

Statistics Lab Reader SPSS

Kevin H. Kim

February 11, 2004

Contents

1	Data Entry	1
1.1	Introduction	1
1.2	Defining Variables	1
2	Data Management	5
2.1	Introduction	5
2.2	Transform Menu	6
2.2.1	Recoding Variables	6
2.2.2	Creating a Composite Variable	8
2.2.3	Categorizing a Variable	9
2.3	Small Set	9
2.3.1	Use Sets	10
2.3.2	Create Smaller Dataset	10
2.4	Select Cases	11
2.5	Split File	11
3	Data Screening	15
3.1	Descriptive Information	15
3.1.1	SPSS Explore	15
3.1.2	Example Output of the Explore	16
3.2	Linearity and Homoscedasticity	18
3.3	Detecting Multivariate Outliers	19
3.4	What if the data is not Normal?	20
3.4.1	Robustness of a method	20
3.4.2	Transformation	21
4	Matrix Command in SPSS	25
4.1	Introduction	25

4.2	Defining a Matrix	26
4.2.1	from syntax	26
4.2.2	from SPSS data window	26
4.3	Simple Computation	27
4.4	Generating output	28
5	Multiple Regression	31
5.1	Assumptions	31
5.1.1	Dependent Variable	31
5.1.2	Independent Variables	31
5.1.3	Relationship between DV and IV	32
5.2	Setup in SPSS	32
5.3	Output in SPSS	35
5.3.1	Model summary	35
5.3.2	ANOVA table	36
5.3.3	Regression Coefficients	37
5.3.4	Collinearity Diagnostics	37
6	Regression Diagnostics	39
6.1	Outliers	39
6.1.1	Standardized Residuals (ZRESID)	40
6.1.2	Studentized Residual (SRESID)	40
6.1.3	Studentized Deleted Residuals (SDRESID)	41
6.2	Influence Analysis	42
6.2.1	Leverage	42
6.2.2	Cook's D	43
6.2.3	DFBETA	43
6.2.4	Standardized STDFBETA	44
7	Between-Subjects Analysis of Variance	45
7.1	Assumptions	45
7.2	One-way Between-Subjects ANOVA	46
7.3	Two-way Between-Subjects ANOVA	47
7.3.1	Example of the Design	47
7.3.2	Setup in SPSS	47
7.3.3	Output in SPSS	47
7.3.4	Research hypotheses and How to test them	48
7.4	Contrast Matrices	54

7.4.1	Deviation Coding	54
7.4.2	Difference Contrast	54
7.4.3	Helmert Contrast	54
7.4.4	Simple Contrast	55
7.4.5	Repeated	55
8	Within and Mixed Analysis of Variance	57
8.1	Assumptions	57
8.2	Within-Subjects ANOVA	58
8.3	Mixed ANOVA	58
8.3.1	Example of the Design	58
8.3.2	Setup in SPSS	58
8.3.3	Output in SPSS	60
8.3.4	Research hypotheses and How to test them	61
9	Analysis of Covariance	65
9.1	Introduction	65
9.1.1	What is a covariate?	65
9.1.2	How do you choose a covariate?	65
9.1.3	Assumptions	66
9.1.4	If a covariate is not	66
9.2	Example of the Design	66
9.3	Setup in SPSS	67
9.4	Output in SPSS	67

Chapter 1

Data Entry

1.1 Introduction

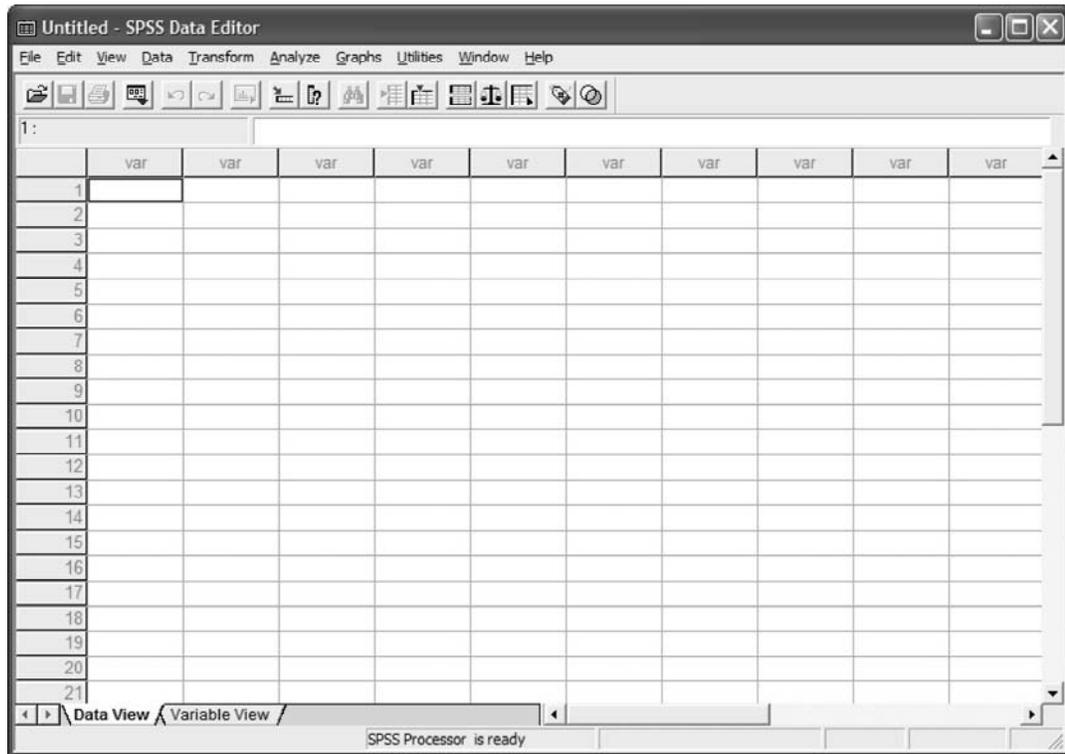
- SPSS for Windows is just like any other Windows program. Many familiar Windows features are available in SPSS (e.g., menu bar, tool bar, cut, copy, paste; Figure 1.1).
- Many features of the data editor are similar to those found in spreadsheet applications (e.g., Microsoft Excel). One row per subject and one column per variable (observation).
- Before entering data: define each variable in the dataset.

1.2 Defining Variables

SPSS introduced Data View and Variable View in the data editor since version 10. You can switch between the two views by clicking on the tap at the bottom of SPSS window,

Switch to Variable View, then click on the box under Name. Type in the

Figure 1.1: SPSS for Windows



variable name then type enter. SPSS will automatically assign its default setting for other columns (e.g., type, width, decimals).

There are few restrictions on variables names:

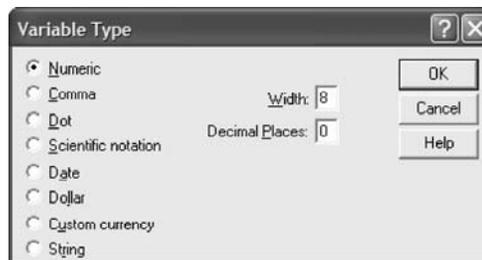
- the name must begin with a letter. The remaining character can be any letter, digit, a period, or symbols (@, #, -, or \$).
- variable names can not end with a period.
- variable names that end with an underscore should be avoided (to avoid conflict with variables automatically created by some procedure).
- the length of the name cannot exceed eight characters.

- blanks and special characters cannot be used.
- each variable name must be unique.

To change the default type and other settings, either click on the cell and type or click on [...] button to open a dialog box.

- Define Variable Type (Figure 1.2). A variable does not always have to be a number. Width and Decimals columns can be typed in from the Define Variable dialog box or in the Variable view from data editor (along with columns, align, and measure). If there are several variables with same type. Type definition along with other columns can be copied and pasted. Since variable name is limited to 8 character a longer name can be specified in the Label column (upto 256 character with spaces). Measures columns identifies variables as 3 different type (scale, ordinal, and nominal).

Figure 1.2: Variable Type

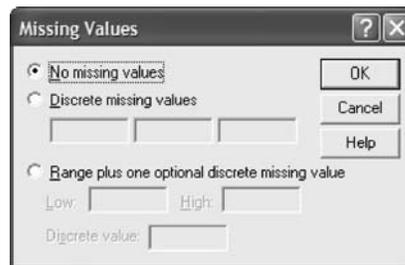


- Value Labels (Figure 1.3) is used to assign descriptive labels for each value of a variable. This is particularly useful for categorical variables (e.g., sex, 1 for male and 2 for female).
- Missing Values (Figure 1.4) is used to assign missing value codes. A discrete, range, or both can be used to define missing value codes.

Figure 1.3: Value Label



Figure 1.4: Missing Values



Chapter 2

Data Management

2.1 Introduction

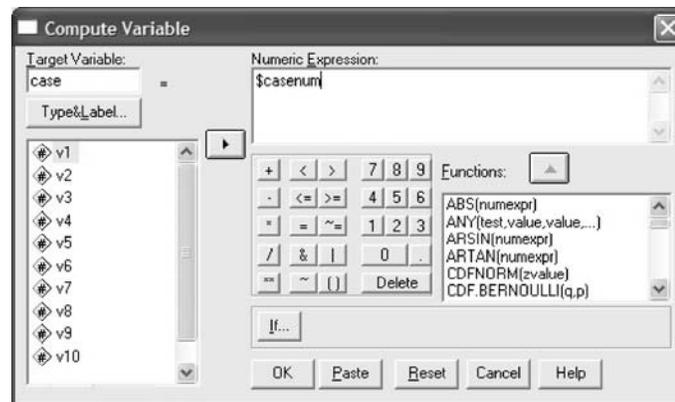
A data file should be created such that it can be accessed in the future with ease. Here are some general guidelines.

- Label each levels of variable (not always; when you should vs. not)
- Create a codebook (a document that explains the dataset, including definitions of all variables, how were they changed, how composite variable were created, any recoding of the variable, etc.). Keep a copy of the file in the same folder as the data file.
- Keep a copy of the original data file. Always work on a copy, NOT the original.
- Give descriptive file names. File names in Windows could be as long as 256 characters.
- Include any and all information you have. They may be useful in the future analyses.
- Each case or subject should have a unique identifier (e.g., id number). If the data file does not have a unique identifier, you could use case numbers.

To create case numbers in your data; click Transform → Compute. Type a name under Target Variable and type \$casenum in the Numeric Expression box (Figure 2.1). Its equivalent syntax would be:

```
compute case = $casenum.
execute.
```

Figure 2.1: Compute Variable



2.2 Transform Menu

Transform menu offers many different method to modify or create variables.

2.2.1 Recoding Variables

SPSS can recode a variable into same variable or different variable.

