mean*	*Variance*
Group 1	5	2.000	0.500
Group 2	5	2.200	0.700
Group 3	5	3.200	0.700
Group 4	5	3.800	1.700

Table 18: ANOVA for a Four-group Design

Source	*SS*	*df*	*MS*	*F*	*P-value*
Between Groups	10.800	3	3.600	4.000	0.027
Within Groups	14.400	16	0.900		
Total	25.200	19			

The results indicate a significant treatment (Between Groups) effect: $F(3,16) = 4.000$; $p = .027$. Ordinarily, a post-hoc test comparing pairs of group means (e.g., Tukey's HSD, Newman-Keuls, etc.), would be indicated, but is ignored here.

Dummy Coding in One-way ANOVA

As mentioned earlier, dummy coding is, perhaps, the easiest system of coding treatment effects (i.e., group membership effects) in ANOVA. It uses columns of 1's and 0's to identify group membership. For instance, in Table 14, there are four "indicator" columns, or variables, i.e., the X's. For $X1$, an individual is coded 1 if the individual belongs to Group 1, and zero otherwise. For $X2$, an individual is coded 1 if the individual belongs to Group 2, and so on. In this way the group to which an individual belongs is uniquely identified. For instance, an individual in Group 2 has the dummy code: 0 1 0 0.

Note, however, that, as we saw in the case of the two-group design, there is a redundancy in the coding displayed in Table 19. Only three of the four X columns are required to identify the groups to which individuals belong. We can ignore any one of the X columns and still be able to identify unique group membership. For instance, if we ignore $X4$ then, Individuals belong to Group 1 have the dummy code: 1 0 0,

Individuals belong to Group 2 have the dummy code: 0 1 0,

Individuals belong to Group 3 have the dummy code: 0 0 1, and

Individuals belong to Group 4 have the dummy code: 0 0 0.

Table 19: Dummy Variable Coding

Group	$X1$	$X2$	$X3$	$X4$	Y
1	1	0	0	0	2
1	1	0	0	0	3
1	1	0	0	0	2
1	1	0	0	0	1
1	1	0	0	0	2
2	0	1	0	0	3
2	0	1	0	0	1
2	0	1	0	0	2
2	0	1	0	0	3
2	0	1	0	0	2
3	0	0	1	0	3
3	0	0	1	0	2
3	0	0	1	0	4
3	0	0	1	0	4
3	0	0	1	0	3
4	0	0	0	1	5
4	0	0	0	1	3
4	0	0	0	1	4
4	0	0	0	1	2
4	0	0	0	1	5

In fact, if all four of the X columns were included in a regression analysis, the redundancy would lead to an indeterminate solution. In general, we need only as many Xs as there are independent variable degrees of freedom (number of groups -1).

To analyze the data in Table 19, we set up the following regression equation.

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 , \qquad \text{(Eq. 3)}$$

using only the first three predictors ($X1$, $X2$, and $X3$). The regression analysis computed in Excel, yielded the results shown in Tables 20, 21, and 22. As can be seen, the model accounted for 43% of the variance in the dependent variable, Y, i.e., R^2 = .429 (the fact that this is identical to the R^2 given in the two-group examples, earlier, is purely coincidental.)

Table 20: Regression Statistics

Multiple R	0.655
R Square	0.429
Standard Error	0.949
Observations	20.000

The analysis of variance table (Table 21) confirms the statistical significance of the regression model, $F(3,19) = 4.000$, $p = .027$.

Table 21: ANOVA

	df	*SS*	*MS*	*F*	*Sig*
Regression	3	10.800	3.600	4.000	0.027
Residual	16	14.400	0.900		
Total	19	25.200			

As in the case of a two-group design, when using dummy coding, the group with all zeros for the Xs is called the *reference* (sometimes the *control*) *group*. (Obviously, with some rearranging of the 1s and 0s, any one of the groups can be designated the reference group.) When examining the regression coefficients (Table 22), the coefficient for the intercept is equal to the sample mean of the reference group (in this case, Group 4). Hence, 3.800 is the mean of Group 2 (see Table 17). Each of the other coefficients are equal to the difference between the mean of the reference group and the mean of the group corresponding to the coefficient. For instance, the coefficient for $X2$ is equal to the mean of Group 2 minus the mean of the mean of the reference (2.2 – 3.8 = -1.6).

