Example Using Missing Data

Creating the Missing Data Variable (Miss)

Here is a data set (achieve subset MANOVAmiss.sav) with the actual missing data on the outcomes. We can use the Compute and Select If command to create the missing data variable (Miss). Refer to the end of this handout for instructions.

<table>
<thead>
<tr>
<th>Table 1. miss</th>
<th>Frequency</th>
<th>Percent</th>
<th>Valid Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Valid 0</td>
<td>150</td>
<td>25.9</td>
<td>25.9</td>
<td>25.9</td>
</tr>
<tr>
<td>1</td>
<td>430</td>
<td>74.1</td>
<td>74.1</td>
<td>100.0</td>
</tr>
<tr>
<td>Total</td>
<td>580</td>
<td>100.0</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>

Here we can see about 26% of the cases contain at least one missing value on the dependent variables.

Now we can use Miss as a dependent variable in the model and the other predictors can be added to the model. I created an interaction term for the predictors to see if there might be some relationship with subsets of the predictors and the outcomes. I created this by using Compute and put lowses*female in the dialog box. This saves an interaction term (femlowses) in the data set. Then I used logistic regression to see whether there was a relationship between the predictors and the pattern of missing data on the outcomes. (Refer to the end of this handout for instructions to create the lowses*female interaction and logistic regression.)

As you can see in this case, there is no significant predictor such that we can likely conclude the missing values are MAR.

<table>
<thead>
<tr>
<th>Table 2. Variables in the Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>Step 1a</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

a. Variable(s) entered on step 1: female, lowses, femlowses.

1 Three data sets are used for this handout: achieve subset MANOVAmiss.sav; achievementMANOVAvertical.sav; and achieve subset MANOVA.sav data file.
Using Mixed Modeling

If we use this data set with a vertical format, we can compare the estimates with the full vertical data set we will be using next week. Here are the results of the data set with missing values included under the assumption of MAR. (Refer to the end of this handout for instructions.)

### Table 3. Model Dimension

<table>
<thead>
<tr>
<th></th>
<th>Number of Levels</th>
<th>Covariance Structure</th>
<th>Number of Parameters</th>
<th>Subject Variables</th>
<th>Number of Subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed Effects</td>
<td>Index1</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>female</td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>lowses</td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>female * lowses</td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Repeated Effects</td>
<td>Index1</td>
<td>3</td>
<td>Unstructured</td>
<td>6</td>
<td>580</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Correlations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>9</td>
<td></td>
<td>12</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Dependent Variable: achieve.

First, we can see above that analysis retains all 580 individuals. This is useful in adding to the assumption that the values that are missing are missing at random.

Next are the resulting model estimates (using one fixed effect for each predictor). You can compare the results in the table below to Table 13 (see p. 12) in the handout on multivariate models I gave you last week. We can see the intercepts are a bit different (i.e., they are all a bit higher in the table below). The effect for female is 8.19 versus 11.87 in the other handout. For low SES the estimate below is -21.99 versus -19.36. The estimate for the interaction is -1.84 against -1.96 in the other handout. The pattern of results is the same however; that is, the main effects are significant and the interaction is not.

### Table 4. Estimates of Fixed Effects

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>df</th>
<th>t</th>
<th>Sig.</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Index1=1]</td>
<td>654.899928</td>
<td>2.368743</td>
<td>633.367</td>
<td>276.476</td>
<td>.000</td>
<td>650.248388-659.551468</td>
</tr>
<tr>
<td>[Index1=2]</td>
<td>669.553217</td>
<td>2.506249</td>
<td>715.416</td>
<td>267.153</td>
<td>.000</td>
<td>664.632734-674.473700</td>
</tr>
<tr>
<td>[Index1=3]</td>
<td>656.533663</td>
<td>2.328033</td>
<td>595.802</td>
<td>282.012</td>
<td>.000</td>
<td>651.961514-661.105812</td>
</tr>
<tr>
<td>female</td>
<td>8.190010</td>
<td>3.298291</td>
<td>574.605</td>
<td>2.483</td>
<td>.013</td>
<td>1.711832-14.668188</td>
</tr>
<tr>
<td>female * lowses</td>
<td>-1.836824</td>
<td>4.671193</td>
<td>574.279</td>
<td>- .393</td>
<td>.694</td>
<td>-11.011530 7.337882</td>
</tr>
</tbody>
</table>