- Click on Transform → Recode → Into Same Variable (Figure 2.2). Move over the variable(s) from left into variables box.

- Click on Transform → Recode → Into Different Variable (Figure 2.3). Move over the variable from the left into the input variable → output variable box. Type in the new name in the output variable name box then click change.

Figure 2.2: Recode Same

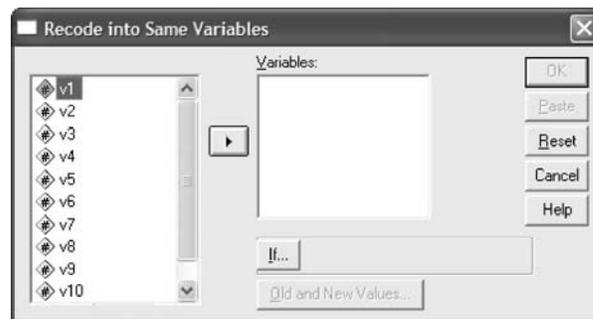
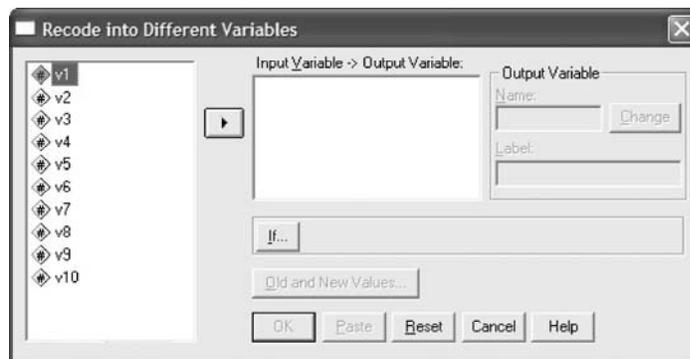


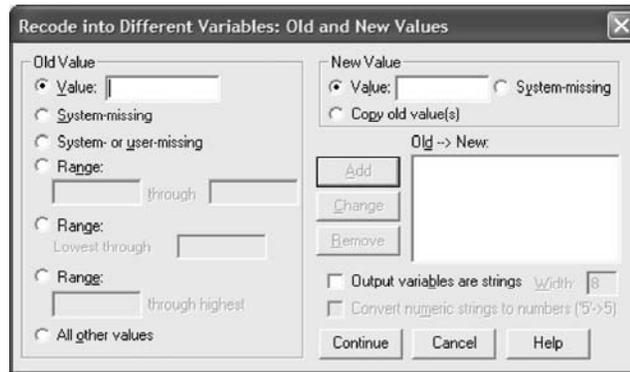
Figure 2.3: Recode Different



- From either Figure 2.2 or Figure 2.3, click on old and new values (Figure 2.4). Use the options under Old Values to specify which values to recode and how.
- To fast reverse code a variable, try using compute statement where

$$\text{recode variable} = (\text{minimum} + \text{maximum value}) - \text{variable}.$$

Figure 2.4: Recode Old and New Values



2.2.2 Creating a Composite Variable

A composite variable are usually created by either computing the mean or the sum of items (variables). Make sure to check how a composite score is compute for a scale. It does not make any difference as far as an analysis. However, some scales have norms that you might want to compare it against.

For these purpose, SPSS has two built in command, Mean and Sum. Not to mention, the old fashion method of computing a mean or sum; add all variables to compute sum and divide by number of variables to compute the mean. Both methods are accomplished in the Compute Variable windows (Figure 2.1).

For example,

```
selfest = mean(se1, se2, se3, se4, se5).
```

if se1 through se5 are consecutive, the above equation can be replace by:

```
selfest = mean(se1 to se5).
```

Note, if mean or sum function is used, it will compute a mean or sum using all available case if there are missing data. You can place a restriction on SPSS

such that it only computes a mean or sum if there is a minimum of observations by using the functions; MEAN.x and SUM.x where x is replace by minimum number.

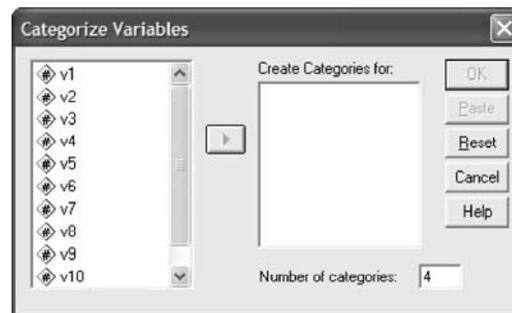
For example,

```
selfest = mean.3(se1 to se5).
```

2.2.3 Categorizing a Variable

SPSS can automatically categorize a variable to a specific number of levels. Click on Transform → Categorize Variables (Figure 2.5). Move over variable(s) from the left to Create Categories For box, then type in the number of categories desired at the bottom.

Figure 2.5: Categorize Variables



2.3 Small Set

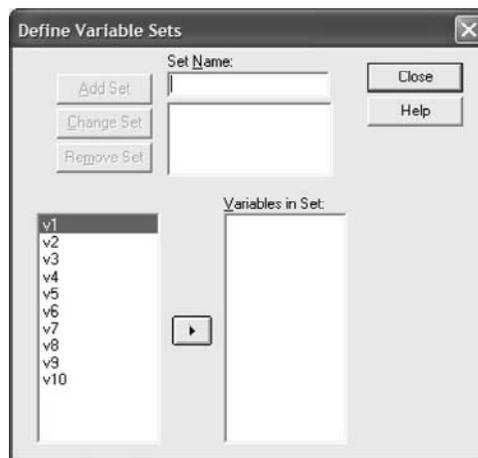
If you have a data set with many variables, however, you don't want to see all the variables each time you go to the variables list, then:

- use sets or
- create a new smaller data set

2.3.1 Use Sets

To use sets click on Utilities → Define Sets (Figure 2.6). You must define a set before you could use them. You only need to define a set one time.

Figure 2.6: Define Sets



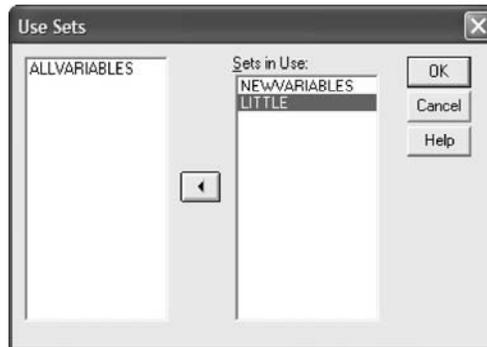
Type a name for your set you are defining (e.g., little). Move over variable(s) in your set, then click Add Set button and Close.

Once you have define your set, goto Utilities → Use Sets (Figure 2.7). Move over your set name to Set in Use box and remove ALLVARIABLES. Keep NEWVARIABLES in the Set in Use box.

2.3.2 Create Smaller Dataset

To create a smaller data set, click on File → Save As (Figure 2.8), type in a file name. Then click on Variables (Figure 2.9). Remove or Add check mark next to variables you want to keep or delete.

Figure 2.7: Use Sets



2.4 Select Cases

Select Cases (Figure 2.10) under Data Menu allows SPSS to perform analysis on only select cases. Cases can be selected on certain criteria that can be specified. A criteria can be specified according some condition, random sample, and based on time or case range. Unselected case can be temporary filtered or deleted. Don't forget to go back and turn off the filter once you are finished.

One useful criteria is If condition is satisfied (Figure 2.11). Conditional statement can be specified using any built-in or custom functions.

2.5 Split File

Often times, it is necessary to analyze the data within each group separately. Instead of running an analysis multiple times. Split file (Figure 2.12) under Data Menu can be used to analyze within each group simultaneously. Output will be generated for each group. Don't forget to turn off the split file once you are finished.

Figure 2.8: Save As

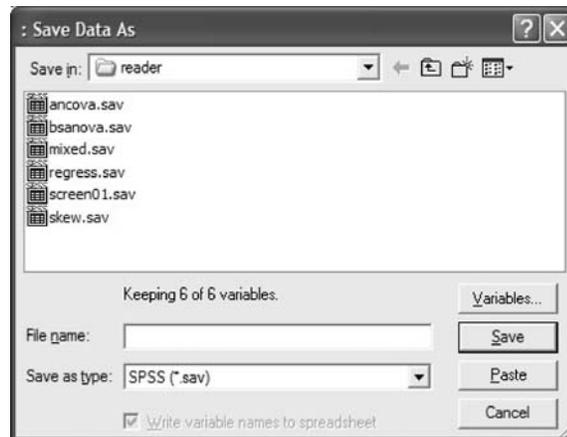


Figure 2.9: Save As Variables

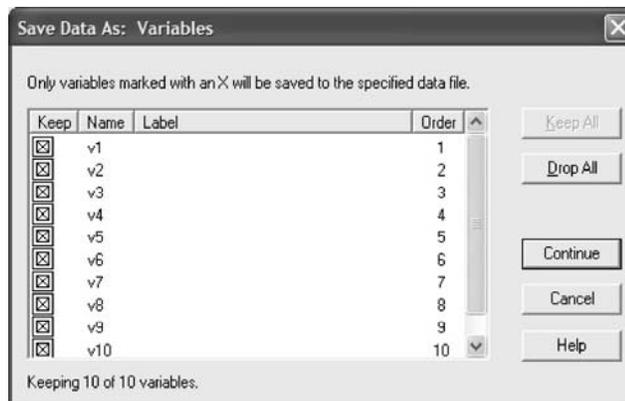


Figure 2.10: Select Cases

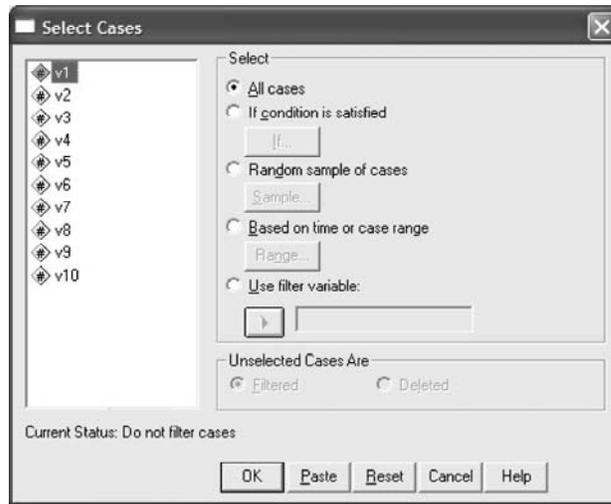


Figure 2.11: Select Cases - Condition

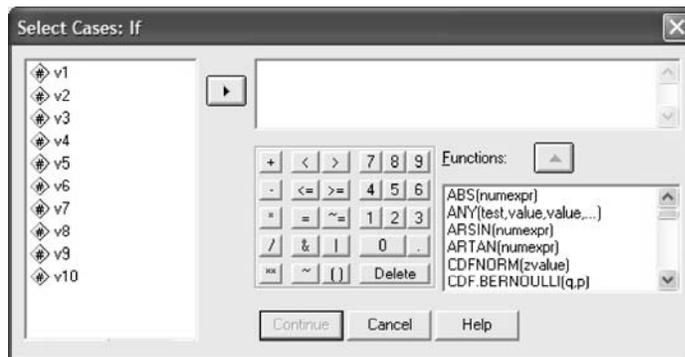
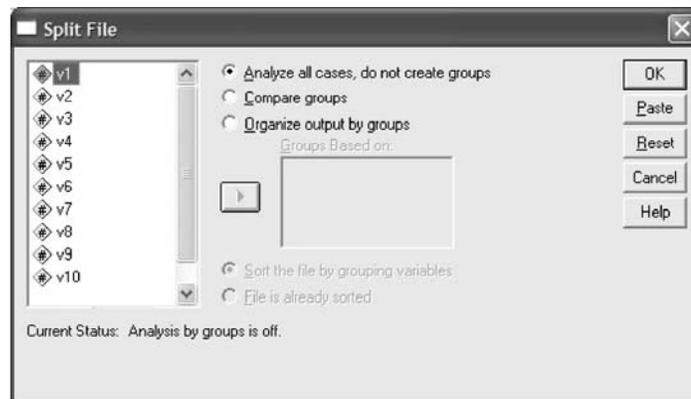


