Chapter 8

Multivariate Regression Analysis

8.3 Multiple Regression with K Independent Variables

8.4 Significance tests of Parameters

Population Regression Model

The principles of bivariate regression can be generalized to a situation of several independent variables (predictors) of the dependent variable

For *K* independent variables, the population regression and prediction models are:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_K X_{Ki} + \varepsilon_i$$

$$\hat{Y}_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_K X_{Ki}$$

The sample prediction equation is:

$$\hat{Y}_i = a + b_1 X_{1i} + b_2 X_{2i} + \ldots + b_K X_{Ki}$$

Predict number of children ever born (Y) to the 2008 GSS respondents (N=1,906) as a linear function of education (X_1), occup'l prestige (X_2), no. of siblings (X_3), and age (X_4):

$$\hat{Y}_i = 1.118 - .080X_{1i} - .001X_{2i} + .0678X_{3i} + .035X_{4i}$$

People with more education and higher-prestige jobs have fewer children, but older people and those raised in families with many siblings have more children.

Use the equation to predict the expected number of kids by a person with X_1 = 12; X_2 = 40; X_3 = 8; X_4 = 55:

$$\hat{Y}_i = 1.118 - .080(12) - .001(40) + .067(8) + .035(55) = \underline{\qquad\qquad}$$

For X_1 = 16; X_2 = 70; X_3 = 1; X_4 = 25:

$$\hat{Y}_i = 1.118 - .080(16) - .001(70) + .067(1) + .035(25) = \underline{\qquad\qquad}$$

OLS Estimation of Coefficients

As with bivariate regression, the computer uses Ordinary Least Squares methods to estimate the intercept (a), slopes (b_{YX}), and multiple coefficient of determination (R^2) from sample data.

OLS estimators minimize the sum of squared errors for the linear prediction:

$$\textbf{min} \sum e_i^2$$

See SSDA#4 Boxes 8.2 and 8.3 for details of best linear unbiased estimator (BLUE) characteristics and the derivations of OLS estimators for the intercept <u>a</u> and slope <u>b</u>

Nested Equations

A set of nested regression equations successively adds more predictors to an equation to observe changes in their slopes with the dependent variable

Predicting children ever born (Y) by adding education (X_1); occupational prestige (X_2); siblings (X_3); age (X_4). (Standard errors in parentheses)

(1)
$$\hat{Y}_i = 3.606 - 0.124X_{1i} \qquad R^2 = 0.051$$
(0.165) (.012)

(2)
$$\hat{Y}_i = 3.473 - 0.133X_{1i} + 0.006X_{2i} \qquad R^2 = 0.052$$
(0.173) (.014) (.003)

(3)
$$\hat{Y}_i = 2.865 - 0.109X_{1i} + 0.006X_{2i} + 0.073X_{3i} \qquad R^2 = 0.066$$
(0.199) (.015) (.003) (.012)

(4)
$$\hat{Y}_i = 1.118 - 0.080X_{1i} - 0.001X_{2i} + 0.067X_{3i} + 0.035X_{4i} \qquad R^2 = 0.193$$
(0.211) (.014) (.003) (.011) (.002)

F-test for ρ^2

The hypothesis pair for the multiple coefficient of determination remains the same as in the bivariate case:

$$H_0: \rho^2 = 0$$
$$H_1: \rho^2 > 0$$

But the F-test must also adjust the sample estimate of R^2 for the *df* associated with the *K* predictors:

$$\mathbf{F}_{\mathbf{K},\mathbf{N}-\mathbf{K}-1} = \frac{\mathbf{MS}_{\mathbf{REGRESSION}}}{\mathbf{MS}_{\mathbf{ERROR}}} = \frac{\mathbf{R}^2/\mathbf{K}}{(1-\mathbf{R}^2)/(\mathbf{N}-\mathbf{K}-1)}$$

As you enter more predictors into the equation in an effort to pump up your R^2, you must pay the higher “cost” of an additional *df* per predictor to get that result.

