Linear Regression in SPSS

Data: mangunkill.sav
Goals:
- Examine relation between number of handguns registered (nhandgun) and number of man killed (mankill)
- Model checking
- Predict number of man killed using number of handguns registered

I. View the Data with a Scatter Plot

To create a scatter plot, click through Graphs > Scatter > Simple > Define. A Scatterplot dialog box will appear. Click Simple option and then click Define button. The Simple Scatterplot dialog box will appear. Click mankill to the Y-axis box and click nhandgun to the x-axis box. Then hit OK. To see fit line, double click on the scatter plot, click on Chart > Options, then check the Total box in the box labeled Fit Line in the upper right corner.
II. Regression Analysis

To perform the regression, click on Analyze\Regression\Linear. Place nhandgun in the Dependent box and place mankill in the Independent box. To obtain the 95% confidence interval for the slope, click on the Statistics button at the bottom and then put a check in the box for Confidence Intervals. Hit Continue and then hit OK.

The independent variable (nhandgun) is said to be useful in predicting the dependent variable (mankill) when the level of significance (P-value labeled with Sig. on the Output) is below 0.05.
III. Residual Analysis

To assess the fit of the model, when performing the regression, also click on the Save button at the bottom of the dialogue box. In the boxes labeled Predicted Values and Residuals, click Unstandardized in both boxes. In the Hit Continue and then hit OK. There will now be a column labeled res_1 and a column labeled pre_1 in the data window. To make a residual plot, click Graphs\Scatter\Simple\Define and put res_1 in the Y-axis box and again put pre_1 in the X-axis box. Hit OK.

The goal of a residual plot is to see a random scatter of residuals. The normality test in the Explore… option can be used to check for normality.

IV. Prediction Intervals

To calculate the mean prediction intervals and the individual prediction intervals, use the Save button that appears after clicking Analyze\Regression\Linear. Now in the box labeled Prediction Values, click on Unstandardized. This will give the predicted Y-values from the model. The data window will have a column labeled pre_1. For the prediction intervals, in the boxes near the bottom labeled Prediction Intervals, put check marks in front of Mean and Individual. In the data window, will now be columns, labeled lmci_1, umci_1, lici_1, and uici_1. The first two columns are for the lower and upper bounds for the 95% mean prediction interval. (We are 95% sure that the average y for the given x values is within that interval.) The second two columns are the 95% prediction intervals. To do a prediction, simply enter the value of the predictor variable at the last row of the data sheet under the predictor variable and go through the model building. For instance, to predict the average number of people killed if the number of handguns registered is 700. The confidence interval for mean response is (41.49, 50.45), and for an individual response is (35.63, 56.31).
Multiple Regression in SPSS

Data: WeightbyAgeHeight.sav

Goals:
- Examine relation between weight (response) and age and height (explanatory)
- Model checking
- Predict weight

I. View the Data with a Scatter Plot

To create a scatter plot, click through Graphs\Scatter\Simple\Define. A Scatterplot dialog box will appear. Click Matrix option and then click Define button. The Simple Scatterplot dialog box will appear. Click age, weight and height variables into matrix variables box.

II. Regression Analysis

To perform the regression, click on Analyze\Regression\Linear. Place weight in the Dependent box and place age and height in the Independent box. Click the Statistics button to select the collinearity diagnostics and click Continue. To obtain the 95% confidence interval for the slope, click on the Statistics button at the bottom and then put a check in the box for Confidence Intervals. Hit Continue and then hit OK.

The independent variables (age, height) are useful in predicting the dependent variable (weight) when the level of significance (P-value labeled with Sig. on the Output) is below 0.05.

<table>
<thead>
<tr>
<th>Model</th>
<th>(Constant)</th>
<th>Age</th>
<th>Height</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
</tr>
<tr>
<td></td>
<td>-127.820</td>
<td>12.099</td>
<td>-10.565</td>
</tr>
<tr>
<td>1</td>
<td>.240</td>
<td>.055</td>
<td>.228</td>
</tr>
<tr>
<td></td>
<td>3.090</td>
<td>.257</td>
<td>.627</td>
</tr>
</tbody>
</table>

a. Dependent Variable: Weight
Stepwise Regression

To perform stepwise regression for automatically selecting significant variables, check the Method drop down list and choose the desired one and click OK. SPSS will produce an output table to present the final model with a coefficients table.

Interaction Term

To examine the interaction between age and height variables, first create the interaction variable (intageht). Click Transform\Compute, and in the compute variable dialog box, enter a name for the interaction term, intageht. Click and enter age variable and click “*” for multiplying and then click and enter height variable in the Numeric Expression box and then click OK.

