

SPSS Analysis and Multivariate Statistics

Multivariate analyses are invariably done by computer because computations are complex. To illustrate computer analyses for three multivariate techniques, we will continue to use the example described in Chapters 17 and 18, involving a prenatal intervention for low-income pregnant young women. Data for these analyses are shown in a screenshot from a data file from the Statistical Package for the Social Sciences (SPSS),¹ reproduced here as Figure 1.

EXAMPLE OF MULTIPLE REGRESSION

In the Supplement to Chapter 18, we tested the hypothesis that older mothers in the sample had infants with higher birth weights than younger mothers, using Pearson's r . The calculated value of r (.594) was highly significant, supporting the research hypothesis.

Suppose that we wanted to test whether we could significantly *improve* our ability to predict infant birth weight by adding two predictor variables in a multiple regression. The two additional predictors are whether the mother smoked while pregnant (coded 1 for yes, 0 for no), and her number of prior pregnancies. The SPSS commands for running a regression analysis are:

ANALYZE → REGRESSION → LINEAR

Figure 2 presents part of the SPSS printout for a multiple regression analysis in which infant birth weight is the outcome variable and maternal age (AGE), smoking status (SMOKE), and number of prior pregnancies (PRIORS) are predictor variables. We will explain a few noteworthy aspects of this printout. Values that we mention in the text are shaded on the printout.

Panel A of Figure 2 (labeled “Variables Entered/Removed”) shows that we used hierarchical regression to predict the outcome, infant birth weight (footnote b). Mother’s age was entered first (Model 1), and then smoking status and prior pregnancies were entered in a second block (Model 2).

Panel B (“Model Summary”) indicates that, in Model 1, $R = .594$ —the same as the bivariate correlation shown in Figure 3 of the Chapter 18 Supplement. The value of R^2 is $.353$ ($.594^2$), which represents the proportion of variance in birth weight accounted for by mother’s age. The **adjusted R^2** of $.330$ in Model 1 is the R^2 after it has been adjusted to reflect more closely the goodness of fit of the regression model in the population, through a formula that involves sample size and number of predictors. Next, the *standard error of the estimate* (8.9702) is shown. (The standard error of estimate is a measure of the accuracy of predictions that we do not describe further, except to note that smaller values are preferred to larger ones.) The next few columns present information about changes to R^2 . In Model 1, the change is from

¹SPSS Version 21 was used to create all output.

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons							
	ID	GROUP	AGE	PRIORS	SMOKE	BWEIGHT	REPEAT
1	1	1	17	1	1	107	1
2	2	1	14	0	0	101	0
3	3	1	21	3	0	119	0
4	4	1	20	2	0	128	1
5	5	1	15	1	1	89	0
6	6	1	19	0	1	99	0
7	7	1	19	1	0	111	0
8	8	1	18	1	1	117	1
9	9	1	17	0	0	102	1
10	10	1	20	0	0	120	0
11	11	1	13	0	1	76	0
12	12	1	18	0	1	116	0
13	13	1	16	0	0	100	1
14	14	1	18	0	0	115	0
15	15	1	21	2	1	113	0
16	16	2	19	0	0	111	1
17	17	2	21	1	0	108	0
18	18	2	19	2	1	95	0
19	19	2	17	0	1	99	0
20	20	2	19	0	0	103	1
21	21	2	15	0	1	94	0
22	22	2	17	1	0	101	1
23	23	2	21	2	0	114	0
24	24	2	20	1	0	97	0
25	25	2	18	0	1	99	1
26	26	2	18	0	1	113	0
27	27	2	19	1	0	89	0
28	28	2	20	0	0	98	0
29	29	2	17	0	0	102	0
30	30	2	19	1	1	105	0

NOTES:

- GROUP: Group status, 1 = Experimental group 2 = Control group
- AGE: Mother's age in years
- PRIORS: Number of prior pregnancies
- SMOKE: Mother's smoking status, 1 = Smokes 0 = Does not smoke
- BWEIGHT: Infant's birth weight, in ounces
- REPEAT: Had repeat pregnancy within 18 months, 1 = Yes 0 = No

FIGURE 1 Fictitious Dataset for Intervention Study with Low-Income Pregnant Adolescents (Screenshot of an SPSS Data File).

