Presentation of data using
statistical procedures learned in class
(dependent variable = self esteem)

164 ◀ APPLIED MULTIVARIATE RESEARCH

Table 5a.3 Summary of the Example for Multiple Regression

Variable	r	b	beta	Squared Semipartial Correlation	Structure Coefficients	t
Positive affect	.55	2.89	.40	.14	.80	10.61*
Negative affect	−.57	−2.42	−.43	.16	−.82	−11.50*
Openness	.22	.11	.06	.00	.32	1.64
Constant (Y intercept)		56.66				

*p < .01. $R^2 = .48$

bivariate correlation

Unique Contribution to the variance explained

bivariate Correlation between indep. var. and predicted score

(standardized) the weight applied to the variable to maximize least-squares

$.69 = \frac{.55}{\sqrt{.48}} = .80$

also present a correlation matrix of all vars.

uniquely explained (squared semipartial correlation) ...

Material left to cover in semester:

1. Assumptions of regression

2. Diagnostics to determine whether the assumptions are being met

3. Solutions to use when an assumption isn't being met, including variable transformations, interactions, and Dummy Variables

---

Curvilinear transformations,
Dummy Variables,
and
Interactions

Curvilinear transformations

Most common are polynomial transformations which are simply models with the X transformed to a certain "power" (e.g., squared, cubed).

For example, a quadratic polynomial is:

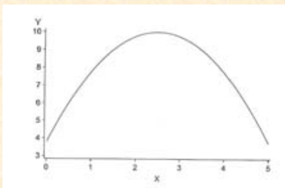
$Y = a + b_1X + b_2X^2$

and a cubic model is:

$Y = a + b_1X + b_2X^2 + b_3X^3$

Note that whenever you add a higher power, you must always include terms for all the lower order powers.

The quadratic polynomial is used when the relationship between the dependent and independent variable has one curve. For example, this might include the relationship between income and age.



The cubic polynomial is used when the relationship between the dependent and independent variable has two curves.

Multiple Regression:

Using nominal variables (creating dummies)

Nominal variables cannot be included in an OLS regression equation in the same way as interval-ratio variables since they are not linear variables.

However, there is a way of "transforming" the nominal variable so that it can be included.

This requires creating a separate variable for each category of the nominal variable (called dummy variables). These can then be included in the regression equation (although one of the newly created "dummy" variables must be left out of the regression equation)

For example: if we want to include the variable "race" in our regression equation and it is coded:

1 = white
2 = black
3 = other

then we would create three variables, one for each category of the variable "race"

To create the new variables:
In SPSS, go to transform, then "recode into another variable"

we then use the dichotomous variable race to create the variable "white" where:

1 = white
0 = black
0 = other

That is, for the new variable, if a person is white he/she will be coded "1" and coded "0" if anything other than white

Similarly we create a new variable called "black" where:

1 = black
0 = white
0 = other

and the variable "other" where:

1 = other
0 = white
0 = black

With these variables created we can then include the concept of "race" into the regression equation by including two of the three new "race" variables in the regression.

The one variable we leave out can be examined and interpreted by viewing the "a" coefficient.

Rationale for leaving out one of the dummy variables:

One variable must be left out so that the regression equation can calculate the regression line.

If all the dummy variables are included in the regression equation, it will not be mathematically possible to create a regression line.

Interpreting Dummy Variables:
Suppose our output shows the following:

$$\text{Job Satisfaction} = 4.02 + .32X_1 + -.18X_2$$

Where: X_1 = Blacks and X_2 = Others and the t values for both are significant (the left out dummy variable is whites)

Job satisfaction for Whites = 4.02 (average)

Blacks job satisfaction on average is significantly higher than whites at 4.34 (4.02 + .32)

Others job satisfaction on average is significantly less than Whites at 3.84 (4.02 - .18)

Interpreting the "*t*" value for dummy variables in the regression equation

The t test associated with a given dummy variable tests for the difference between the mean of the dummy variable in the equation and the mean of the dummy variable left out of the regression equation.

(remember that the t tests for continuous variables tests whether the independent variable significantly increases the variance explained in the dependent variable)

How do we know if the dummy variables significantly reduce error in the dependent variable?

We can use an f test to compare the R^2 when only the continuous variables are included in the regression equation, to the R^2 when the continuous variables and the dummies are included. If there is a significant difference, then the dummy variables have significantly increased the variation explained beyond that of the continuous variables.

Which dummy variable should be omitted?

One choice is to omit that variable that you have the most interest in statistically comparing to those that are included.

It has been found that leaving out a dummy variable with a small number of cases, can create biased regression coefficients.