Table 22: Regression Coefficients

	Coef.	*Std Err*	*t Stat*	*P-value*
Intercept (b_0)	3.800	0.424	8.957	0.000
X1	-1.800	0.600	-3.000	0.008
X2	-1.600	0.600	-2.667	0.017
X3	-0.600	0.600	-1.000	0.332

The *t* tests for the regression coefficients in Table 22 indicate that the difference between Group 1 and Group 4, and difference between Group 2 and Group 4 are significant. The difference between Group 3 and Group 4 is not significant.

The empirical regression equation, using the results obtained from the sample is:

$$\hat{Y}_i = 3.80 + (-1.80)X_{i1} + (-1.60)X_{i2} + (-.60)X_{i3}\cdot$$

Effect Coding in One-way ANOVA

Recall, from the effect coding section under the two-group design, that effect coding uses 1's and -1's to designate group (i.e., treatment group) membership. Since the one-way design being considered here has four groups, there are 3 (number of groups minus one) degrees of freedom. In effect coding, there is one (predictor) column for each degree of freedom. For the current example, effect coding is given in Table 23.

Table 23: Effect Coding

Group	X1	X2	X3	Y
1	1	0	0	2
1	1	0	0	3
1	1	0	0	2
1	1	0	0	1
1	1	0	0	2
2	0	1	0	3
2	0	1	0	1
2	0	1	0	2
2	0	1	0	3
2	0	1	0	2
3	0	0	1	3
3	0	0	1	2
3	0	0	1	4
3	0	0	1	4
3	0	0	1	3
4	-1	-1	-1	5
4	-1	-1	-1	3
4	-1	-1	-1	4
4	-1	-1	-1	2
4	-1	-1	-1	5

The regression model for the analysis using effect coding is the same as that given in Eq. 3 using the new values for the Xs. Computing the regression analysis, using Excel, yields the results given in Tables 24, 25, and 26. Unsurprisingly, the percent of variance in Y explained (R^2) by the effect variables is .43—the same as was found in the regression analysis using dummy variables (Table 20). That this regression model is statistically significant is shown in Table 25: $F(3,16) = 4.00$; $P = .027$. These are identical to the results given in Table 21 for the analysis using dummy variables.

Table 24: Regression Statistics

Multiple R	0.655
R Square	0.429
Standard Error	0.949
Observations	20

Table 25: ANOVA

	df	*SS*	*MS*	*F*	*Sig*
Regression	3	10.800	3.600	4.000	0.027
Residual	16	14.400	0.900		
Total	19	25.200			

What is different between the analysis using dummy coding and the analysis using effect coding is the interpretation of the regression coefficients (Table 26). When using effect coding, the intercept (b_0) gives the overall grand mean. Each of the other coefficients gives the deviation of the mean of its corresponding group from the grand mean. For instance the grand mean, the mean over all cases in the design, is 2.8. The deviation of the Group 1 mean from the grand mean is (2 – 2.8 = -.8).

Table 26: Regression Coefficients

	Coef.	*Std Err*	*t Stat*	*P-value*
Intercept	2.800	0.212	13.199	0.000
X1	-0.800	0.367	-2.177	0.045
X2	-0.600	0.367	-1.633	0.122
X3	0.400	0.367	1.089	0.292

In this analysis, only the deviation of the mean for Group 1 from the overall grand mean is statistically significant: $t(1) = -2.177$; $p = .045$.

Orthogonal Coding in One-Way ANOVA

Orthogonal coding, for the one-way ANOVA design is similar to effect coding except that 0's, 1s, and higher values are also used to identify group membership. Orthogonal coding requires the two conditions, (1) that the sum of the values coded for any given X variable is zero, and (2) that the sum of the cross products (Xi x Xj , for all $I \neq j$) equal zero. The orthogonal coding given in Table 27 satisfies these conditions. Summing the coded values under $X1$ yields zero. This is true, also, for $X2$ and $X3$. Furthermore, summing the cross products ($X1$ x $X2$) also yields zero.