Test the null hypothesis H_0: $\rho^2 = 0$ for Equation 3:

Source	SS	df	MS	F
Regression	354.7			
Error	5,011.1			
Total	5,365.8		--------------------	

α	df_R, df_E	c.v.
.05	3, ∞	2.60
.01	3, ∞	3.78
.001	3, ∞	5.42

Decision about H_0: ____________________

Prob. Type I error: ____________________

Conclusion: ____

Difference in ρ^2 for Nested Equations

We can also test whether adding predictors to a second, nested regression equation increases ρ^2:

$$H_0: \rho_2^2 - \rho_1^2 = 0$$

$$H_1: \rho_2^2 - \rho_1^2 > 0$$

where subscripts "1" and "2" refer to the equations with fewer and more predictors, respectively

The F-statistic tests whether adding predictors increases the population rho-square, relative to the difference in the two nested equations' degrees of freedom:

$$\mathbf{F}_{(\mathbf{K}_2 - \mathbf{K}_1),(\mathbf{N} - \mathbf{K}_2 - 1)} = \frac{(\mathbf{R}_2^2 - \mathbf{R}_1^2) / (\mathbf{K}_2 - \mathbf{K}_1)}{(1 - \mathbf{R}_2^2) / (\mathbf{N} - \mathbf{K}_2 - 1)}$$

Is the ρ^2 for Eq. 2 larger than the ρ^2 for Eq. 1?

$$F_{(2-1),(2648-2-1)} = \frac{(R_2^2 - R_1^2)\,/\,(K_2 - K_1)}{(1 - R_2^2)\,/(N - K_2 - 1)} =$$

α	df_R, df_E	c.v.
.05	1, ∞	3.84
.01	1, ∞	6.63
.001	1, ∞	10.83

____

Decision:____________________

Prob. Type I error: ___________

Interpretation: Adding occupation to the regression equation with education did not significantly increase the explained variance in number of children ever born.
In the population, the two coefficients of determination are equal; each explains about 5% of the variance of Y.

Now test the difference in ρ^2 for Eq. 4 versus Eq. 3:

$$F_{(4-3),(2648-4-1)} = \frac{(R_4^2 - R_3^2)/(K_4 - K_3)}{(1 - R_4^2)/(N - K_4 - 1)} =$$

α	df_R, df_E	c.v.
.05	1, ∞	3.84
.01	1, ∞	6.63
.001	1, ∞	10.83

____

Decision:______________________

Prob. Type I error: _____________

Interpretation: Adding age to the regression equation with three other predictors greatly increases the explained variance in number of children ever born. The coefficient of determination for equation #4 seems to be almost three times larger than for equation #3.

Adjusting R^2 for *K* predictors

The meaning of the multiple regression coefficient of determination is identical to the bivariate case:

$$R^2_{YX} = \frac{\sum (Y_i - \bar{Y})^2 - \sum (Y_i - \hat{Y}_i)^2}{\sum (Y_i - \bar{Y})^2}$$

$$R^2_{YX} = \frac{SS_{TOTAL} - SS_{ERROR}}{SS_{TOTAL}} = \frac{SS_{REGRESSION}}{SS_{TOTAL}}$$

However, when you report the sample estimate of a multiple regression R^2, you must <u>adjust</u> its value by 1 degree of freedom for <u>each</u> of the *K* predictors:

$$\mathbf{R}^2_{\mathbf{adj}} = \mathbf{R}^2 - \left(\frac{(\mathbf{K})(1 - \mathbf{R}^2)}{(\mathbf{N} - \mathbf{K} - 1)} \right)$$

For large sample *N* and low R^2, not much will change.

Adjust the sample R^2 for each of the four nested equations (N = 1,906):

Eq.	R^2	K	Adj. R^2
1:	0.051	1	
2:	0.052	2	
3:	0.066	3	
4:	0.193	4	

Here are those four nested regression equations again with the number of ever-born children as the dependent variable. Now we'll examine their regression slopes.

Predict children ever born (Y) by adding education (X_1); occupational prestige (X_2); siblings (X_3); age (X_4) (Standard errors in parentheses)

(1)

$$\hat{Y}_i = 3.606 - 0.124 X_{1i} \qquad R^2 = 0.051$$

(0.165) (.012)

(2)

$$\hat{Y}_i = 3.473 - 0.133 X_{1i} + 0.006 X_{2i} \qquad R^2 = 0.052$$

(0.173) (.014) (.003)

(3)

$$\hat{Y}_i = 2.865 - 0.109 X_{1i} + 0.006 X_{2i} + 0.073 X_{3i} \qquad R^2 = 0.066$$

(0.199) (.015) (.003) (.012)

(4)

$$\hat{Y}_i = 1.118 - 0.080 X_{1i} - 0.001 X_{2i} + 0.067 X_{3i} + 0.035 X_{4i} \qquad R^2 = 0.193$$

(0.211) (.014) (.003) (.011) (.002)

Interpreting Nested b_{yx}

The multiple regression slopes are partial or net effects. When other independent variables are statistically "held constant," the size of b_{YX} often decreases. These changes occur if predictor variables are correlated with each other as well as with the dependent variable.