How is the "a" coefficient interpreted when there are continuous variables in the regression equation?

Note: a regression equation with both dummy variables and continuous variables is referred to as an analysis of covariance (ANCOVA)

The same as if only the dummy variables were included.

In a regression equation with no dummy variables, the "a" coefficent is the average score of the respondents when each of the independent variables equals zero.

When dummy variables are included the "a" coefficient represents the average for the left out dummy variable after controlling for the independent (continuous) variables.

A third type of "transformation" can be thought of as interactions.

In a linear regression model, a one unit increase in X_1 always produces a change of B_1 units in Y.

Now let's suppose the effect of X on Y depends on the value of another independent variable.

For example, the effect of age on income may depend on the person's education. Or, in other words, there is an interaction between age and education in their effects on income.

We could also say that the slope of income on age is steeper for those with more education.

If we want to test our example we could create the following model:

$Y = a + B_1 age + B_2\ educ. + B_3 age*educ.$

This equation has both age and education entered in the usual way, but also has the product of age and educ. as an additional variable.

Once the regression has been performed the first thing to do is exam the p value for the product variable.

If it is significant, you can conclude that there is strong evidence that the effect of age on income depends on the level of education.

Interpreting regression results with an interaction variable

Each variable involved in the interaction variable (called the main-effect variables) has its own b coefficient and they each have a special (and often not very useful) interpretation when an interaction variable is present.

More specifically, we would say that the *b* coefficient for age can be interpreted as the effect of age when education is zero.

$Y = a + b_1 age + b_2\ educ. + b_3 age*educ.$

Similarly, the coefficient for education can be interpreted as the effect of education when age is zero.

Typically, we are not concerned about the significant effects of the two main effects or of their interpretation.

Rather we are interested in interpreting the b for the interaction variable (however the main effect variables should remain in the equation).

The way to interpret the b for the product variable is to calculate the effect of age on income for different values of education.

Mathmatically, the effect of age on income is:

$Income = b_1 age + b_2 educ + (b_3 * education)$

b_3 represents the estimate for the product term; let's look at this example further:

TABLE 8.2 Regression of Income on Schooling, Age, and Their Interaction

Variable	*Coefficient*	*Standard Error*	*p Value*
Intercept	88,159	33,131	
Schooling	−7,649	2,696	.012
Age	−1,770	659	.008
Schooling × Age	207	55.4	.001
R^2	.50		

The effect of age on income for a specific value of education, let's say 9 years, would be:

($b_1$age + (b_3 * education) = Y)

-1,770 + (207 * 9) = $93

What would be the effect of age on income for 12 years of school?

Here are additional effects that have been calculated

TABLE 8.3

Years of Schooling	*Effect of Age on Income*
9	93 – (9*207) –1,770
12	714
14	1,128
16	1,542
20	2,370

b for interaction term

How to interpret interactions with dummy variables

Dummy variables are treated just like continuous variables by creating a product term. However the interpretation is somewhat different.

In Table 8.5, what is the dependent variable and the independent variables?

TABLE 8.5 Regression of Income (Dollars) on Schooling, Marital Status, and Their Interaction

Variable	*Coefficient*	*p Value*
Age	652	.003
Schooling	−980	.57
Married	−48,592	.04
Schooling × Married	3,912	.04
Intercept	8,404	

Interpetation: The coefficient for the product term ($3,912) is the additional effect of schooling when the person is married, so the effect of schooling for married respondents is:

(b_1 Schooling + (b_2 * Married) = Y

-980 + 3,912 = $2,932

Each additional year of schooling brings $2,932 more income for those married. For those not married, income goes down -980 for each year of schooling.

TABLE 8.5 Regression of Income (Dollars) on Schooling, Marital Status, and Their Interaction

Variable	*Coefficient*	*p Value*
Age	652	.003
Schooling	−980	.57
Married	−48,592	.04
Schooling × Married	3,912	.04
Intercept	8,404	

What about the effect of being married?

Interpetation: The coefficient for married (-$48,592) says that among respondents with no education, those who are married make about $49,000 less than those who are unmarried. Further, among those with, lets say, 6 years of schooling, they make about $25,000 less.

b_1 Married + (b_2 * Schooling) = Y

-48,592 + (3,912 * 6) = -25,120

TABLE 8.5 Regression of Income (Dollars) on Schooling, Marital Status, and Their Interaction

Variable	*Coefficient*	*p Value*
Age	652	.003
Schooling	−980	.57
Married	−48,592	.04
Schooling × Married	3,912	.04
Intercept	8,404	

The End

The End.