a. Dependent Variable: achieve.
MANOVA Results

Following are the MANOVA results (based on only 430 individuals). (Refer to the end of this handout for instructions.) Notably, in the MANOVA results we can see that the interaction is now significant – which is likely the function of missing 26% of the data.

<table>
<thead>
<tr>
<th>Effect</th>
<th>Value</th>
<th>F</th>
<th>Hypothesis df</th>
<th>Error df</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>.998</td>
<td>74324.545b</td>
<td>3.000</td>
<td>424.000</td>
<td>.000</td>
</tr>
<tr>
<td>Wilks' Lambda</td>
<td>.002</td>
<td>74324.545b</td>
<td>3.000</td>
<td>424.000</td>
<td>.000</td>
</tr>
<tr>
<td>Hotelling's Trace</td>
<td>525.881</td>
<td>74324.545b</td>
<td>3.000</td>
<td>424.000</td>
<td>.000</td>
</tr>
<tr>
<td>Roy's Largest Root</td>
<td>525.881</td>
<td>74324.545b</td>
<td>3.000</td>
<td>424.000</td>
<td>.000</td>
</tr>
<tr>
<td>female</td>
<td>.028</td>
<td>3.998b</td>
<td>3.000</td>
<td>424.000</td>
<td>.008</td>
</tr>
<tr>
<td>Wilks' Lambda</td>
<td>.972</td>
<td>3.998b</td>
<td>3.000</td>
<td>424.000</td>
<td>.008</td>
</tr>
<tr>
<td>Hotelling's Trace</td>
<td>.028</td>
<td>3.998b</td>
<td>3.000</td>
<td>424.000</td>
<td>.008</td>
</tr>
<tr>
<td>Roy's Largest Root</td>
<td>.028</td>
<td>3.998b</td>
<td>3.000</td>
<td>424.000</td>
<td>.008</td>
</tr>
<tr>
<td>lowses</td>
<td>.182</td>
<td>31.448b</td>
<td>3.000</td>
<td>424.000</td>
<td>.000</td>
</tr>
<tr>
<td>Wilks' Lambda</td>
<td>.818</td>
<td>31.448b</td>
<td>3.000</td>
<td>424.000</td>
<td>.000</td>
</tr>
<tr>
<td>Hotelling's Trace</td>
<td>.223</td>
<td>31.448b</td>
<td>3.000</td>
<td>424.000</td>
<td>.000</td>
</tr>
<tr>
<td>Roy's Largest Root</td>
<td>.223</td>
<td>31.448b</td>
<td>3.000</td>
<td>424.000</td>
<td>.000</td>
</tr>
<tr>
<td>female * lowses</td>
<td>.024</td>
<td>3.472b</td>
<td>3.000</td>
<td>424.000</td>
<td>.016</td>
</tr>
<tr>
<td>Wilks' Lambda</td>
<td>.976</td>
<td>3.472b</td>
<td>3.000</td>
<td>424.000</td>
<td>.016</td>
</tr>
<tr>
<td>Hotelling's Trace</td>
<td>.025</td>
<td>3.472b</td>
<td>3.000</td>
<td>424.000</td>
<td>.016</td>
</tr>
<tr>
<td>Roy's Largest Root</td>
<td>.025</td>
<td>3.472b</td>
<td>3.000</td>
<td>424.000</td>
<td>.016</td>
</tr>
</tbody>
</table>