Figure 2.12: Split Files



Chapter 3

Data Screening

Before any analysis is performed, you must do everything possible to ensure a data file is clean (e.g., meet the necessary assumptions of a statistical procedure) or at least know the weakness of your data.

3.1 Descriptive Information

3.1.1 SPSS Explore

- The best method for data screening via descriptive information is the Explore procedure. To start, click on Analyze → Descriptive Statistics → Explore (Figure 3.1). Move over the variables to the Dependent List. If you have group data, you should screen the data within each group. An independent variable can be inserted under Factor List. However, if it is a factorial design, a syntax command must be used. A syntax command can be created easily by clicking on Paste instead of OK.
- Click on Statistics button from Figure 3.1. Check Descriptive and Outliers (Figure 3.2). By checking on Outliers, it will produce 5 highest and lowest scores with their case numbers.
- Click on Plots button from Figure 3.1. Uncheck Stem-and-leaf. It is not

Figure 3.1: Explore

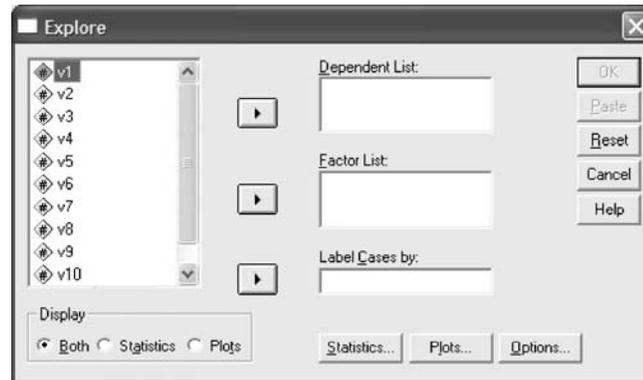
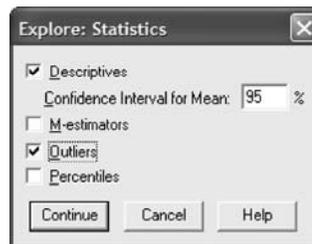


Figure 3.2: Explore - Statistics



useful anymore with computers' ability to generate high resolution charts. Check Histogram (Figure 3.3).

- Click on Options button from Figure 3.1. I would recommend choosing Exclude cases pairwise to start (Figure 3.4). Once you know, there is no silly mistake then you might want to come back and perform listwise deletion.

3.1.2 Example Output of the Explore

- Case Processing Summary is produced first (Figure 3.5). This table shows how many total cases with the number of missing per variable.

Figure 3.3: Explore - Plots

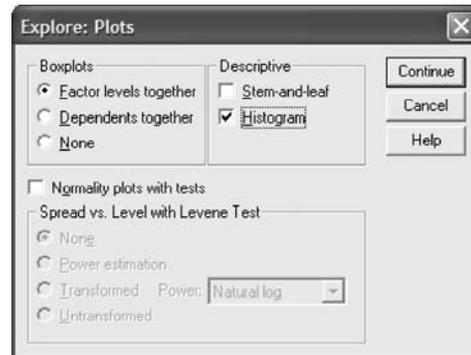
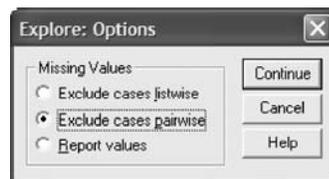


Figure 3.4: Explore - Options



- Next, Descriptive table is produced with the usual mean and standard deviation as well as some distributional information (e.g., skewness and kurtosis) (Figure 3.6).
- Extreme Values (Figure 3.7) table list five highest and lowest values with their case numbers. This is very useful in detecting outliers.
- Histogram and Boxplot are produced next (Figure 3.8 and 3.9). It is always good idea to have a picture.

Figure 3.5: Explore - Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
V1	100	100.0%	0	.0%	100	100.0%
V2	100	100.0%	0	.0%	100	100.0%
V3	100	100.0%	0	.0%	100	100.0%
V4	100	100.0%	0	.0%	100	100.0%
V5	100	100.0%	0	.0%	100	100.0%

Figure 3.6: Explore - Descriptives

			Statistic	Std. Error	
V1	Mean		5.47	.134	
	95% Confidence Interval for Mean	Lower Bound	5.20		
		Upper Bound	5.74		
	5% Trimmed Mean		5.57		
	Median		6.00		
	Variance		1.807		
	Std. Deviation		1.344		
	Minimum		0		
	Maximum		8		
	Range		8		
	Interquartile Range		1.00		
	Skewness		-1.042		.241
	Kurtosis		2.036		.478
V2	Mean		4.23	.426	
	95% Confidence Interval for Mean	Lower Bound	3.38		
		Upper Bound	5.08		
5% Trimmed Mean		3.67			

3.2 Linearity and Homoscedasticity

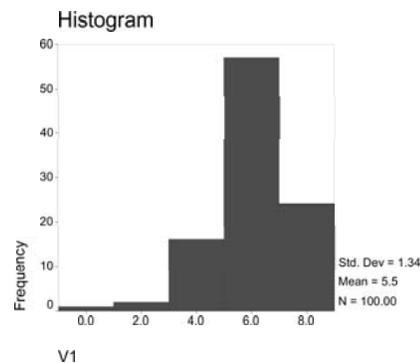
To check for linearity and homoscedasticity, use scatterplots. Click on Graphs → Scatter (Figure 3.10). Four choices are presented. Usually for data screening purposes, start with Matrix (Figure 3.11). if you notice something strange use Simple to get a larger view of the possible problem.

For testing these assumptions, basically as long as there is not non-linearity nor heteroscedasticity, then everything is okay.

Figure 3.7: Explore - Extreme Values

Extreme Values				
		Case Number	Value	
V1	Highest	1	49	8
		2	88	7
		3	96	7
		4	75	7
		5	64	8
	Lowest	1	14	0
		2	24	2
		3	54	2
		4	35	3
		5	44	3
V2	Highest	1	84	23
		2	93	20

Figure 3.8: Explore - Histogram



3.3 Detecting Multivariate Outliers

In order to compute Mahalanobis Distances, you must perform a linear regression (see Chapter 5). Move over the variables you want to screen into independent variables and use any variable as dependent (good option would be id variable). Go to Save option (Figure 5.4) and choose Mahalanobis under Distances section. This will create a new variables in the data set Mah_x where x is a counter (first time you compute it. x will be 1). Ignore any output that is produced. Perform an Explore procedure on the newly created variable. Look for extreme values (Figure 3.13). Mah_x will be χ^2 (degrees of freedom = number of variables) distributed.

Figure 3.9: Explore - Boxplot

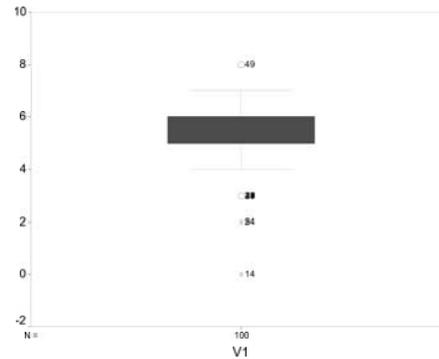
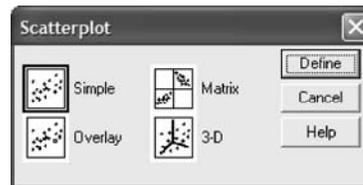


Figure 3.10: Scatterplot



3.4 What if the data is not Normal?

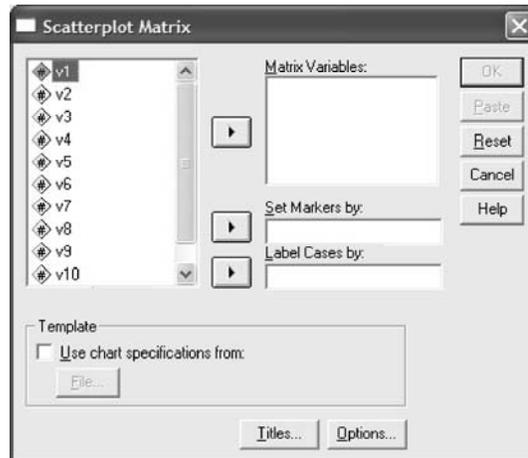
The two most common approaches to non-normal data:

- relying on the robustness of a method (do nothing approach)
- transformation of variables

3.4.1 Robustness of a method

Most basic statistical method (e.g., t-test, ANOVA) are fairly robust against violation of normality. The results will still be meaningful even with some violations.

Figure 3.11: Scatterplot - Matrix



3.4.2 Transformation

There are no firm guidelines when a variable should be transformed. It depends on the statistical methods and interpretation of the transformed variable. If you are not sure whether to transform or not, you could analyze your data both ways; transformed and not transformed. If the results are same, then use the original variables. However, if the results are vastly different transformed variables result will be more reliable and accurate.