Again the regression model to be analyzed is identical to that given in Eq. 3. Only here, the values of the Xs are different. The regression analysis, using the procedure in Excel, yields the same R^2 (Table 28) and ANOVA (Table 29) results as those given earlier in the examples using dummy coding and effect coding. The differences occur in the regression coefficients (Table 30).

Table 27: Orthogonal Coding

Group	*X*1	*X*2	*X*3	*Y*
1	1	1	1	2
1	1	1	1	3
1	1	1	1	2
1	1	1	1	1
1	1	1	1	2
2	-1	1	1	3
2	-1	1	1	1
2	-1	1	1	2
2	-1	1	1	3
2	-1	1	1	2
3	0	-2	1	3
3	0	-2	1	2
3	0	-2	1	4
3	0	-2	1	4
3	0	-2	1	3
4	0	0	-3	5
4	0	0	-3	3
4	0	0	-3	4
4	0	0	-3	2
4	0	0	-3	5

Table 28: Regression Statistics

Multiple R	0.655
R Square	0.429
Standard Error	0.949
Observations	20

Table 29: ANOVA

	df	*SS*	*MS*	*F*	*Sig*
Regression	3	10.800	3.600	4.000	0.027
Residual	16	14.400	0.900		
Total	19	25.200			

In orthogonal coding, the intercept estimates the grand mean. The individual regression coefficients represent weighted contrasts between group means. The coefficient for $X1$ is equal to

$$\frac{1}{2}(\overline{Y}_1 - \overline{Y}_2)\cdot$$

Table 30: Regression Coefficients

	Coef	*Std Err*	*t Stat*	*P-value*
Intercept	2.800	0.212	13.199	0.000
X1	-0.100	0.300	-0.333	0.743
X2	-0.367	0.173	-2.117	0.050
X3	-0.333	0.122	-2.722	0.015

The coefficient for $X2$ is equal to

$$\frac{1}{4}(\overline{Y}_1 + \overline{Y}_2 - 2\overline{Y}_3),$$

and $X3$, to

$$\frac{1}{6}(\overline{Y}_1 + \overline{Y}_2 + \overline{Y}_2 - 3\overline{Y}_4).$$

What is important is that the coefficients are proportional to mean differences. For instance, the coefficient for $X1$ (i.e., b_1) is (1/2)(2.0 - 2.2) or -.1. The coefficient for $X2$ is (1/4)[(2.0 + 2.2) - 2 x 332], or -.367. And, for $X3$, the regression coefficient is (1/6)[(2.0 + 2.2 + 3.2) - 3 x 3.8], or -.333.

In this analysis, using orthogonal coding, only the coefficients for $X2$ and $X3$ are statistically significant. From this we can conclude that the combined mean of Groups 1 & 2 is different from the mean of Group 3 (Group 3 had a higher mean than the *average* of Groups 1 & 2.) Furthermore, the mean of Group 4 was significantly different from the combined mean of Groups 1, 2 & 3.

Factorial ANOVA Design

We will now consider a two-way, 2 x 3, factorial ANOVA design. Aside from a 2 x 2 factorial design, this is simplest factorial design we can consider. The design for this example is shown in Table 33. Factor A, which could be, for example, type of instruction has two levels (e.g., Level A1 = F2F, and Level A2 = Online). Factor B, which might be level of course has three levels (e.g., Level B1 = Elementary, Level B2 = Intermediate, and Level B3 = Advanced).

To analyze these data, I used the *Two-Way Factorial ANOVA for Independent Samples* from VassarStats (http://vassarstats.net/). The procedure yielded summary statistics that allowed me to construct the descriptive statistics table shown in Table 34. Actually, a sufficient table of descriptive statistics needed to include only N,

Table 33: 2 x 3 Factorial ANOVA Design

		Factor B		
		Level B1	Level B2	Level B3
Factor A	Level A1	2	3	3
		3	2	4
		1	2	5
		4	1	4
		3	3	5
	Level A2	6	4	3
		2	4	4
		6	2	2
		3	1	3
		4	2	2

Mean, and SD, although authors often include the other statistics as well. I also used the Excel **ToolPac procedure, ANOVA: 2 Factor with Replication** and obtained the same results (see Appendix C).