Collinearity Diagnostics

To perform the regression, click on Analyze\Regression\Linear. Place weight in the Dependent box and place age, height and intageht in the Independent box. Click the Statistics button to select the collinearity diagnostics and click Continue, and then hit OK. In the coefficients table, VIF are all greater than 10 which implies collinearity. In this example, centralizing the variables
may reduce the possible collinearity from the interaction term that is generated by the product of the two variables.

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>Collinearity Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(Constant)</td>
<td>66.996</td>
<td>106.189</td>
</tr>
<tr>
<td></td>
<td>Age</td>
<td>-.973</td>
<td>.660</td>
</tr>
<tr>
<td></td>
<td>Height</td>
<td>-3.13E-02</td>
<td>1.710</td>
</tr>
<tr>
<td></td>
<td>INTAGEHT</td>
<td>1.936E-02</td>
<td>.010</td>
</tr>
</tbody>
</table>

a. Dependent Variable: Weight

If categorical variables are to be included in the model, the indicator variables will need to be created. Recode under transform option is one of the methods to generate the indicator variables.

### III. Residual Analysis

To assess the fit of the model, when performing the regression, also click on the Save button at the bottom of the dialogue box. In the boxes labeled Predicted Values and Residuals, click Unstandardized in both boxes. In the Hit Continue and then hit OK. There will now be a column labeled res_1 and a column labeled pre_1 in the data window. To make a residual plot, click Graphs\Scatter\Simple\Define and put res_1 in the Y-axis box and again put pre_1 in the X-axis box. Hit OK. The goal of a residual plot is to see a random scatter of residuals. The normality test in the Explore… option can be used to check for normality.

### IV. Prediction Intervals

To calculate the mean prediction intervals and the individual prediction intervals, use the Save button that appears after clicking Analyze\Regression\Linear. Now in the box labeled Prediction Values, click on Unstandardized. This will give the predicted Y-values from the model. The data window will have a column labeled pre_1. For the prediction intervals, in the boxes near the bottom labeled Prediction Intervals, put check marks in front of Mean and Individual. In the data window, will now be columns, labeled lmci_1, umci_1, lici_1, and uici_1. The first two columns are for the lower and upper bounds for the 95% mean prediction interval. (We are 95% sure that the average y for the given x values is within that interval.) The second two columns are the 95% prediction intervals. To do a prediction, simply enter the value of the predictor variable at the last row of the data sheet under the predictor variable and go through the model building. The procedure is similar to that of simple linear regression.
Logistic Regression in SPSS

Data:  logdisea.sav

Goals:
- Examine relation between disease (binary response) and other explanatory variables such as age, socioeconomic status, sector, and savings account.
- Model checking
- Predict probability of getting disease and estimating the odds ratio

To perform the regression, click on Analyze\Regression\Binary Logistic. Place disease in the Dependent box and place age, sciostat, sector and savings in the covariates box. Click the Categorical button for creating indicator variables for the categorical variables. To obtain the predicting probability of getting disease, click on the Save button at the bottom and then put a check in the Probabilities box of Predicted Values and click Continue. Also, click the Option button and check the goodness of fit box to see how well the model fit the data, and check the CI for exp(B) to obtain confidence interval for odds ratio. Hit Continue and then hit OK.

The independent variables (age, sector) are significant in predicting the dependent variable (disease) when the level of significance (P-value labeled with Sig. on the Output) is below 0.05.

### Variables in the Equation

<table>
<thead>
<tr>
<th>Step</th>
<th>AGE</th>
<th>S.E.</th>
<th>Wald</th>
<th>df</th>
<th>Sig.</th>
<th>Exp(B)</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.027</td>
<td>.009</td>
<td>8.646</td>
<td>1</td>
<td>.003</td>
<td>1.027</td>
<td>1.009</td>
<td>1.045</td>
</tr>
</tbody>
</table>

*Variable(s) entered on step 1: AGE, SCIOSTAT, SECTOR, SAVINGS.*

### Hosmer and Lemeshow Test

<table>
<thead>
<tr>
<th>Step</th>
<th>Chi-square</th>
<th>df</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10.853</td>
<td>8</td>
<td>.210</td>
</tr>
</tbody>
</table>

### Stepwise Regression

To perform stepwise regression for automatically selecting significant variables, check the Method drop down list and choose the desired one and click OK. SPSS will produce an output table to present the final model with a coefficients table.