Regression

A Variables Entered/Removed^b

Model	Variables Entered	Variables Removed	Method
1	Mother's age ^a	.	Enter
2	Smoking status, No. of prior pregnancies ^a	.	Enter

a. All requested variables entered.

b. Dependent Variable: Infant birth weight in ounces

B Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1					.353		1	28	
2		.358	.284	9.26997	.005		2	26	

a. Predictors: (Constant), Mother's age

b. Predictors: (Constant), Mother's age, Smoking status, No. of prior pregnancies

C ANOVA^c

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1227.283	1	1227.283		
	Residual	2253.017	28	80.465		
	Total	3480.300	29			
2	Regression	1246.061	3	415.354		
	Residual	2234.239	26	85.932		
	Total	3480.300	29			

a. Predictors: (Constant), Mother's age

b. Predictors: (Constant), Mother's age, Smoking status, No. of prior pregnancies

c. Dependent Variable: Infant birth weight in ounces

D Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	48.040	14.600		3.290	.003	18.133	77.947
	Mother's age	3.119	.799	.594	3.905	.001	1.483	4.755
2	(Constant)	52.170	18.398		2.836	.009	14.354	89.987
	Mother's age	2.916	1.021	.555	2.855	.008	.817	5.016
	No. of prior pregnancies	.394	2.410	.030	.163	.872	-4.561	5.349
	Smoking status	-1.643	3.610	-.076	-.455	.653	-9.064	5.778

a. Dependent Variable: Infant birth weight in ounces

E Excluded Variables^b

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics
						Tolerance
1	No. of prior pregnancies	.020 ^a	.108	.915	.021	.727
	Smoking status	-.072 ^a	-.446	.659	-.086	.910

a. Predictors in the Model: (Constant), Mother's age.

b. Dependent Variable: Infant birth weight in ounces.

FIGURE 2 SPSS Printout for Hierarchical Multiple Regression of Infant Birth weight on Three Predictors.

0.0 to .353—which yields an F (15.252) that, with $df = 1$ and 28, is significant ($p = .001$). In Model 2, after adding SMOKE and PRIORS, the value of R is higher (.598), but the F for change (.109) with 2 and 26 df is not significant ($p = .897$).

Panel C of Figure 2 (“ANOVA”) shows, for Model 1, the F -ratio in which variability due to regression (for the relationship between birth weight and age) is contrasted with residual variability. Again, the printout shows that the value of F (15.252) with 1 and 28 df is significant at $p = .001$. The information for Model 2 is for all *three* predictors that are in the model when variables in the second block are entered. Here, the value of F (4.834) with 3 and 26 df is statistically significant at $p = .008$.

Regression equations are presented in Panel D (“Coefficients”). If we wanted to predict new values of birth weight based on maternal age at birth, the equation from Model 1 would be:

$$\text{Birth weight}' = (3.119 \times \text{Age}) + 48.040$$

The predicted birth weights in a new sample of young mothers would equal the regression coefficient ($b = 3.119$) times the value of maternal age (X_1), plus the value of the intercept constant ($a = 48.040$). When values of b are divided by the standard error (.799 for maternal age in Model 1), the result is a t statistic, which indicates the significance of each predictor. In Model 1, $t = 3.905$, which is significant ($p = .001$). The standardized Beta weight (β) for maternal age is .594. In the far right we see that the 95% CI for the regression coefficient b is 1.483 and 4.755—we can be 95% confident that the true coefficient is in this interval. We can see from this confidence interval that the value of b is statistically significant because it does not include 0.0.

If we wanted to use all three predictors to predict infant birth weight, the equation for Model 2 shows the b coefficients for each of the three predictors. In model 2, neither number of prior pregnancies nor smoking status is significant: $p = .872$ and .653, respectively. The 95% CI for both these predictors *does* include 0.0. When we compare the standardized coefficients (the β s) for the three variables, we see that Beta for maternal

age is substantial (.555), while those for the other two predictors are negligible (.030 and $-.076$).

Panel E (“Excluded Variables”) shows the two predictors that were not yet in the equation in Model 1, i.e., number of prior pregnancies and smoking status. The printout shows that the t values associated with the regression coefficients for the two predictors are both nonsignificant ($p = .915$ and .659, respectively), once variation due to maternal age is taken into account. This reinforces what we have already learned—that neither of the two predictors in block 2 would add significantly to the prediction of birth weight, over and above what was achieved with maternal age.