Figure 3.12: Scatterplot - Matrix Output



Figure 3.13: Mahalanobis Distances: Histogram

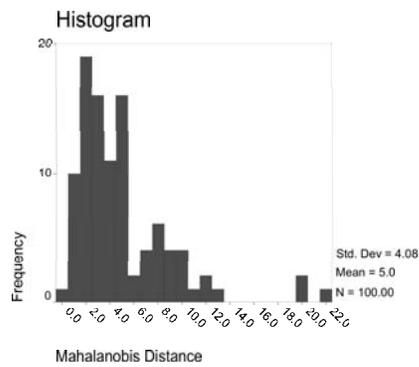
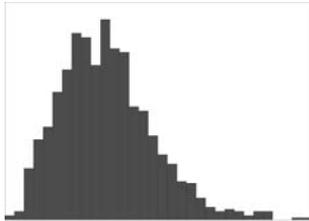


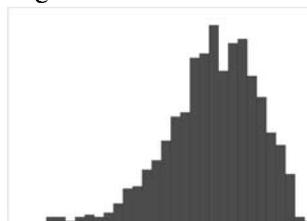
Figure 3.14: Several Common Transformations

Positive Moderate



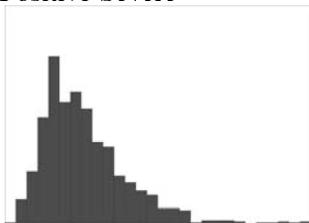
new variable = $\sqrt{\text{old variable}}$

Negative Moderate



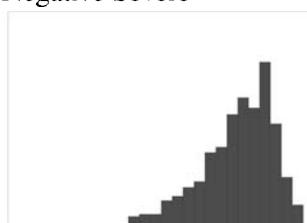
new variable = $\sqrt{K - \text{old variable}}$ ^a

Positive Severe



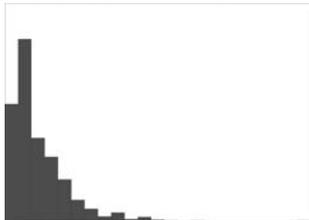
new variable = $\lg_{10}(\text{old variable})$
 new variable = $\lg_{10}(C + \text{old variable})$ ^b

Negative Severe



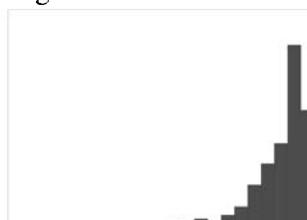
new variable = $\lg_{10}(K - \text{old variable})$

Positive Substantial



new variable = $1/\text{old variable}$
 new variable = $1/(C + \text{old variable})$

Negative Substantial



new variable = $1/(K - \text{old variable})$

^a $K = \text{largest score} + 1$

^bIf zero is in the range use this equation, where $C = \text{smallest score} + 1$.

Chapter 4

Matrix Command in SPSS

4.1 Introduction

SPSS offers Matrix commands allowing a researcher to compute any custom equation using matrix algebra. Full list of commands are listed in `spssbase.pdf` file in the `spss` folder. This chapter will only cover select few commands that will be useful for GLM (Table 4.1).

To start programming your own matrix commands, open a syntax window by clicking on File → New → Syntax. After commands have been typed in, to execute, click on Run → All or Selection.

Matrix commands in SPSS start and end with special commands letting SPSS know that you are starting your own customize computations.

```
matrix.  
***insert commands here***  
end matrix.
```

4.2 Defining a Matrix

A matrix (or vector) can be defined by typing in the elements of the matrix in the syntax or have SPSS read in a matrix from a file (e.g., SPSS data window). Although, there are more than one method of reading in a matrix from external source, only one method will be discussed here (i.e., reading from SPSS data window).

4.2.1 from syntax

A matrix is defined using { }. Columns in a matrix is separated by commas and rows by semi-colon.

to define the following matrix:

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 3 \end{pmatrix} \quad (4.1)$$

```
matrix.  
compute A = {1,2,3;2,3,3}.  
end matrix.
```

compute keyword must be specified before defining a matrix (of course, enclosed by matrix-end matrix commands).

4.2.2 from SPSS data window

Often, it is not reasonable to input a data from syntax windows. It is easier to use the SPSS data window. Data from SPSS data windows can then be read in by columns. for instance, if there are y, x1, and x2 in your data windows and you wish to read them into matrix command, use a get statement:

```
matrix.
```

```
get maty /variable = y.  
end matrix.
```

get statement is followed by a name of a matrix, maty. Any name can be used for matrix name as long as it is not a built-in keywords. A matrix name, maty, is followed by a keyword /variable = and name of the variable in the data window. Get statement can be used to read in multiple columns at the same time.

```
matrix.  
get matx /variable = x1 x2.  
end matrix.
```

maty has one column, while matx has two columns with number of rows equal to number of cases in the data window.

4.3 Simple Computation

Any computation can be performed as long as it is defined.

```
matrix.  
compute A = {1,2,3;2,3,3}.  
compute B = {2,5,4;2,4,3}.  
compute C = A + B.  
end matrix.
```

Regression coefficients (regression coefficient and intercept) for a simple linear regression can be computed as:

$$\hat{\beta} = (X'X)^{-1}X'Y \quad (4.2)$$

```
matrix.  
get y /variable = y.  
get x1 /variable = x1.  
compute n = nrow(y).  
compute x0 = make(n,1,1).  
compute x = {x0,x1}.  
compute b = inv(t(x)*x)*t(x)*y.  
end matrix.
```

In the commands above, there are slightly more complicated than simple computation.

- `nrow(y)` command was defined to computed number of rows.
- `make(n,1,1)` command was defined to create the dummy variable of ones for the design matrix.
- `x` was created by concatenating `x0` and `x1`.
- `t(x)` command computes transpose of `x`.

4.4 Generating output

Now that computation has been complete. There must be a method to output the results. This is done by using the `print` command. For example, to output the regression coefficient and intercept add

```
print b /title 'regression coefficients'.
```

before `end matrix` command. It will print `b` with the title regression coefficients.

another method is to save a matrix as a SPSS data file using `save` command.

```
save matvar /outfile = 'filename' /variables = varnames.
```

where `matvar` is matrix variable, `filename` is name of the file (use `.sav` extension), and `varnames` is name of the variables.

Table 4.1: Some useful Matrix Commands

CDFNORM	cumulative normal distribution function
CHICDF	cumulative chi-squared distribution function
CMAX	column maxima
CMIN	column minima
CSSQ	column sum of square
DESIGN	create design matrix
DET	determinant
DIAG	diagonal of matrix
FCDF	cumulative F distribution function
IDENT	create identity matrix
INV	inverse
KRONECKER	Kronecker product of two matrices
MAKE	create a matrix with all elements equal
MMAX	maximum element in matrix
MMIN	minimum element in matrix
MSSQ	matrix sum of squares
MSUM	matrix sum
NCOL	number of columns
NROW	number of rows
RESHAPE	change shape of matrix
RMAX	row maxima
RMIN	row minima
RSSQ	row sum of squares
RSUM	row sums
SQRT	square roots of matrix elements
SSCP	sums of squares and cross-products
T	(synonym for TRANSPOS)
TCDF	cumulative normal t distribution function
TRACE	calculate trace
TRANSPOS	transposition of matrix

Chapter 5

Multiple Regression

5.1 Assumptions

5.1.1 Dependent Variable

- interval/ratio scale
- normally distributed (technically, errors/residuals should be normally distributed)
- no univariate outliers

5.1.2 Independent Variables

- any level of measurement. However, nominal and ordinal variables must be dummy coded.
- multicollinearity/singularity
 - $tolerance = 1 - SMC$
 - $VIF = 1/tolerance$

5.1.3 Relationship between DV and IV

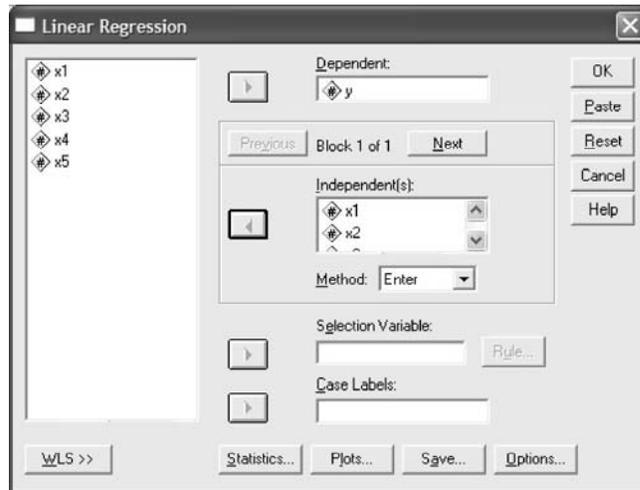
- linearity
- homoscedasticity (in group data, e.g., ANOVA, this is known as homogeneity of variance). Heteroscedasticity standard errors can be used if a data violate this assumption (not available in SPSS).
- independence of residuals
- no multivariate outliers
 - “what if a population I am studying has multivariate outliers and I don’t want to delete cases?”
 - * with outliers, it has undue influence on parameters estimated (i.e., regression coefficients and standard errors). We want cases to be equally weighted. But, what we really want is cases to be equally important/influential on estimating parameters. weighting variables can be used.

5.2 Setup in SPSS

Multiple Regression is used to predict an interval or ratio dependent variable from a set of independent variables (most likely interval or ratio, but it is capable of using nominal and ordinal variables as long as the variables are dummy coded).

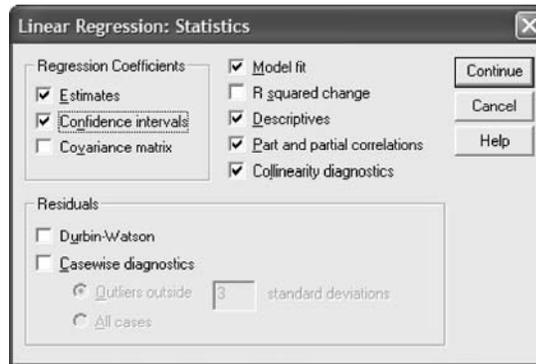
- To start: click on Analyze → Regression → Linear (Figure 5.1).
- SPSS is capable of performing 5 different method of multiple regression which could be classified into 3 categories:
 - standard multiple regression
 - * enter
 - sequential multiple regression
 - * stepwise
 - * forward

Figure 5.1: Linear Regression



- * remove
- * backward
- hierarchical multiple regression
 - * use Block feature in SPSS. Within each Block, there can be different methods (standard or sequential).
- click on Statistics in the Linear Regression (Figure 5.1) dialog box (Figure 5.2):
 - always check:
 - * estimates
 - * model fit
 - * confidence intervals
 - * part and partial correlations
 - first run
 - * descriptives
 - * collinearity diagnostics
 - sequential

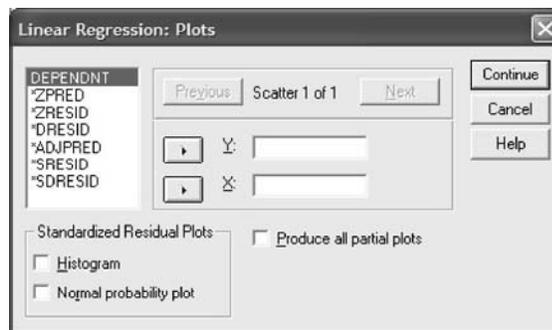
Figure 5.2: Linear Regression - Statistics



* R squared change

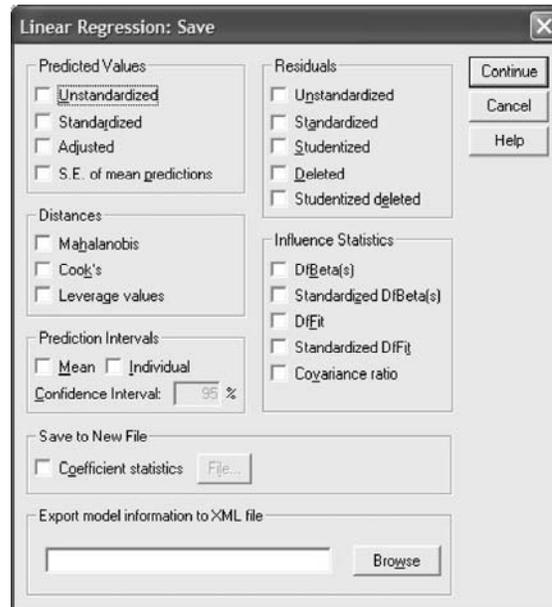
- Plots and Save in regression can be used for regression diagnostics.
- No plots. If you want to: click on Plots in the Linear Regression (Figure 5.1) dialog box (Figure 5.3). Move over the variables from the left to appropriate boxes in the middle. To draw more than one graph, click on Next at the top.