Two correlated predictors divide their joint impact on the dependent variable between both b_{yx} coefficients.

For example, age and education are negatively correlated ($r = -.17$): older people have less schooling. When age was entered into equation #4, the net effect of education on number of children decreased from $b_1 = -.124$ to $b_1 = -.080$. So, controlling for respondent's age, an additional year of education decreases the number of children ever born by a much smaller amount.

t-test for Hypotheses about β

t-test for hypotheses about K predictors uses familiar procedures

A hypothesis pair about the population regression coefficient for *j*th predictor could have a two-tailed hypothesis:

$$H_0: \beta_j = 0$$
$$H_1: \beta_j \neq 0$$

Or, a hypothesis pair could indicate the researcher's expected direction (sign) of the regression slope:

$$H_0: \beta_j \leq 0$$
$$H_1: \beta_j > 0$$

Testing an hypothesis about β_j uses a t-test with N-K-1 degrees of freedom (i.e., a Z-test for a large sample)

$$t_{N\text{-}K\text{-}1} = \frac{b_j - \beta_j}{s_{b_j}}$$

where b_j is the sample regression coefficient & denominator is the standard error of the sampling distribution of β_j

(see formula in SSDA#4, p. 266)

Here are two hypotheses, about education (β_1) and occupational prestige (β_2), to be tested using Eq, 4:

Test a two-tail hypothesis about $\boldsymbol{\beta_1}$:

α	1-tail	2-tail
.05	**1.65**	**±1.96**
.01	**2.33**	**±2.58**
.001	**3.10**	**±3.30**

$t_{2648-4-1} =$ ______________________________

Decision: ____________________ Prob. Type I error: ___________

Test a one-tail hypothesis about $\boldsymbol{\beta_2}$:

$t_{2648-4-1} =$ ______________________________

Decision: ____________________ Prob. Type I error: ___________

Test one-tailed hypotheses about expected positive effects siblings (β_3) and age (β_4) on number of children ever born:

$t_{2648-4-1} =$ ____________________

Decision: ________________ Prob. Type I error: _________

$t_{2648-4-1} =$ ____________________

Decision: ________________ Prob. Type I error: _________

Interpretation: These sample regression statistics are very unlikely to come from a population whose regression parameters are zero ($\beta_j = 0$).

Standardizing regression slopes (β*)

Comparing effects of predictors on a dependent variable is difficult, due to differences in units of measurement

Beta coefficient (β*) indicates effect of an X predictor on the Y dependent variable in standard deviation units

$$\beta^*_{YX_i} = b_{YX_i}\left(\frac{s_{X_i}}{s_Y}\right)$$

1. Multiply the b_{YX} for each X_i by that predictor's standard deviation
2. Divide by the standard deviation of the dependent variable, Y

The result is a standardized regression equation, written with *Z*-score predictors, but no intercept term:

$$\hat{Z}_Y = \beta^*_1 Z_1 + \beta^*_2 Z_2 + \ldots + \beta^*_K Z_K$$

Standardize the regression coefficients in Eq. 4

$$\hat{Y}_i = 1.118 - 0.080X_{1i} - 0.001X_{2i} + 0.07X_{3i} + 0.035X_{4i}$$

Use these stnd. devs. to change all the b_{YX} to β^*:

Variable	s.d.
Y Children	1.70
X_1 Educ.	3.08
X_2 Occup.	13.89
X_3 Sibs	3.19
X_4 Age	17.35

$$(\mathbf{X}_1): \quad \boldsymbol{\beta}^*_{\mathbf{YX}_1} = -0.080\left(\frac{3.08}{1.70}\right) = \_\_\_\_\_\_$$

$$(\mathbf{X}_2): \quad \boldsymbol{\beta}^*_{\mathbf{YX}2} = -0.001\left(\frac{13.89}{1.70}\right) = \_\_\_\_\_\_$$

$$(\mathbf{X}_3): \quad \boldsymbol{\beta}^*_{\mathbf{YX}_3} = +0.067\left(\frac{3.19}{1.70}\right) = \_\_\_\_\_\_$$

$$(\mathbf{X}_4): \quad \boldsymbol{\beta}^*_{\mathbf{YX}_4} = +0.035\left(\frac{17.35}{1.70}\right) = \_\_\_\_\_\_$$