Table 34: Descriptive Statistics for the 2 x 3 ANOVA Design

			Factor B			
		Statistic	Level B1	Level B2	Level B3	TOTAL
actor A	Level A1	N	5	5	5	15
		Mean	2.6	2.2	4.2	3.0
		SD	1.14	0.84	0.84	1.25
	Level A2	N	5	5	5	15
		Mean	4.2	2.6	2.8	3.2
		SD	1.79	1.34	0.84	1.47
	Total	N	10	10	10	30
		Mean	3.4	2.4	3.5	3.1
		SD	1.65	1.07	1.08	1.35

The analysis of variance performed by the VassarStats procedure is summarized in the Table 35. The table shows that the *main effect* for Factor A is not significant, $F(1,24) = 0.21$; $p = .651$. There is no basis for concluding that Factor A (type of instruction, in this case) has any on the outcome variable (Y). Similarly the main effect of Factor B is not significant, $F(2,24) = 2.64$; $p = .090$. Again, we have no basis for concluding that Factor B (Level of instruction) has *main* effect on the outcome variable. The presence of a significant A x B interaction, $F(2,24) = 4.07$, $p = .030$, however complicates matters. An inspection of the cell means given in Table 34 shows that an F2F environment was more effective for students receiving an elementary level of instruction while an online environment was more effective for students receiving an advanced level of instruction. There was little difference in the means for F2F and online environments for students receiving an intermediate level of instruction.

Table 35: Analysis of Variance Table

Source	SS	df	MS	F	*p*
Factor A	0.30	1	0.30	0.21	.651
Factor B	7.40	2	3.70	2.64	.090
A x B	11.40	2	5.70	4.07	.030
Residual (error)	33.60	24	1.40		
Total	52.70	29			

Dummy Variable Coding in Factorial ANOVA

I now turn to analyzing the 2 x 3 ANOVA problem using Regression Analysis.

As in the case of a one-way ANOVA, dummy coding is probably the easiest system of coding treatment effects. Again, It uses columns of 1's and 0's to identify group membership. The Dummy Variable Coding scheme is given in Table 36. The coding is for two factors, Factor A having two levels and Factor B having three levels. As shown, an individual belonging to the treatment classification, A1B1 had, for D1 thru D5, the values 1, 1, 0, 1, and 0. An individual in treatment classification, A2B2, has the dummy codes 0, 0, 1, 0, and 0. In this way every individual is uniquely identified as belonging to one, and only one, cell in in the 2 x 3 table (Table 33). Note also that the redundancies described earlier in the case of two-group dummy coding have been removed.

One way of determining how many dummy codes are needed is to use the degrees of freedom. The Factor A effect has one degree of freedom (number of levels of Factor A minus 1). Hence D1 codes the Factor A effect. Factor B has two degrees of freedom (number of levels of Factor B minus 1). Hence, D2 and D3 code the Factor B effect. The number of degrees for the A x B interaction is the product of the degrees of freedom for the two factors in the interaction: 2 x 1 = 2. In fact the coding of the interactions is easily established by multiplying the codes for the variables involved in the interaction. For instance, the first interaction dummy variable is D4. The codes for each individual's D4 are obtained by multiplying D1 times D2. The codes for their D5 variable are obtained by multiplying D1 times D2.

Table 36: Dummy Coding for a 2 x 3 Factorial ANOVA

Factor	Factor	Effect Variables				
A	B	D1	D2	D3	D4	D5
A1	B1	1	1	0	1	0
		1	1	0	1	0
		1	1	0	1	0
		1	1	0	1	0
		1	1	0	1	0
	B2	1	0	1	0	1
		1	0	1	0	1
		1	0	1	0	1
		1	0	1	0	1
		1	0	1	0	1

Continued on next page.

Continued from previous page.