Figure 5.3: Linear Regression - Plots



- Don't save anything. if you want to: click on Save in the Linear Regression (Figure 5.1) dialog box (Figure 5.4). Many of these variables are used for regression diagnostic purposes.

Figure 5.4: Linear Regression - Save



- Don't change the Options (Figure 5.5).

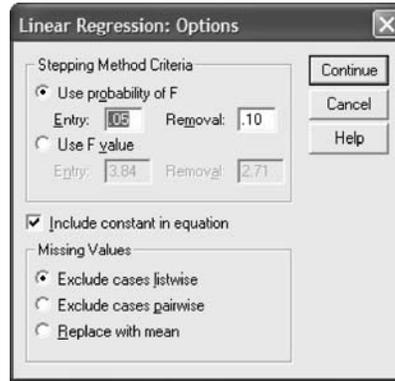
5.3 Output in SPSS

5.3.1 Model summary

Model Summary (Figure 5.6):

- R = correlation between the DV and set of IVs
- R square = proportion (amount) of variability in the dependent variable explained by the set of independent variables in the sample
- Adjusted R square = estimated proportion of variability in the dependent

Figure 5.5: Linear Regression - Options



variable explained by the set of independent variables in the population (adjusting for sample size and number of independent variables)

- Std. Error of the Estimate = standard deviation of the sampling distribution of the residuals (dependent variable - predicted dependent variable)

Figure 5.6: Linear Regression - Model Summary

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.867 ^a	.753	.739	2.271

a. Predictors: (Constant), X5, X4, X1, X2, X3

5.3.2 ANOVA table

ANOVA table (Figure 5.7):

- there is a significant prediction of the dependent variable by the independent variables.

Figure 5.7: Linear Regression - ANOVA

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1474.769	5	294.954	57.167	.000 ^a
	Residual	484.991	94	5.159		
	Total	1959.760	99			

a. Predictors: (Constant), X5, X4, X1, X2, X3
b. Dependent Variable: Y

5.3.3 Regression Coefficients

Regression Coefficients (Figure 5.8):

- correlations
 - zero-order = Pearson product moment correlation
 - partial
 - part (semi-partial)
- tolerance (do not want it to be close to zero)
- VIF (do not want it to be too large)

5.3.4 Collinearity Diagnostics

Collinearity Diagnostics (Figure 5.9)

- examine the last row: do not want Condition Index to be larger than 30 and two of the variance proportions to be larger than .50 excluding the constant.

Figure 5.8: Linear Regression - Coefficients

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	13.229	4.257		3.108	.002	4.777	21.680
	X1	.487	.055	.537	8.854	.000	.378	.596
	X2	.231	.029	.500	8.074	.000	.174	.287
	X3	-.134	.054	-.154	-2.472	.015	-.241	-.026
	X4	.259	.173	.094	1.492	.139	-.086	.603
	X5	.534	.084	.381	6.326	.000	.366	.701

Coefficients^a

Model		Correlations			Collinearity Statistics	
		Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)					
	X1	.629	.674	.454	.717	1.395
	X2	.240	.640	.414	.687	1.455
	X3	-.399	-.247	-.127	.681	1.469
	X4	.546	.152	.077	.668	1.497
	X5	.480	.546	.325	.728	1.374

a. Dependent Variable: Y

Figure 5.9: Linear Regression - Collinearity Diagnostics

Collinearity Diagnostics^a

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions					
				(Constant)	X1	X2	X3	X4	X5
1	1	5.903	1.000	.00	.00	.00	.00	.00	.00
	2	5.625E-02	10.244	.00	.00	.06	.06	.00	.29
	3	1.926E-02	17.506	.00	.00	.34	.37	.06	.02
	4	1.418E-02	20.400	.00	.16	.27	.02	.08	.57
	5	5.528E-03	32.678	.02	.33	.26	.15	.85	.00
	6	2.197E-03	51.830	.97	.51	.07	.39	.00	.11

a. Dependent Variable: Y

Chapter 6

Regression Diagnostics

All the computations described in this chapter can be computed and saved into the data file using Save option (Figure 5.4) under Multiple Regression (Chapter 5). After the variables have been computed, Explore procedure (Section 3.1.1) can be used to detect any influential cases.

The issue of diagnostics involves the detection of outliers and extremely influential data points that may distort the interpretation of regression output. The issue of diagnostics can be broken down into two separate subtopics:

- outliers
- Influence analysis

6.1 Outliers

An outlier is an extreme datum that may bias the interpretation of the parameter estimates (i.e., regression coefficients and standard errors) in a regression analysis. Outliers may arise because of:

- Recording or input error.

- Measurement error.
- Inappropriate or misunderstood instructions.
- A variety of other reasons.

There are three major approaches to the detection of outliers:

- Use of standardized residuals.
- Use of studentized residuals.
- Use of studentized deleted residuals.

6.1.1 Standardized Residuals (ZRESID)

A standardized residual is simply computed as the residual score (i.e., the difference between the \hat{Y}_i , predicted Y -score, and the actual Y -score from a regression equation), divided by the standard error for the regression (i.e., $s_{Y \cdot X}$). In other words,

$$\text{ZRESID}_i = \frac{Y_i - \hat{Y}_i}{s_{Y \cdot X}} \quad (6.1)$$

6.1.2 Studentized Residual (SRESID)

Therefore, the studentized residual approach divides each residual score by an estimate of its own standard error. In the case of the simple linear regression, this standard error is given by:

$$s_{e_i} = s_{Y \cdot X} \sqrt{1 - \left[\frac{1}{N} + \frac{(X_i - \bar{X})^2}{SS_X} \right]} \quad (6.2)$$

Where SS_X is the sum of squares for X . Note that the more X_i deviates from the \bar{X} , the smaller the standard error of the residual.

Also, note that when the assumptions of the model are reasonably met, the SRESID follows a t -distribution with $df = N - K - 1$, where K is the number of independent variables.

6.1.3 Studentized Deleted Residuals (SDRESID)

The greater the extent to which a given data point is an outlier, the more its retention in the analysis will lead to upward bias in the standard error of estimate $s_{Y \cdot X}$ and thereby running the risk of failing to identify it as an outlier. To correct for this potential problem, we can compute the standard error of a deleted residual as:

$$s_{e_i} = s_{Y \cdot X_i} \sqrt{1 - \left[\frac{1}{N} + \frac{(X_i - \bar{X})^2}{SS_X} \right]} \quad (6.3)$$

Where $s_{Y \cdot X_i}$ is the standard error of estimate when person i has been deleted from the analysis.

However, this approach would involve computing N separate regression analyses, one for every deleted individual. An alternative approach would be to compute:

$$\text{SDRESID}_i = \text{SRESID}_i \sqrt{\frac{N - K - 2}{N - K - 1 - \text{SRESID}_i^2}} \quad (6.4)$$

Note also that SDRESID is distributed as t with $df = N - K - 2$.

6.2 Influence Analysis

An influential observation is one which has, either alone or together with other observations, a much larger impact on regression outcomes (i.e., regression coefficients, standard errors, t -values, etc.) than most other observations.

Note, an outlier can be, but is not necessarily, an influential observation.

6.2.1 Leverage

The way to assess an observations' influence is to compute its leverage¹ (h_i).

$$h_i = \frac{1}{N} + \frac{(X_i - \bar{X})^2}{SS_X} \quad (6.5)$$

Note the following about Leverage:

1. Leverage is solely a function of the independent variables. Thus a case that may be influential by virtue of its status as a dependent variables may not be detected on the basis of its leverage.
2. The greater the deviation of X_i from the \bar{X} , the greater the leverage.
3. The maximum value of leverage is 1 and the minimum value is $1/N$.
4. The average leverage for a set of scores is $K + 1/N$, where K is the number of independent variables.
5. It has been suggested that leverage can be considered high if it is $h_i > \frac{2(K + 1)}{N}$

For those familiar with matrix algebra, the coefficient h_i can also be found as the i^{th} diagonal element in the matrix $X(X'X)^{-1}X'$, where X is the $N \times (K + 1)$ matrix of independent variables augmented by a column of 1's.

¹SPSS computes leverage using a slightly different equation. It does not add $1/N$ to the computation of leverage

6.2.2 Cook's D

As we recall, leverage only identifies a case as independent variables rather than dependent variables as influential. To detect influential scores, regardless of whether or not they are independent or dependent variables, we can use Cook's D (distance) measure:

$$D_i = \left(\frac{\text{SRESID}_i^2}{K + 1} \right) \left(\frac{h_i}{1 - h_i} \right) \quad (6.6)$$

Although there are significance tests for D , for diagnostic purposes it is sufficient to look for large D values in relation to the rest of the data.

6.2.3 DFBETA

One can also look at the effects on specific parameter estimates of a given observation. In DFBETA_{k_i} indicates the change in k (intercept or regression coefficient) when i is deleted.

In the case of a simple regression, we will refer to DFBETA_{a_i} as a change in the intercept a , and DFBETA_{b_i} as a change in the regression coefficient b when i is deleted.