Write the standardized equation:

$$\hat{Z}_Y = -0.14Z_1 - 0.01Z_2 + 0.13Z_3 + 0.36Z_4$$

Interpreting β^*

Standardizing regression slopes transforms predictors' effects on the dependent variable from their original measurement units into standard-deviation units. Hence, you must interpret and compare the β^* effects in standardized terms:

Education $\beta^* = -0.14$ ⇨ a 1-standard deviation difference in education levels <u>reduces</u> the number of children born by one-seventh st. dev.

Occupational $\beta^* = -0.01$ ⇨ a 1-standard deviation difference in prestige <u>reduces</u> N of children born by one-hundredth st. dev.

Siblings $\beta^* = +0.13$ ⇨ a 1-standard deviation difference in siblings <u>increases</u> the number of children born by one-eighth st. dev.

Age $\beta^* = +0.36$ ⇨ a 1-standard deviation difference in age <u>increases</u> the number of children born by more than one-third st. dev.

Thus, age has the largest effect on number of children ever born; occupation has the smallest impact (and it's not significant)

Let's interpret a standardized regression, where annual church attendance is regressed on X_1 = religious intensity (a 4-point scale), X_2 = age, and X_3 = education:

$$\hat{Y}_i = -20.21 + 12.13X_{1i} + 0.12X_{2i} + 0.09X_{3i} \qquad R^2_{adj} = 0.269$$

$$(3.05) \quad (0.50) \qquad (0.03) \qquad (0.17)$$

The standardized regression equation:

$$\hat{Z}_i = +0.50Z_{1i} + 0.08Z_{2i} + 0.01Z_{3i}$$

Interpretations:

- ✓ Only two predictors significantly increase church attendance
- ✓ The linear relations explain 26.9% of attendance variance
- ✓ Religious intensity has strongest effect (1/2 std. deviation)
- ✓ Age effect on attendance is much smaller (1/12 std. dev.)

Dummy Variables in Regression

Many important social variables are not continuous but measured as discrete categories and thus cannot be used as independent variables without recoding

Examples of such variables include gender, race, religion, marital status, region, smoking, drug use, union membership, social class, college graduation

Dummy variable coded "1" to indicate the presence of an attribute and "0" its absence

1. Create & name one dummy variable <u>for each of the K categories</u> of the original discrete variable

2. For each dummy variable, code a respondent "1" if s/he has that attribute, "0" if lacking that attribute

3. Every respondent will have a "1" for only one dummy, and "0" for the K-1 other dummy variables

GSS codes for SEX are arbitrary: 1 = Men & 2 = Women

Recode SEX as two new dummies ⇨	**MALE**	**FEMALE**
1 = Men	**1**	**0**
2 = Women	**0**	**1**

MARITAL five categories from 1 = Married to 5 = Never

MARITAL ⇨	**MARRYD**	**WIDOWD**	**DIVORCD**	**SEPARD**	**NEVERD**
1 = Married	**1**	**0**	**0**	**0**	**0**
2 = Widowed	**0**	**1**	**0**	**0**	**0**
3 = Divorced	**0**	**0**	**1**	**0**	**0**
4 = Separated	**0**	**0**	**0**	**1**	**0**
5 = Never	**0**	**0**	**0**	**0**	**1**

SPSS RECODE to create K dummy variables (1-0) from MARITAL

The ORIGINAL 2008 GSS FREQUENCIES:

marital MARITAL STATUS

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 MARRIED	972	48.0	48.2	48.2
	2 WIDOWED	164	8.1	8.1	56.3
	3 DIVORCED	281	13.9	13.9	70.2
	4 SEPARATED	70	3.5	3.5	73.7
	5 NEVER MARRIED	531	26.2	26.3	100.0
	Total	2018	99.8	100.0	
Missing	9 NA	5	.2		
Total		2023	100.0		

RECODE STATEMENTS:

```
COMPUTE marryd=0.
COMPUTE widowd=0.
COMPUTE divord=0.
COMPUTE separd=0.
COMPUTE neverd=0.
IF (marital EQ 1) marryd=1.
IF (marital EQ 2) widowd=1.
IF (marital EQ 3) divord=1.
IF (marital EQ 4) separd=1.
IF (marital EQ 5) neverd=1.
```

Every case is coded 1 on one dummy variable and 0 on the other four dummies. The MARITAL category frequencies above appear in the “1” row for the five marital status dummy variables below:

RECODE	MARRYD	WIDOWD	DIVORD	SEPARD	NEVERD
1	972	164	281	70	531
0	1,046	1,854	1,737	1,948	1,487
TOTAL	2,018	2,018	2,018	2,018	2,018

Linear Dependency among Dummies

Given *K* dummy variables, if you know a respondent's codes for *K* - 1 dummies, then you <u>also know</u> that person's code for the *K*th dummy!