An additional piece of information in Panel E concerns **multicollinearity**, which is a problem that can occur when predictors are too highly intercorrelated. When multicollinearity is present, the computations required for regression coefficients are compromised and results tend to be unstable. Multicollinearity can be diagnosed by computing an index of **tolerance**. If predictors are totally uncorrelated, tolerance is 1.0, and if they are perfectly intercorrelated, tolerance is 0.0. Thus, higher values are more desirable. The computer can be instructed to exclude predictors whose tolerance falls below a specified level (e.g., .10). In our example, tolerance values of .727 and .910 for the two variables not yet in the model are acceptable.

EXAMPLE OF ANALYSIS OF COVARIANCE

In the Supplement to Chapter 18, we tested the hypothesis that infants in the experimental group would have higher birth weights than infants in the control group, using a t -test. The computer calculated t to be 1.44, which was nonsignificant with 28 df . The research hypothesis was therefore rejected.

Through ANCOVA, we can test the same hypothesis controlling for maternal age, which, as we have just seen, is significantly correlated with birth weight. The SPSS commands for an analysis of covariance are:

ANALYZE → GENERAL LINEAR MODEL → UNIVARIATE

Figure 3 presents the printout for ANCOVA for this analysis, with birth weight as the outcome variable, maternal age (AGE) as the covariate, and GROUP (experimental vs. control) as the independent variable. Panel A (“Between Subjects Factors”) shows that the treatment group variable involves 15 mothers in the experimental group and 15 control group mothers. Panel B (“Descriptive Statistics”) presents means and *SDs* for the infant birth weights in the two groups and in the overall sample of 30 mothers. As this panel shows, the mean for the experimental group was 107.5 ounces, compared to a mean of 101.9 ounces for those in the control group.

In Panel C (“Tests of Between-Subjects Effects”), we see that the *F*-value for the overall model is highly significant (14.088, $p = .000$). The value of *F* for the covariate AGE is 24.358, significant at the .000 level (i.e., beyond the .001 level). The value of partial (adjusted) eta squared for age (i.e., the effect size) is .474, and the observed power to detect this effect, for $\alpha = .05$, is quite high, .997. After controlling for age, the *F*-value for the independent variable Group is 8.719, which is significant at the .006 level. In other words, once AGE is controlled, the research hypothesis about experimental versus control differences in infant birth weight is supported rather than rejected. Moreover, the effect size for the intervention (.244) is fairly high. The unadjusted R^2 for predicting birth weight (footnote a), based on both Age and Group, is .511—substantially more than the R^2 between maternal age and birth weight alone (.352).

Panel D (“Estimated Marginal Means: Grand Mean”) shows the overall mean (104.70) for the sample, the standard error (1.45), and the 95% confidence interval for the estimated population mean (95% CI = 101.725, 107.675). Finally, Panel E (“Treatment Group”) shows group means *after they are adjusted for maternal age*. The original, unadjusted means for the experimental and control groups were 107.53 and 101.87, respectively (Panel B). After adjusting for maternal age, however, the experimental mean is 109.08, and the control mean is 100.32, a more sizable difference.

Note that the reason that controlling maternal age affected the comparison between the two

groups is that, despite random assignment, those in the experimental group were younger than those in the control group (means of 17.7 and 18.6, respectively). Since younger maternal age is a risk factor for lower birth weight, controlling maternal age through ANCOVA changed the conclusions of the statistical test.

EXAMPLE OF LOGISTIC REGRESSION

The output for a logistic regression analysis in SPSS is complex, and so we include only a few key panels from an analysis in which we used three predictors—smoking status (SMOKE), maternal age (AGE), and number of prior pregnancies (PRIORS)—to predict whether the young mother had a repeat pregnancy within 18 months of delivering the index infant. The SPSS commands for doing a logistic regression analysis are:

ANALYZE → REGRESSION → BINARY LOGISTIC

Panel A of Figure 4 (“Omnibus Test of Model Coefficients”) presents the results of the chi-square goodness of fit test for the overall model with three predictors (i.e., the test based on the likelihood ratio). The chi-square of 1.107 ($df = 3$) was not significant, $p = .775$, suggesting that the null hypothesis should be retained. In Panel B (“Model Summary”), we see that $-2LL = 35.544$, and that the value of Nagelkerke R^2 is .051. In Panel C (“Hosmer and Lemeshow Test”), the Hosmer–Lemeshow test was nonsignificant ($p = .587$), a result that conflicts with the earlier chi-square goodness of fit test in terms of our ability to infer that the model was adequate in predicting likelihood of a repeat pregnancy. The next two panels suggest that the model is, in fact, disappointing. Panel D (“Classification Table”) indicates how cases would have been classified as having or not having a repeat pregnancy, based on the logistic regression prediction equation (“predicted”), compared to the women’s actual status (“observed”). Only 66.7% of cases were correctly classified, and the model did not predict a repeat pregnancy for a single case where one was actually observed.