While there are more laborious formulas one might use, the easiest and the one requiring only a single run of the data is:

$$\begin{aligned} \text{DFBETA}_{a_i} &= a - a_i \\ &= \left[\frac{\sum X_i^2}{N \sum X_i^2 - (\sum X_i)^2} + \left(\frac{-\sum X_i}{N \sum X_i^2 - (\sum X_i)^2} \right) X_i \right] \left(\frac{e_i}{1 - h_i} \right) \end{aligned} \quad (6.7)$$

where e_i is the residual.

Turning our attention to the regression coefficient, we have:

$$\begin{aligned} \text{DFBETA}_{b_i} &= b - b_i \\ &= \left[\frac{-\sum X_i}{N \sum X_i^2 - (\sum X_i)^2} + \left(\frac{N}{N \sum X_i^2 - (\sum X_i)^2} \right) X_i \right] \left(\frac{e_i}{1 - h_i} \right) \end{aligned} \quad (6.8)$$

6.2.4 Standardized STDFBETA

We must now face the problem of deciding what constitutes a large DFBETA. One way to answer this question is by means of standardization. We will call this STDFBETA.

$$\text{STDFBETA}_i = \frac{\text{DFBETA}_{a_i}}{\text{MSR}_i \left(\frac{\sum X_i^2}{N \sum X_i^2 - (\sum X_i)^2} \right)} \quad (6.9)$$

Where MSR_i is the mean square residual when i is deleted. MSR_i can be found by:

$$\text{MSR}_i = \frac{SS_{res} - \frac{e_i^2}{1-h_i}}{N - K - 2} \quad (6.10)$$

Chapter 7

Between-Subjects Analysis of Variance

7.1 Assumptions

- Normality: the dependent variable (technically speaking, the errors) should be normally distributed within each cell. Usually ANOVA is robust against violations of normality as long as a variable is skewed in the same direction across all cells of the design.
- Homogeneity of variance: variances of the dependent variable are same in each cell in the design. ANOVA must meet this assumption. There is no way around this violation of assumption.
- Outliers: no univariate outliers. Outliers will bias results.
- Sample size: good rule of thumb is minimum of 10 subjects per cell. also, the ratio between largest sample size to smallest size should be less than 3:1.

7.2 One-way Between-Subjects ANOVA

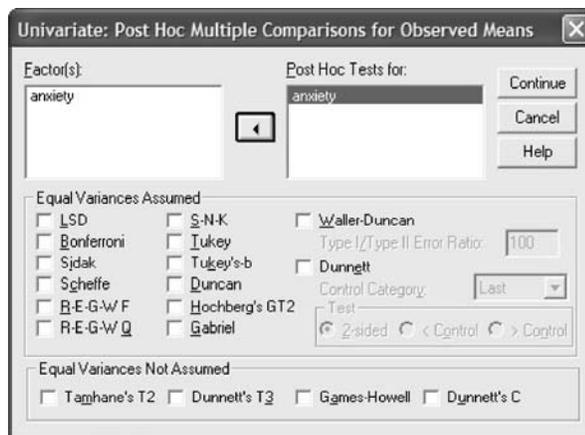
Both one-way and factorial ANOVAs can be performed through the GLM procedure in SPSS. It does not make a difference for SPSS whether there are one IV or multiple IVs.

A one-way between-subjects ANOVA can be performed using two different procedure in SPSS (GLM and One-way ANOVA under Compare Means). Both procedure generate the same results. One-way ANOVA does offer easy method of performing contrasts while GLM produces richer information (e.g., effect size).

The outputs for both one-way and factorial ANOVAs are similar. Examine the two-way between-subject ANOVA outputs (Section 7.3.3) for help.

If there is a significant difference (i.e., main effect), then a researcher should consider performing a post hoc comparisons. A post hoc comparisons can be requested by clicking on Post Hoc button from the GLM dialog box. SPSS provides vast number of post hoc comparisons (Figure 7.1).

Figure 7.1: General Linear Model - Post Hoc



7.3 Two-way Between-Subjects ANOVA

7.3.1 Example of the Design

A between-subjects analysis of variance is demonstrated using a simple dataset.

		Anxiety		
		Low	Medium	High
Sex	Male			
	Female			

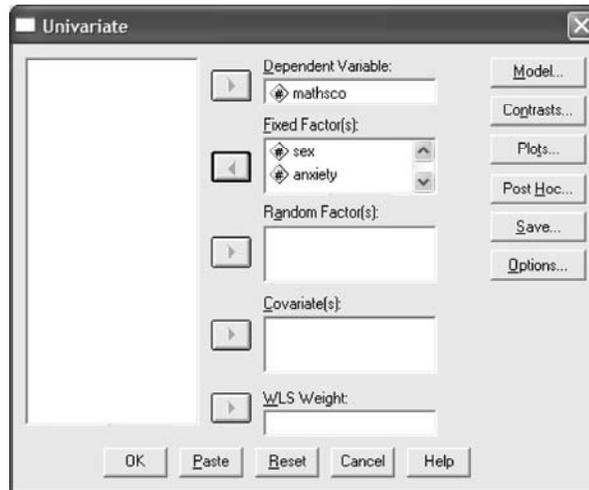
7.3.2 Setup in SPSS

- Click on Analyze → General Linear Model → Univariate (Figure 7.2). Fill in the variables as shown above.
- Click Options (Figure 7.3). Check Descriptive Statistics, Estimate of effect size (will report partial η^2), and homogeneity tests (will report Levene's test of homogeneity of variance).
- Click Plots from the main window (Figure 7.4). Move over the variables in to the middle. Usually, move over the variable with more levels as Horizontal Axis. For this example, move over Anxiety into Horizontal Axis and Sex as Separate Lines. Don't forget to click on Add button at the bottom. SPSS is capable of generating multiple plots.

7.3.3 Output in SPSS

Figure 7.5 displays SPSS test of homogeneity of variance (Levene's Test of Equality of Error Variance). The test should not be significant. However, this test is liberal. Worry if it is significant at $\alpha = .001$. If it is significant at $\alpha = .001$, then examine the variance of each cell to determine what is causing the problem. A better test, which is not available in SPSS, would be Brown-Forsythe.

Figure 7.2: General Linear Model - Univariate



7.3.4 Research hypotheses and How to test them

Starting Questions

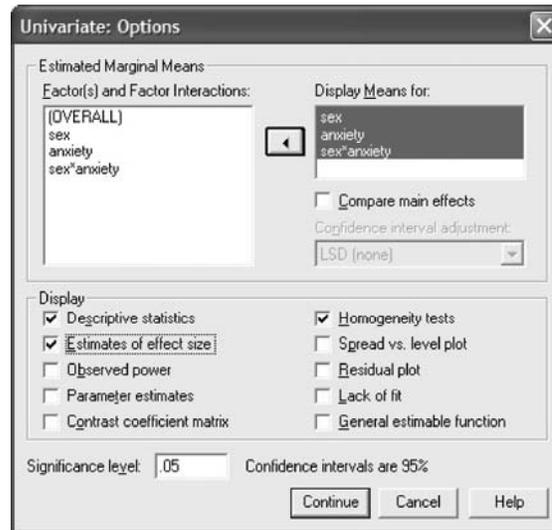
- Is there a significant difference on math scores between male and female college students averaged across anxiety levels? (Main effect of Sex)
- Is there a significant difference on math scores among anxiety levels averaged across sex? (Main effect of anxiety)
- Is the pattern of difference on the math scores among anxiety levels different between male and female college students? (Interaction of sex and anxiety)

What's Next?

If you have a specific hypothesis, test it. If you do not have a specific hypothesis, there are general guideline (Figure 7.7).

- for overall analysis without any option, the syntax would be:

Figure 7.3: General Linear Model - Options



GLM

```
mathsco by sex anxiety
/design = sex anxiety sex*anxiety.
```

- is the pattern of difference on math scores at low and medium anxiety levels different between male and female college students? (interaction contrast)

GLM

```
mathsco by sex anxiety
/lmatrix = "interaction contrast"
sex*anxiety 1 -1 0 -1 1 0
/design = sex anxiety sex*anxiety.
```

		Anxiety		
		Low	Medium	High
Sex	Male	1	-1	0
	Female	-1	1	0

Figure 7.4: General Linear Model - Plots

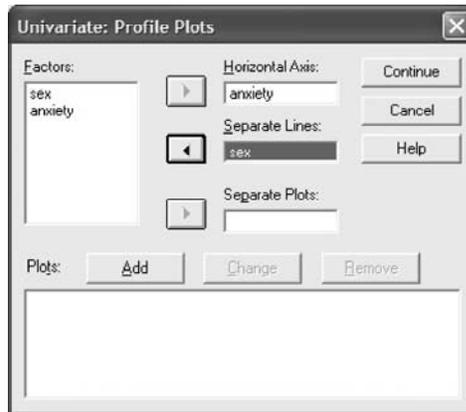


Figure 7.5: General Linear Model - Levene's Test

Levene's Test of Equality of Error Variances^a

Dependent Variable: MATHSCO

F	df1	df2	Sig.
3.827	5	84	.004

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept+SEX+ANXIETY+SEX * ANXIETY

To perform any contrast, you would need to understand /matrix. The general format of the command is:

```
/matrix = "label"
effect ***contrast coefficients***
```

effect in the example above was sex*anxiety. The effect is followed by contrast coefficient for each cell. In the example above the order was:

low male, medium male, high male, low female, medium female, high female (without commas)

the output for the above contrast is displayed in the Custom Hypothesis Tests section (Figure 7.8).

Figure 7.6: General Linear Model - ANOVA

Tests of Between-Subjects Effects

Dependent Variable: MATHSCO

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	2897.580 ^a	5	579.516	19.721	.000	.540
Intercept	477575.697	1	477575.697	16251.837	.000	.995
SEX	39.931	1	39.931	1.359	.247	.016
ANXIETY	368.553	2	184.276	6.271	.003	.130
SEX * ANXIETY	2489.096	2	1244.548	42.352	.000	.502
Error	2468.420	84	29.386			
Total	482941.697	90				
Corrected Total	5366.000	89				

a. R Squared = .540 (Adjusted R Squared = .513)

- is the pattern of difference on math scores at low compared to average of medium and high anxiety levels different between male and female college students? (interaction contrast)

GLM

```

mathsco by sex anxiety
/lmatrix = "interaction contrast"
sex*anxiety 1 -1/2 -1/2 -1 1/2 1/2
/design = sex anxiety sex*anxiety.

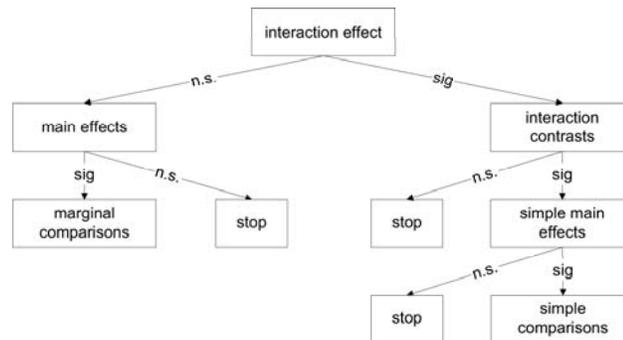
```

		Anxiety		
		Low	Medium	High
		1	-1/2	-1/2
Sex	Male	1	-1/2	-1/2
	Female	-1	1/2	1/2

It is easier to answer multiple comparisons than simple main effects. Simple main effects are made-up of simple comparisons just like main effects are made-up of marginal comparisons.