This **linear dependency** is similar to the degrees of freedom problem in ANOVA.

Thus, to use a set of *K* dummy variables as predictors in a multiple regression equation, you must omit one of them. Only K-1 dummies can be used in an equation.

The omitted dummy category serves as the **reference category** (or baseline), against which to interpret the *K*-1 dummy variable effects (*b*) on the dependent variable

Use four of the five marital status dummy variables to predict annual sex frequency in 2008 GSS. WIDOWD is the omitted dummy, serving as the reference category.

$$\hat{Y}_i = 8.8 + 52.4\, D_{MARR} + 32.8\, D_{DIV} + 21.1\, D_{SEP} + 53.0\, D_{NEVER} \qquad R^2_{adj} = 0.054$$

$$(5.5)\quad (6.0) \qquad\qquad (6.9) \qquad\qquad (10.3) \qquad\qquad (6.3)$$

Widows are coded "0" on all four dummies, so their prediction is:

$$\hat{Y}_i = 8.8 + 52.4\,(0) + 32.8\,(0) + 21.1\,(0) + 53.0\,(0) = \_\_\_\_\_ \quad per\ year$$

Married: $$\hat{Y}_i = 8.8 + 52.4\,(1) + 32.8\,(0) + 21.1\,(0) + 53.0\,(0) = \_\_\_\_\_ \quad per\ year$$

Divorced: $$\hat{Y}_i = 8.8 + 52.4\,(0) + 32.8\,(1) + 21.1\,(0) + 53.0\,(0) = \_\_\_\_\_ \quad per\ year$$

Separated: $$\hat{Y}_i = 8.8 + 52.4\,(0) + 32.8\,(0) + 21.1\,(1) + 53.0\,(0) = \_\_\_\_\_ \quad per\ year$$

Never: $$\hat{Y}_i = 8.8 + 52.4\,(0) + 32.8\,(0) + 21.1\,(0) + 53.0\,(1) = \_\_\_\_\_ \quad per\ year$$

Which persons are the least sexually activity? Which the most?

ANCOVA

Analysis of Covariance (ANCOVA) equation has <u>both</u> dummy variable and continuous predictors of a dependent variable

Marital status is highly correlated with age (widows are older, never marrieds are younger), and annual sex activity falls off steadily as people get older.

Look what happens to the marital effects when age is controlled, by adding AGE to the marital status predictors of sex frequency:

$$\hat{Y}_i = 127.2 + 15.5\,D_{MARR} + 0.1\,D_{DIV} - 23.4\,D_{SEP} - 10.4\,D_{NEVER} - 1.7\,X_{AGE} \qquad R^2_{adj} = 0.172$$

$$(9.2) \quad (6.1) \qquad (6.9) \qquad (10.1) \qquad (7.2) \qquad (0.1)$$

Each year of age reduces sex by –1.7 times per year.

Among people of same age, marrieds have more sex than others, but never marrieds now have <u>less</u> sex than widows!

What would you predict for: Never marrieds aged 22? Marrieds aged 40? Widows aged 70?

Add FEMALE dummy to regression of church attendance on X_1 = religious intensity, X_2 = age, and X_3 = education:

$$\hat{Y}_i = -20.92 + 11.96X_{1i} + 0.10X_{2i} + 0.09X_{3i} + 2.20D_{FEMi} \qquad R^2_{adj} = 0.270$$

$$(3.06) \quad (0.50) \qquad (0.03) \qquad (0.17) \quad (1.05)$$

The standardized regression equation:

$$\hat{Z}_i = +0.49Z_{1i} + 0.08Z_{2i} + 0.01Z_{3i} + 0.04D_{FEMi}$$

Interpretations:

- Women attend church 2.20 times more per year than men
- Other predictors' effects unchanged when gender is added
- Age effect is twice as larger as gender effect
- Religious intensity remains strongest predictor of attendance