Univariate Analysis of Variance

A Between-Subjects Factors

	Value Label	N
Treatment group 0	Control	15
1	Experimental	15

B Descriptive Statistics

Dependent Variable: Infant birth weight in ounces

Treatment group	Mean	Std. Deviation	N
Control	101.8667	7.23944	15
Experimental	107.5333	13.37838	15
Total	104.7000	10.95492	30

C Tests of Between-Subjects Effects

Dependent Variable: Infant birth weight in ounces

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Observed Power ^b
Corrected Model	1777.228 ^a	2	888.614	14.088	.000	.511	.997
Intercept	572.734	1	572.734	9.080	.006	.252	.828
Age	1536.395	1	1536.395	24.358	.000	.474	.997
Group	549.945	1	549.945	8.719	.006	.244	.812
Error	1703.072	27	63.077				
Total	332343.000	30					
Corrected Total	3480.300	29					

a. R Squared = .511 (Adjusted R Squared = .474).

b. Computed using alpha = .05.

Estimated Marginal Means

D 1. Grand Mean

Dependent Variable: Infant birth weight in ounces

Mean	Std. Error	95% Confidence Interval	
		Lower Bound	Upper Bound
104.700 ^a	1.450	101.725	107.675

a. Covariates appearing in the model are evaluated at the following values: Mother's age = 18.1667.

E 2. Treatment group

Dependent Variable: Infant birth weight in ounces

Treatment group	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Control	100.320 ^a	2.074	96.063	104.576
Experimental	109.080 ^a	2.074	104.824	113.337

a. Covariates appearing in the model are evaluated at the following values: Mother's age = 18.1667.

FIGURE 3 SPSS Printout for ANCOVA of Group Differences in Infant Birth weight, Controlling Maternal Age.

Logistic Regression
Block 1: Method = Enter

A Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	1.107	3	.775
	Block	1.107	3	.775
	Model	1.107	3	.775

B Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	35.544	.036	.051

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

C Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	6.540	8	.587

D Classification Table^{a,b}

Observed			Predicted		
			Repeat pregnancy		
			No (0)	Yes (1)	Percentage Correct
Step 1 Repeat pregnancy	No (0)	20	1	95.2	
	Yes (1)	9	0	.0	
Overall Percentage				66.7	

a Constant is included in the model.

b The cut value is .500

E Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I. for EXP(B)	
							Lower	Upper
Step 1 ^a Age	-.131	.239	.298	1	.585	.878	.549	1.402
Priors	-.097	.593	.027	1	.870	.907	.284	2.901
Smoke	-.781	.895	.762	1	.383	.458	.079	2.645
Constant	1.891	4.283	.195	1	.659	6.628		

^aVariable(s) entered on step 1: AGE, PRIORS, SMOKE.

FIGURE 4 Partial SPSS Printout for Logistic Regression Predicting an Early Repeat Pregnancy.

Panel E (“Variables in the Equation”) shows the logistic regression equation—i.e., the value of the *b* weights and the constant—in the column headed B. The Wald statistics indicate that none of the three predictors was significant. For example, the Wald statistic for predicting a repeat pregnancy based on

the number of prior pregnancies (Priors) was .027, *p* = .870. The column headed Exp(B) is a particularly important one—the values in this column are the adjusted odds ratios associated with each predictor. The OR of .458 for the smoking status variable, for example, indicates that those women

19-8 • PART 3 Designing and Conducting Quantitative Studies to Generate Evidence for Nursing

in this sample who smoked were 46% less likely to have a repeat pregnancy, with age and prior pregnancies controlled. However, this *OR* was not significant. The 95% CI extends from .079 to 2.645,

and, because this encompasses the value of 1.0, this means that we cannot reject the null hypothesis that smokers were as likely as nonsmokers to have a repeat pregnancy.