A1	B3	1	0	0	0	0
		1	0	0	0	0
		1	0	0	0	0
		1	0	0	0	0
		1	0	0	0	0
A2	B1	0	1	0	0	0
		0	1	0	0	0
		0	1	0	0	0
		0	1	0	0	0
		0	1	0	0	0
	B2	0	0	1	0	0
		0	0	1	0	0
		0	0	1	0	0
		0	0	1	0	0
		0	0	1	0	0
	B3	0	0	0	0	0
		0	0	0	0	0
		0	0	0	0	0
		0	0	0	0	0
		0	0	0	0	0

There is a problem in using dummy coding in a two-way design (and higher) that we did not encounter with the analysis of a one-way design. Whereas, with one-way ANOVA designs, the dummy variables are *independent* (i.e., not correlated), this is not the case with the dummy variables in a two-way design. We can see this by correlating the dummy variables given in Table 36 (I used Excel to do this). I got the correlation matrix shown in Table 37. As is readily seen the correlations between D1 (the dummy code for Factor A) and both D2 and D3 (the dummy codes for the Factor B effect) are both zero. Hence the codes for Factor A and Factor B are independent. However, the two dummy codes for the Factor B (D2 and D3) *are* correlated ($r_{D2,D3}$ = -.5). This means that the two dummy codes are *confounded*. The same is true for the correlations between D1, D2, D3 and the dummy codes for the interaction, D4 and D5. Hence, by virtue of the dummy codes used in this scheme, the interaction is confounded with both main effects (Factor A and Factor B). Because of this (and for other reasons beyond the scope of this development) we cannot use dummy coded regression models with Factorial ANOVA designs. We will see that we do not have this problem with effect and orthogonal coding schemes.

Table 37: Correlations Among the Dummy Variables

	D1	*D2*	*D3*	*D4*	*D5*
D1	1.000				
D2	0.000	1.000			
D3	0.000	-0.500	1.000		
D4	0.447	0.632	-0.316	1.000	
D5	0.447	-0.316	0.632	-0.200	1.000

Effect Variable Coding in Factorial ANOVA

You will recall, from the earlier discussion of regression analysis using effect coding that that effect coding uses 1's and -1's to designate group membership. For a two-factor design we need $df_{FactorA}$ effect variables for the first Factor and $df_{FactorB}$ effect variables for the second factor. For the current example, this means 1 effect variable for the Factor A and two effect variables for Factor B.

In a two-factor design there is an interaction, Factor A x Factor B, also. As was shown in the section on Dummy Variable Coding in Factorial ANOVA, the number of degrees for the A x B interaction is the product of the degrees of freedom for the two factors in the interaction: 1 x 2 = 2. In fact the coding of the interactions is easily established by multiplying the codes for the variables involved in the interaction. For instance, the first interaction effect variable is E4. The codes for each individual's E4 are obtained by multiplying E1 times E2. The codes for their E5 variables are obtained by multiplying E1 times E2. For the current example, effect coding is given in Table 38.

Table 38. Effect Coding for a 2 x 3 Factorial ANOVA

Factor	Factor	Effect Variables					
A	B	E1	E2	E3	E4	E5	Y
		1	1	1	1	1	2
		1	1	1	1	1	3
	B1	1	1	1	1	1	1
		1	1	1	1	1	4
		1	1	1	1	1	3
		1	-1	0	-1	0	3
		1	-1	0	-1	0	2
A1	B2	1	-1	0	-1	0	2
		1	-1	0	-1	0	1
		1	-1	0	-1	0	3
		1	0	-1	0	-1	3
		1	0	-1	0	-1	4
	B3	1	0	-1	0	-1	5
		1	0	-1	0	-1	4
		1	0	-1	0	-1	5
		-1	1	1	-1	-1	6
		-1	1	1	-1	-1	2
	B1	-1	1	1	-1	-1	6
		-1	1	1	-1	-1	3
A2		-1	1	1	-1	-1	4
		-1	-1	0	1	0	4
		-1	-1	0	1	0	4
	B2	-1	-1	0	1	0	2
		-1	-1	0	1	0	1
		-1	-1	0	1	0	4

Continued on next page.

Continued from previous page.