- a. Design: Intercept + female + lowses + female * lowses
- b. Exact statistic

This procedure outlined would be one way of dealing with considerable missing data, as some individuals have complete data on the outcomes, while others are missing one or two values. You can see, however, that the vertical format retained all the data, while the MANOVA format would be missing data on 26% of the study’s participants. Of course, in this case I only created missing data on the dependent variable. Missing values on the predictors would contribute possible bias. Often, however, the dependent variable is more likely to have missing data that are harder to “work around” in terms of the data analysis. The point is that we assessed the proportion of missing cases, then developed a strategy to deal with the possible effects of the missing data. We also compared alternative ways of examining the data (MANOVA versus Mixed Modeling) and compared the differences in the quality of the analyses and results using both approaches. My guess is that a journal would accept this as a legitimate way of dealing with the missing data problem in this instance.
Creating the Missing Data Variable (Miss) with IBM SPSS Menu Commands

| IBM SPSS Syntax: | IF (read >= 0 & math >= 0 & lang >= 0) Miss = miss = 1. |

Launch the IBM SPSS program* then go to the toolbar to select File > Open > Data to open the achieve subset MANOVAmiss.sav data file.

1. Go to the toolbar and select TRANSFORM, COMPUTE VARIABLE.

This command will open the Compute Variable dialog box.

*Note: Instructions and screenshots based on IBM SPSS version 22 and SPSS Classic display setting.

2a. Enter Miss as the Target Variable.

b. We will set-up a condition for case selection so click the “If” button to open the Compute Variable: If Cases dialog box.
c. Click “include if case satisfied condition”.

d. Type into the dialog box:
   read $\geq 0 \&$ math $\geq 0 \&$
   lang $\geq 0$

   This will identify individuals that have no missing data on any outcome. Using “&” gives the message that each one has to have no missing data.

Click the CONTINUE button to close this dialog box and return to the Compute Variable main screen.

e. In the Compute Variable dialog box type:
   miss $= 1$

   This will tell the computer that if there is no missing data on read, math, or language then assign a 0 to the new variable “miss.” Anyone who does not fulfill that condition (i.e., that has missing data on either reading, math, or language will be assigned a 0.

Click the OK button to perform the function.
3. A warning appears notifying you that Miss will over-write the pre-existing miss variable in the data set.

Click the OK button.
Creating the Interaction \((femlowses)\) Term with IBM SPSS Menu Commands

**SPSS Syntax:**

```
COMPUTE femlowses=female * lowses.
```

Continue using the *achieve subset MANOVAmiss.sav* data file.

1. Go to the toolbar and select **TRANSFORM, COMPUTE VARIABLE.**

   This command will open the **Compute Variable** dialog box.

2a. Click the **RESET** button to clear prior settings if you are continuing after computing the *Miss* variable.

   b. Enter *femlowses* as the **Target Variable**.

   c. Click to select *female* then click the right arrow button (or drag the variable) to place it in the **Numeric Expression** box.

   d. Click the asterick button to insert the multiplier sign.

   e. Click to select *lowsess* then click the right arrow button (or drag the variable) to place it in the **Numeric Expression** box.

   Click the **OK** button to perform the function.
3. A warning appears notifying you that Miss will overwrite the pre-existing miss variable in the data set.

Click the OK button.
Defining Logistic Regression Model (Table 2) with IBM SPSS Menu Commands

**IBM SPSS Syntax:**

```
LOGISTIC REGRESSION VARIABLES miss
/METHOD=ENTER female lowses femlowses
/CONTRAST (female)=Indicator
/CRITERIA=PIN(.05) POUT(.10) ITERATE(20) CUT(.5).
```

Continue using the *achieve subset MANOVAmiss.sav* data file.

1. Go to the toolbar and select ANALYZE, REGRESSION. BINARY LOGISTIC.

   This command will open the Logistic Regression dialog box.

2a. We will designate *miss* as the dependent variable. Click to select *miss* then click the right arrow button (or drag the variable) to place it in the Dependent box.

   b. We will designate three predictors for the model. Click to select *female*, *lowses*, and *female* then click the right arrow button (or drag the variables) to place them in the Covariates box.
c. To specify *female* as a categorical variable first click the CATEGORICAL button.

d. Click to select *female* then click the right arrow button (or drag the variable) to place it in the *Categorical Covariates* box.