- is there a significant difference on math scores among anxiety levels for male college students? (simple main effect)

Figure 7.7: General Linear Model - Flow Chart



GLM

```

mathsco by sex anxiety
/lmatrix = "simple main effect of anxiety for male"
anxiety 1 -1 0
sex*anxiety 0 0 0 1 -1 0;
anxiety 0 1 -1
sex*anxiety 0 0 0 0 1 -1;
/design = sex anxiety sex*anxiety.

```

- is there a significant difference on math scores between male and female college students for medium anxiety levels? (simple main effect)

GLM

```

mathsco by sex anxiety
/lmatrix = "simple main effect of sex"
sex 1 -1
sex*anxiety 0 1 0 0 -1 0
/design = sex anxiety sex*anxiety.

```

- is there a significant difference on math scores between low and medium anxiety levels for female college students? (simple comparisons)

GLM

```

mathsco by sex anxiety

```

Figure 7.8: General Linear Model - Custom Hypothesis Tests

Contrast Results (K Matrix)^a

Contrast		Dependent Variable	
		MATHSCO	
L1	Contrast Estimate	10.937	
	Hypothesized Value	0	
	Difference (Estimate - Hypothesized)	10.937	
	Std. Error	2.799	
	Sig.	.000	
	95% Confidence Interval for Difference	Lower Bound Upper Bound	5.370 16.504

a. Based on the user-specified contrast coefficients (L') matrix: interaction contrast

Test Results

Dependent Variable: MATHSCO

Source	Sum of Squares	df	Mean Square	F	Sig.
Contrast	448.549	1	448.549	15.264	.000
Error	2468.420	84	29.386		

```

/lmatrix = "simple comparison"
anxiety 1 -1 0
sex*anxiety 0 0 0 1 -1 0
/design = sex anxiety sex*anxiety.

```

- is there a significant difference on math scores between low and average of medium and high anxiety levels for male students? (simple comparison)

```

GLM
mathsco by sex anxiety
/lmatrix = "simple comparison"
anxiety 1 -1/2 -1/2
sex*anxiety 1 -1/2 -1/2 0 0 0
/design = sex anxiety sex*anxiety

```

- is there a significant difference on math scores between low and average of medium and high anxiety levels averaged across sex? (marginal comparison)

```

GLM
mathsco by sex anxiety
/lmatrix = "marginal comparison"
anxiety 1 -1/2 -1/2
sex*anxiety 1/2 -1/4 -1/4 1/2 -1/4 -1/4
/design = sex anxiety sex*anxiety.

```

7.4 Contrast Matrices

7.4.1 Deviation Coding

Compares the effect for each category of the IV, except one, to the grand mean. Select either first or last as the omitted category.

For example with 3 groups: $b_1 = \bar{y}_1 - \bar{y}$ and $b_2 = \bar{y}_2 - \bar{y}$.

7.4.2 Difference Contrast

The effect for each category of the IV except the first is compared to the average of the previous categories.

For example with 3 groups: $b_1 = \bar{y}_2 - \bar{y}_1$ and $b_2 = \bar{y}_3 - \bar{y}_1 + \bar{y}_2/2$.

7.4.3 Helmert Contrast

The effect for each category of the IV except the last is compared to the average of subsequent categories.

For example with 3 groups: $b_1 = \bar{y}_1 - \bar{y}_2 + \bar{y}_3/2$ and $b_2 = \bar{y}_2 - \bar{y}_3$.

7.4.4 Simple Contrast

Each category except one of the IV is compared to a reference category. Select either first or last as the reference category.

For example with 3 groups: $b_1 = \bar{y}_2 - \bar{y}_1$ and $b_2 = \bar{y}_3 - \bar{y}_1$.

7.4.5 Repeated

Compares adjacent categories. Each category of the predictor variable or factor except the first category is compared to the category that proceed it.

For example with 3 groups: $b_1 = \bar{y}_1 - \bar{y}_2$ and $b_2 = \bar{y}_2 - \bar{y}_3$.

Chapter 8

Within and Mixed Analysis of Variance

8.1 Assumptions

- Normality: the dependent variable (technically speaking, the errors) should be normally distributed within each cell. Usually ANOVA is robust against violations of normality as long as a variable is skewed in the same direction across all cells of the design.
- Homogeneity of variance: variances of the dependent variable are same in each cell in the design. ANOVA must meet this assumption. There is no way around this violation of assumption.
- Homogeneity of Covariance: covariances (relationships) among the dependent variable at different levels of within-subjects independent variable is same.
- Outliers: no univariate outliers. Outliers will bias results.
- Sample size: good rule of thumb is minimum of 10 subjects per cell. also, the ratio between largest sample size to smallest size should be less than 3:1.

8.2 Within-Subjects ANOVA

Both within-subjects and mixed ANOVA are performed using the same procedure in SPSS (GLM → Repeated Measures).

If there are only one within-subjects IV (factor), then only define one within-subjects factor at Define dialog box (Figure 8.1). This dialog box does offer ability to define multiple within-subjects IV.

One difference between a between-subjects and within-subjects ANOVA as far as procedure in SPSS is in post hoc comparisons. The usual Tukey and Scheffé are not available for within-subject IV.

8.3 Mixed ANOVA

Mixed ANOVA is when there are combination of both within-subjects and between-subjects IV.

8.3.1 Example of the Design

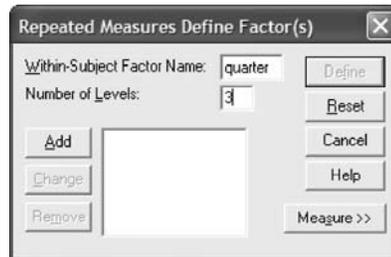
A mixed analysis of variance is demonstrated using a simple dataset.

		Quarter		
		Fall	Winter	Spring
Sex	Male			
	Female			

8.3.2 Setup in SPSS

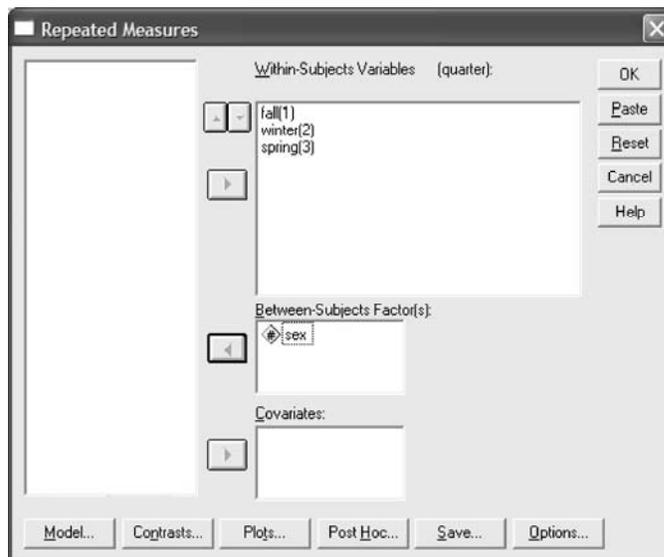
- Click on Analyze → General Linear Model → Repeated Measures. Fill in the Within-Subject Factor name with Number of Levels and click Add then Define (Figure 8.1).

Figure 8.1: Repeated Measures Define



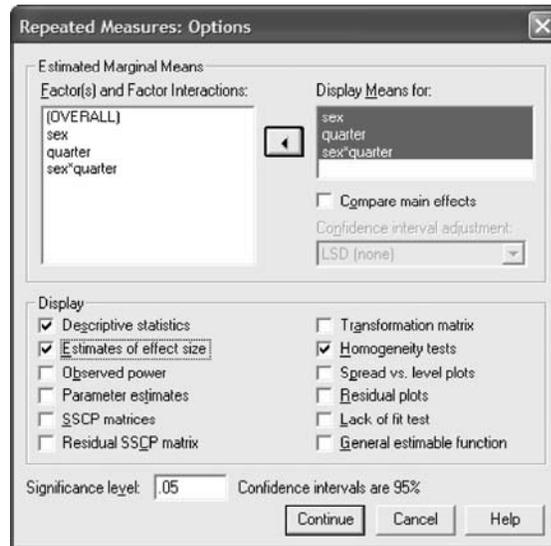
- Repeated Measure dialog box is displayed next (Figure 8.2). Move over Sex into the Between-Subjects Factor and fall to spring into Within-Subjects Variables box.

Figure 8.2: Repeated Measures



- Click Options (Figure 8.3). Check Descriptive Statistics, Estimate of effect size (will report partial η^2), and homogeneity tests (will report Levene's test of homogeneity of variance).

Figure 8.3: Repeated Measures - Options



- Click Plots from the main window (Figure 8.4). Move over the variables in to the middle. Usually, move over the within-subject independent variable as Horizontal Axis. For this example, move over Quarter into Horizontal Axis and Sex as Separate Lines. Don't forget to click on Add button at the bottom. SPSS is capable of generating multiple plots.

8.3.3 Output in SPSS

SPSS report both multivariate and univariate results whenever within-subjects (repeated) independent variable is used. A researcher must decide a priori, whether to perform an analysis univariately or multivariately and ignore the other outputs.

Mauchly's Test of Sphericity (Figure 8.5) is a test of homogeneity of covariance. The test should be not significant. This table also displays Epsilon for adjustment to degrees of freedom for violations of homogeneity of covariance.

If the assumption of homogeneity of covariance is met, interpret the sphericity

Figure 8.4: Repeated Measures - Plots

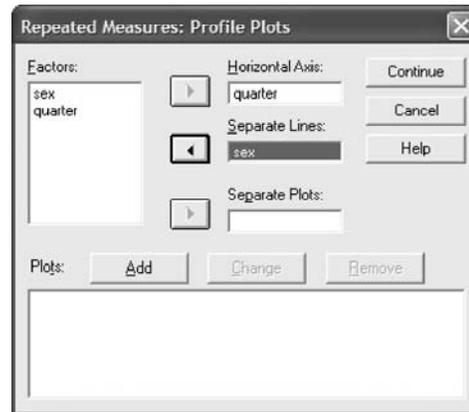


Figure 8.5: Mixed ANOVA - Test of Sphericity

Mauchly's Test of Sphericity^b

Measure: MEASURE_1

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon ^a		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
QUARTER	.737	8.240	2	.016	.792	.861	.500

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

b. Design: Intercept+SEX
Within Subjects Design: QUARTER

assumed row in Figure 8.6 else other rows that adjust degrees of freedom. The between-subjects effect is not printed in the same table (Figure 8.7).