	-1	0	-1	0	1	3
	-1	0	-1	0	1	4
B3	-1	0	-1	0	1	2
	-1	0	-1	0	1	3
	-1	0	-1	0	1	2

The analysis of these data using regression analysis is moderately complicated. It involves partialing out the proportion of variance that each effect explains or accounts for. Using the data in Table 38, the *full* regression model is given by,

$$\hat{Y}_{Full} = b_0 + b_1E_1 + b_2E_2 + b_3E_3 + b_4E_4 + b_5E_5, \quad \text{(Eq. 4)}$$

where the b_i regression weights (b_0 is the intercept), and the E_i are the effect variables. To test the effect of Factor A, say, we compute a restricted model:

$$\hat{Y}_{Restricted} = b'_0 + b'_2E_2 + b'_3E_3 + b'_4E_4 + b'_5E_5, \quad \text{(Eq. 5)}$$

where the apostrophes (ʻ) denote that the coefficients in the restricted model are different from the coefficients in the full model. Note that in the restricted model the effect variable for Factor A (i.e., E_1) is omitted. Since the full model has more explanatory terms (or predictors) in it than does the restricted model, it stands to reason that the proportion of variance accounted for (or explained), R^2, by the full model will be equal to or greater than the R^2 for the restricted model. Hence, we can use the difference in variance explained, $R^2_{Full} - R^2_{Restricted}$ as a measure of the effect due to Factor A. The significance test for this effect is,

$$F = \frac{(R^2_{Full} - R^2_{Restricted}) / (\text{df}_{Full} - \text{df}_{Restricted})}{(1 - R^2_{Full}) / (N - \text{df}_{Full} - 1)}, \quad \text{(Eq. 6)}$$

where

df_{Full} = the degrees of freedom for the full model and

$\text{df}_{Restricted}$ = the degrees of freedom for the restricted model.

I used the Regression procedure in the Excel ToolPac to compute the full model and the three restricted models (one for each effect, in which the orthogonal variables for the factor under consideration were omitted. The regression results for the Full Model, yielded by the Regression procedure in Excel are given in Tables 39 &40.

Table 39: Full Model Regression Analysis (E1 E2 E3 E4 E5)

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.602
R Square	0.362
Adj R Square	0.230
Std. Err.	1.183
Observations	30.000

Table 40: ANOVA of the Full Model

	df	*SS*	*MS*	*F*	*Sig*
Regression	5.000	19.100	3.820	2.729	0.043
Residual	24.000	33.600	1.400		
Total	29.000	52.700			

	Coef	*Std Err*	*t Stat*	*Sig*
Intercept	3.100	0.216	14.350	0.000
E Variable 1	-0.100	0.216	-0.463	0.648
E Variable 2	0.700	0.306	2.291	0.031
E Variable 3	-0.400	0.306	-1.309	0.203
E Variable 4	0.100	0.306	0.327	0.746
E Variable 5	-0.800	0.306	-2.619	0.015

As in the case of 4-level One Way ANOVA example given earlier, the coefficients of the Effects Coded analysis can be used to compute the cell means in the 2 x 3 factorial design. Recall that in Effect Coding the intercept gives the overall grand mean. Here, this is the overall grand mean for all groups in the analysis (3.100). Each of the other coefficients gives the deviation of the mean of its corresponding group from the grand mean. Thus, for the group receiving Level A1 of Factor 1, the mean is equal to the intercept *plus* the coefficient for O1 (the Effect Code for Factor 1:

Mean for Level A1 = 3.1 + (-1) = 3.0.

The mean for Level A2 is equal to the intercept *minus* the coefficient for O1:

Mean for Level A2= 3.1 - (-1) = 3.2.

For Factor B, the mean for cases in Level B1 is computed equal to the Intercept plus the sum of the coefficients for O2 and O3, or

Mean for Level B1= 3.1 + (.7 - .3) = 3.4.

The means for Levels B2 and B3 are obtained by subtracting coefficients for O2 and O3 respectively:

Mean for Level B2= 3.1 - .7 = 2.4, and

Mean for Level B3= 3.1 – (-.4) = 3.5.

These agree with the marginal means given for the 2 x3 ANOVA design (Table 27). Computing the individual cell means in Table 27 considerably more complicated and will not be addressed here (at this time.)

The analyses yielded the R^2 statistics shown in Table 41. Applying the equation for the F statistic, given earlier. Only the F statistic for the AB interaction, 4.071, is statistically significant at the .05 level. This agrees with the analysis from Vassar Stats.