Click the CONTINUE button to close this dialog box and return to the *Logistic Regression* main screen.

3. Click the OK button to perform the function.
Defining MIXED Model (Tables 3, 4) with IBM SPSS Menu Commands

**IBM SPSS syntax:**

```
MIXED achieve BY Index1 WITH female lowses
/FIXED=Index1 female lowses female*lowses | NOINT
SSTYPE(3)
/METHOD=REML
/PRINT=G SOLUTION TESTCOV
/REPEATED=Index1 | SUBJECT(id) COVTYPE(UNR).
```

Launch the IBM SPSS program application and select the *achievementMANOVAvvertical.sav* data file.

1. Go to the toolbar and select **ANALYZE, MIXED MODELS, LINEAR.**

This command enables access to the *Linear Mixed Models: Specified Subjects and Repeated* dialog box...
2. The Linear Mixed Models: Specify Subjects and Repeated screen displays options for defining variables as subjects, repeated observations, and type of covariance structure in a model.

a. A subject is an observational unit that may be independent of other subjects. For this model we will designate \textit{id} as a subject identifier for the model. Click to select \textit{id} then click the right arrow button (or drag the variable) into the Subjects box.

b. The Repeated box allows specifying variables that identify repeated observations. For this model, \textit{Index1} identifies three test scores nested within individuals. Click to select \textit{Index1} then click the right arrow button (or drag the variable) to move the variable into the Repeated box.

The combination of values for \textit{id} and \textit{Index1} defines a particular student across 3 test scores.

c. The Repeated Covariance Type specifies a model’s covariance structure. For this model we will use the \textit{Unstructured Correlation Metric} (UNR) covariance type. Click the pull-down menu and select: \textit{Unstructured Correlation Metric}.

The \textit{Unstructured Correlation Metric} covariance structure has heterogeneous variances and correlations (IBM SPSS, 2011).

Click the CONTINUE button to display the Linear Mixed Models dialog box.
3. The **Linear Mixed Models** main screen enables specifying the dependent variable, factors, covariates, and access to dialog boxes for defining *Fixed* and *Random* effects, and options for *Estimation*, *Statistics*, *EM Means*, and *Save*.

a. For this model we will use *achieve* as the dependent variable. Click to select *achieve* from the left column listing. Then click the right arrow button (or drag) to transfer *achieve* into the *Dependent Variable* box.

b. Click to select *Index1* then click (or drag the variable) into the *Factor(s)* box.

c. Click to select *female* and *lowses* then click (or drag the variables) into the *Covariate(s)* box.

We will next add fixed effects to this model. Click the FIXED button to access the *Linear Mixed Models: Fixed Effects* dialog box.
4a. Within the *Linear Mixed Models: Fixed Effects* dialog box, we will use the default *Build terms* setting.

b. Click the pull-down menu to change the factorial setting to *Main Effects*.

c. Now click to select the 3 variables (*Index1, female, lowses*) from the *Factors and Covariates* box then click the ADD button (or drag the variables) to move them into the *Model* box.

**Interaction: female*lowses**

d. We will add one cross-level interaction (or nested term) to the model: *female*lowses. The interaction will tell us if gender and socioeconomic status predicts student achievement. First click the pull-down menu and select: *Interaction*.

e. To create the interaction click to select *female* and *lowses* then click the ADD button (or drag the variables). The interaction *female*lowses is displayed in the *Model* box.

f. Since this will be a “no intercept' model, de-select the default *Include intercept* option.

Click the CONTINUE button to return to the *Linear Mixed Models* dialog box.
5. In the *Linear Mixed Models* dialog box, click the ESTIMATION button to access the *Linear Mixed Models: Estimation* dialogue box. The estimation method choices are maximum likelihood (ML) or restricted maximum likelihood (REML). In ML, both regression coefficients and variance components are included in maximizing the likelihood function; that is, the process of minimizing the difference between the sample covariance matrix and the model-implied covariance matrix. In REML, only the variance components are included in estimating the likelihood function; thus, REML is a restricted solution.