8.3.4 Research hypotheses and How to test them

Starting Questions

- Is there a significant difference on math scores between male and female college students averaged across three quarters? (Main effect of Sex)
- Is there a significant difference on math scores among three quarters aver-

Figure 8.6: Mixed ANOVA - Within Subjects Effects

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
QUARTER	Sphericity Assumed	368.553	2	184.276	5.849	.005	.173
	Greenhouse-Geisser	368.553	1.584	232.745	5.849	.009	.173
	Huynh-Feldt	368.553	1.723	213.952	5.849	.007	.173
	Lower-bound	368.553	1.000	368.553	5.849	.022	.173
QUARTER * SEX	Sphericity Assumed	2489.096	2	1244.548	39.501	.000	.585
	Greenhouse-Geisser	2489.096	1.584	1571.894	39.501	.000	.585
	Huynh-Feldt	2489.096	1.723	1444.965	39.501	.000	.585
	Lower-bound	2489.096	1.000	2489.096	39.501	.000	.585
Error(QUARTER)	Sphericity Assumed	1764.391	56	31.507			
	Greenhouse-Geisser	1764.391	44.338	39.794			
	Huynh-Feldt	1764.391	48.233	36.581			
	Lower-bound	1764.391	28.000	63.014			

Figure 8.7: Mixed ANOVA - Between Subjects Effects

Tests of Between-Subjects Effects

Measure: MEASURE_1
Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	477575.697	1	477575.697	18993.706	.000	.999
SEX	39.931	1	39.931	1.588	.218	.054
Error	704.029	28	25.144			

aged across sex? (Main effect of quarter)

- Is the pattern of difference on the math scores among three quarters different between male and female college students? (Interaction of sex and quarter)

What's Next?

If you have a specific hypothesis, test it. If you do not have a specific hypothesis, there are general guideline (Figure 7.7).

- for overall analysis without any option, the syntax would be:

```
GLM
  fall winter spring by sex
  /wsfactor = quarter 3
```

```

/wsdesign = quarter
/design = sex.

```

- is the pattern of difference on math scores at fall and winter quarter different between male and female college students? (interaction contrast). /mmatrix statement is required to contrast within-subjects IV while /lmatrix contrasts between-subjects IV.

```

GLM
  fall winter spring by sex
  /wsfactor = quarter 3
  /wsdesign = quarter
  /lmatrix sex 1 -1
  /mmatrix all 1 -1 0
  /design = sex.

```

Unlike in between-subjects ANOVA, you only need to specify the marginal contrast coefficients, not contrast coefficients for each cell. Actual contrast output is similar to between-subjects output (Figure 7.8).

- is the pattern of difference on math scores at fall quarter compared to average of winter and spring quarters different between male and female college students? (interaction contrast)

```

GLM
  fall winter spring by sex
  /WSFACTOR = quarter 3
  /WSDESIGN = quarter
  /lmatrix sex 1 -1
  /mmatrix all 1 -1/2 -1/2
  /design = sex.

```

- is there a significant difference on math scores among three quarters for male college students? (simple main effect). Select only male using Select Cases (Section 2.4) and simply run a one-way within-subjects ANOVA.
- is there a significant difference on math scores between male and female college students for winter quarter? (simple main effect)

```
GLM
  fall winter spring by sex
  /wsfactor = quarter 3
  /wsdesign = quarter
  /lmatrix sex 1 -1
  /mmatrix all 0 1 0
  /design = sex.
```

- is there a significant difference on math scores between fall and winter quarters for female college students? (simple comparisons). Simply select only females and perform a paired t-test.
- is there a significant difference on math scores between fall and average of winter and spring quarters for male students? (simple comparisons). Select only male students and perform a one-way within-subjects ANOVA with following syntax.

```
GLM
  fall winter spring
  /wsfactor = quarter 3
  /wsdesign = quarter
  /mmatrix all 1 -1/2 -1/2.
```

- is there a significant difference on math scores between fall and average of winter and spring quarters averaged across sex? (Marginal comparisons). Perform a one-way within-subjects ANOVA with the following syntax:

```
GLM
  fall winter spring
  /wsfactor = quarter 3
  /wsdesign = quarter
  /mmatrix all 1 -1/2 -1/2.
```

Chapter 9

Analysis of Covariance

9.1 Introduction

Only Between-Subjects ANCOVA will be discussed in this chapter. However, ANCOVA can be performed with any design as long as they meet the following conditions.

9.1.1 What is a covariate?

A covariate is a variable that will influence or impact the dependent variable that cannot be manipulated which you want to account for in a study. A covariate is interval or ratio scale and should be normally distributed without any outliers.

9.1.2 How do you choose a covariate?

- A covariate should be correlated to a dependent variable.
- A covariate should not be correlated to independent variable(s).

- More than one covariate can be used. However, limit the number of covariates. For each covariate used, you lose a degree of freedom from the error term. If a covariate is not related to the dependent variable, you lose power. Also, there is higher chance of multicollinearity/singularity among covariates and IV.

9.1.3 Assumptions

- Same assumptions as ANOVA (Section 7.1).
- Homogeneity of Regression: the relationship between a covariate and the dependent variable is same across all cells in the study. A covariate and the dependent variable must be linearly related.
- If more than one covariate is used, multicollinearity/singularity among covariates and IV.

9.1.4 If a covariate is not

- interval/ratio scale - just use the variable as another IV, randomized block design.
- Linearly related to DV - one option would be to categorize a covariate and use it as another IV, randomized block design.

9.2 Example of the Design

A between-subjects ANCOVA is demonstrated using a simple dataset similar to ANOVA example (Section 7.3.1).

		Anxiety		
		Low	Medium	High
Sex	Male			
	Female			

Let say, we performed a 3X2 between-subjects analysis of covariance to determine the effect of sex and anxiety on math scores of college students after adjusting for amount of study time. It has two independent variables: sex with 2 levels (male, female) and anxiety levels (low, medium, high). The dependent variable is math scores. A covariate is amount of study time.

9.3 Setup in SPSS

- Click on Analyze → General Linear Model → Univariate (Figure 7.2).
- Fill in the variables. Move over the covariate to Covariate(s) box.
- Click Options (Figure 7.3). Choose all the same settings from Section 7.3.2 plus Parameter estimates.
- choose the same settings for Plots.

9.4 Output in SPSS

The Descriptive Statistics (Figure 9.1 displayed at the beginning is the unadjusted (observed) means.

In ANCOVA, Levene's test (Figure 9.2) is a test of both homogeneity of variance and homogeneity of regression. Unfortunately, if it is significant, you don't know if there is a violation of homogeneity of variance, homogeneity of regression, or both. Further test would be required. A test of homogeneity of regression can be performed by computing the regression coefficient(s) for each cell in the design and comparing them (or by computing and testing the interaction effect of covariate and IV).

Interpretation of the Results based on Figure 9.3:

- The amount of study time was significantly predict math scores.

Figure 9.1: ANCOVA - Descriptive Statistics

Descriptive Statistics

Dependent Variable: MATH

SEX	ANXIETY	Mean	Std. Deviation	N
male	low	8.95	3.300	20
	medium	9.61	3.696	18
	high	11.00	2.784	17
	Total	9.80	3.341	55
female	low	9.25	4.297	16
	medium	8.28	2.675	18
	high	8.33	3.395	18
	Total	8.60	3.443	52
Total	low	9.08	3.722	36
	medium	8.94	3.251	36
	high	9.63	3.353	35
	Total	9.21	3.429	107

Figure 9.2: ANCOVA - Levene's Test

Levene's Test of Equality of Error Variances^a

Dependent Variable: MATH

F	df1	df2	Sig.
2.175	5	101	.063

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.
a. Design: Intercept+STIME+SEX+ANXIETY+SEX * ANXIETY

- There was no significant difference on the math scores between male and females after adjusting for the amount of study time.
- There was a significant difference on the math scores among the levels of anxiety after adjusting for the amount of study time.
- The effect of anxiety on the math scores was not different between males and females after adjusting for the amount of study time.

According to Parameter Estimates (Figure 9.4), the math scores was predicted to go down by .652 points for each increase in the study time. Also, the amount of study time explains 36.7% of the variance of the math scores. Ignore other

Figure 9.3: ANCOVA - Table

Tests of Between-Subjects Effects

Dependent Variable: MATH

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	512.642 ^a	6	85.440	11.650	.000	.411
Intercept	3123.969	1	3123.969	425.949	.000	.810
STIME	424.425	1	424.425	57.870	.000	.367
SEX	20.380	1	20.380	2.779	.099	.027
ANXIETY	60.990	2	30.495	4.158	.018	.077
SEX * ANXIETY	19.037	2	9.519	1.298	.278	.025
Error	733.414	100	7.334			
Total	10332.000	107				
Corrected Total	1246.056	106				

a. R Squared = .411 (Adjusted R Squared = .376)

parameter estimates (for now). SPSS uses indicator coding by default. In factorial design, they usually do not test the desired effects.

The estimated marginal means printed at the end is the adjusted means. The adjusted means are also used in a plot instead of observed means.

Figure 9.4: ANCOVA - Parameter Estimates

Parameter Estimates

Dependent Variable: MATH

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval		Partial Eta Squared
					Lower Bound	Upper Bound	
Intercept	14.020	.983	14.263	.000	12.070	15.970	.670
STIME	-.652	.086	-7.607	.000	-.822	-.482	.367
[SEX=1]	-2.176	1.115	-1.951	.054	-4.389	3.679E-02	.037
[SEX=2]	0 ^a	-	-	-	-	-	-
[ANXIETY=1]	.935	.931	1.005	.318	-.911	2.781	.010
[ANXIETY=2]	.488	.906	.539	.591	-1.309	2.284	.003
[ANXIETY=3]	0 ^a	-	-	-	-	-	-
[SEX=1] *							
[ANXIETY=1]	2.365	1.468	1.611	.110	-.547	5.278	.025
[SEX=1] *							
[ANXIETY=2]	1.155	1.327	.871	.386	-1.477	3.788	.008
[SEX=1] *							
[ANXIETY=3]	0 ^a	-	-	-	-	-	-
[SEX=2] *							
[ANXIETY=1]	0 ^a	-	-	-	-	-	-
[SEX=2] *							
[ANXIETY=2]	0 ^a	-	-	-	-	-	-
[SEX=2] *							
[ANXIETY=3]	0 ^a	-	-	-	-	-	-

a. This parameter is set to zero because it is redundant.