Table 41. R2 Summaries for the Full and Restricted Models Using Effect Coding

Model	R^2	df	R^2 Change*	F
Full model	0.362	5		
Restricted model for Factor A (Omitting O1)	0.357	4	0.006	0.214
Restricted model for Factor B (Omitting, O2 and O3	0.222	3	0.140	2.643
Restricted model for AB Interaction (Omitting O4 and O5)	0.146	3	0.216	4.071*

*$R^2_{Full} - R^2_{Restricted}$

**p<.05*

Orthogonal Variable Coding in Factorial ANOVA

This section has yet to be added.

APPENDIX A

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

APPENDIX B
Excel version of the Example

Excel Version of the Example Given in "An Introduction to Linear Regression Analysis"

The first thing I did was compute a table of descriptive statistic using Descriptive Statistics in the ToolPac
The values needed for a sufficient table of descriptive statistics are bolded

The Data

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

Score		*X1*		*X2*	
Mean	5.700	Mean	4.350	Mean	5.500
Standard Error	0.612	Standard Error	0.519	Standard Error	0.373
Median	5.500	Median	4.000	Median	6.000
Mode	4.000	Mode	4.000	Mode	6.000
Standard Deviation	2.736	Standard Deviation	2.323	Standard Deviation	1.670
Sample Variance	7.484	Sample Variance	5.397	Sample Variance	2.789
Kurtosis	-1.047	Kurtosis	-0.638	Kurtosis	-0.408
Skewness	0.041	Skewness	0.198	Skewness	0.151
Range	9.000	Range	8.000	Range	6.000
Minimum	1.000	Minimum	1.000	Minimum	3.000
Maximum	10.000	Maximum	9.000	Maximum	9.000
Sum	114.000	Sum	87.000	Sum	110.000
Count	20.000	Count	20.000	Count	20.000

The next thing I did was comute a table of correlations (a correlation matrix) using Correlation in the ToolPac

	Score	*X1*	*X2*
Score	1.000		
X1	0.771	1.000	
X2	0.657	0.522	1.000

I used the formula in for determining the significance of a correlation, given in HyperstatOnline, Chapter 15

t test for sig of r	Score	X1	X2
Score	1.000	3.693	3.693
X1	5.136	1.000	2.598
X2	3.693	2.598	1.000

Then, using a t table, with 20-2 = 18 df, I determined that all three correlations were statiscially significant
My next step was to comput a regression analysis using Regression in the ToolPac

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.826
R Square	0.683
Adjusted R Square	0.646
Standard Error	1.628
Observations	20.000

ANOVA

	df	*SS*	*MS*	*F*	*Sig of F*
Regression	2.000	97.132	48.566	18.319	0.000
Residual	17.000	45.068	2.651		
Total	19.000	142.200			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-0.461	1.285	-0.359	0.724
X1	0.693	0.189	3.676	0.002
X2	0.572	0.262	2.181	0.044

These statistics agree with what I computed using SPSS. The only thing missing are the colinearity statistics, which could have been addressed using the options available in Regression

APPENDIX C

Excell Replication of the VassarStats 2 x 3 ANOVA Example

2 x 3 ANOVA Example

Factor B		
Level B1	Level B2	Level B3
2	3	3
3	2	4
1	2	5
4	1	4
3	3	5
6	4	3
2	4	4
6	2	2
3	1	3
4	2	2

As a first step, I used ANOVA: 2 Factor with Replication. For th Input Range I gave it $A$3:$D$14. This returned the following results

Anova: Two-Factor With Replication

SUMMARY	Level B1	Level B2	Level B3	Total
Level A1				
Count	5.000	5.000	5.000	15.000
Sum	13.000	11.000	21.000	45.000
Average	2.600	2.200	4.200	3.000
Variance	1.300	0.700	0.700	1.571
Level A2				
Count	5.000	5.000	5.000	15.000
Sum	21.000	13.000	14.000	48.000
Average	4.200	2.600	2.800	3.200
Variance	3.200	1.800	0.700	2.171
Total				
Count	10.000	10.000	10.000	
Sum	34.000	24.000	35.000	
Average	3.400	2.400	3.500	
Variance	2.711	1.156	1.167	

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*
Sample	0.300	1.000	0.300	0.214	0.648
Columns	7.400	2.000	3.700	2.643	0.092
Interaction	11.400	2.000	5.700	4.071	0.030
Within	33.600	24.000	1.400		
Total	52.700	29.000			

These results agree with the results from VassarStata.