With sufficient numbers of groups, the differences in estimation methods are negligible. Restricted maximum likelihood (REML), the default estimation method for both IBM SPSS techniques, which also is the better choice with small data sets will be used to develop the variance component estimates.

For this model we will use the default REML to estimate the models, which facilitates making comparisons between successive models using model fit criteria (Hox, 2002).

Click the CONTINUE button to return to the *Linear Mixed Models* dialog box.
6. In the **Linear Mixed Models** dialog box, click the STATISTICS button to access the **Linear Mixed Models: Statistics** dialog box.

   Click and select the following three statistics to be included in the output: *Parameter estimates, tests for covariance parameters, and covariances of random effects.*

   Click the CONTINUE button to return to the **Linear Mixed Models** dialog box.

7. Finally, in the **Linear Mixed Models** dialog box, click the OK button to run the model.
Defining MANOVA Model (Table 5) with IBM SPSS Menu Commands

**IBM SPSS syntax:**

```
USE ALL.
COMPUTE filter_$=(missing = 1).
VARIABLE LABELS filter_$ 'missing = 1 (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMATS filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE.

GLM read math lang BY female lowses
/METHOD=SSTYPE(3)
/INTERCEPT=INCLUDE
/EMMEANS=TABLES(OVERALL)
/PRINT=DESCRIPTIVE
/CRITERIA=ALPHA(.05)
/DESIGN=female lowses female*lowses.
```

Go to the toolbar to select File > Open > Data to open the *achieve subset MANOVA.sav* data file.

1. Go to the toolbar and select DATA, SELECT CASES.

This command will open the *Select Cases* dialog box.
2a. Within the *Select Cases* dialog box, select the option *if condition is satisfied.*

b. Click the *If* button to access the *Multivariate: Options* dialog box.

c. In the *Select Cases* main dialog box select *missing* then click the right arrow button (or drag the variable) to move it into box.

d. Click the equal sign.

e. Click the number “1” to complete the condition:  
   missing = 1

Click the CONTINUE button to close this dialog box and return to the *Select Cases* main screen.
The condition `missing = 1` appears in the dialog box.

Click the OK button to close the main `Select Cases` dialog box and return to the SPSS main screen.
3. Go to the toolbar, select ANALYZE, GENERAL LINEAR MODEL, MULTIVARIATE.

This command opens the Generalized Linear Model Multivariate main dialog box.

4. In the Multivariate main dialog box we will select 3 variables (read, math, and lang) as dependent variables.

a. Click to select read, math, and lang then click the right arrow button to move them into the Dependent Variables box.

b. Click to select female and lowses then click the right arrow button to move them into the Fixed Factor(s) Variables box.

Now click the MODEL button to access the Multivariate: Model dialog box for designating various statistical output.
5a. The **Multivariate: Model**. Click the pull-down menu to change the factorial setting to **Main Effects**.

b. Now click to select *female* and *lowses* from the **Factors and Covariates** box then click the right arrow button (or drag the variables) to move them into the **Model** box.

**Interaction: female*lowses**

We will add one cross-level interaction (or nested term) to the model: *female*lowses.

c. First click the pull-down menu and select: **Interaction**

d. To create the interaction click to select *female* and *lowses* then then click the right arrow button (or drag the variables) to move them into the **Model** box.

Click the CONTINUE button to return to the **Multivariate** main dialog box.
6. Now click the OPTIONS button to access the Multivariate: Options dialog box for designating various statistical output.

   a. The Multivariate: Options dialog box provides various statistical options. For this example we want to have the means displayed in the output.

      Click to select (OVERALL) then click the right arrow button (or drag the factor) to the Display Means for box.

   b. For this example we want to have the model’s descriptive statistics displayed in the Multivariate Tests portion of the output. Click to select Descriptive Statistics.

      Click the CONTINUE button to return to the Multivariate main dialog box.

7. From the Multivariate main dialog box click the OK button to generate the output results.