**How to Perform a Multiple Regression Analysis in SPSS**

Multiple regression is used to assess the relationship between one dependent variable and several independent variables. In this example, the Bar Height cleared in high jump will be the dependent variable, while the three factors that determine this height will represent the dependent or predictor variables. In high jumping, the height an athlete clears (Bar Height) may be regarded as the sum of three separate heights (Hay, 1993):

****

H1: Takeoff Height – the height of the athlete’s center of gravity at the instant of takeoff.

H2: Flight Height – the max height that the athlete raises the center of gravity during the flight.

H3: Clearance Height – the difference between the maximum height reached by the center of gravity and the height of the crossbar.

For the purpose of this exercise H1 will be measured in feet, H2 in meters, and H3 in inches. Therefore, the following equation gives Bar Height in centimeters:

Bar Height (cm) = 30.48\*H1 + 100\*H2 + 2.54\*H3

The number in front of each height represents the conversion factor from each unit of measurement into centimeters. The conversion factors will be revealing when interpreting the output from the regression analysis.

**The objective of this analysis is to determine the relative importance of each sub height in determining Bar Height.**

The dependent and independent variables are each given their own column in SPSS. In this example there are four columns (Figure 1).

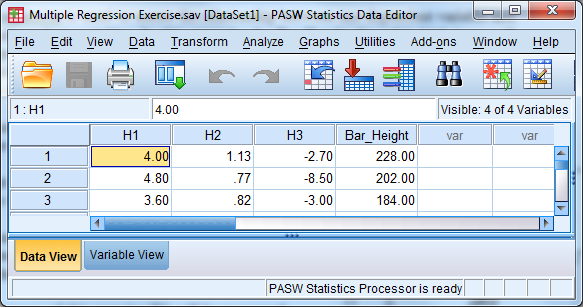


Figure Data view

How did I name the variables? There are 2 tabs at the bottom of the Data Editor, one labeled Data View, the other Variable View, as shown in Figure 2:

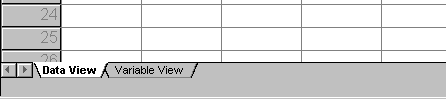


Figure Data view and variable view

You can toggle back and forth between the Data View (see Figure 1) and the Variable View, which is illustrated in Figure 3. In the Name column, you can type whatever labels you wish for your variables. If you don't type in labels, SPSS will use labels like **VAR001** and **VAR002** by default.

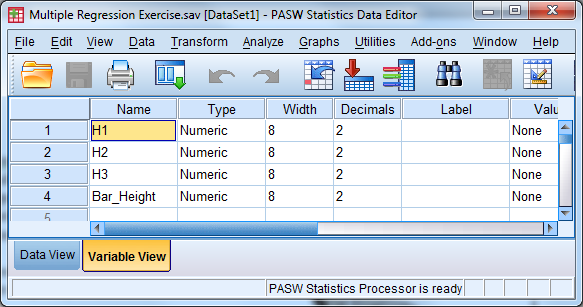


Figure Variable view

To perform the regression, you need to click on the Analyze menu, select Regression and then Linear, as in Figure 4.

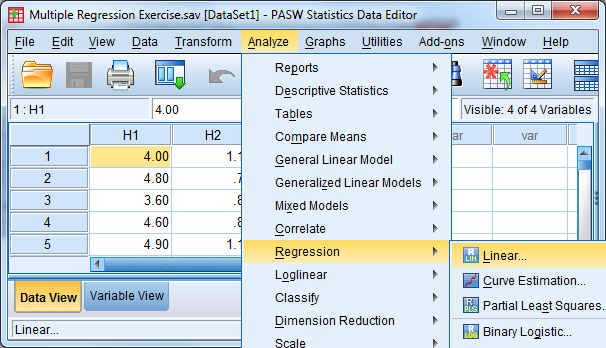


Figure Commands to perform multiple linear regression

After this, a dialog box will appear (Figure 5). In this box, you'll be able to identify the dependent and independent variables as shown. **Initially run the analysis with only two independent variables (H1 and H3).** Select *Enter* for type of “Method”. This means that both H1 and H3 will be included in the final model and analyzed simultaneously.

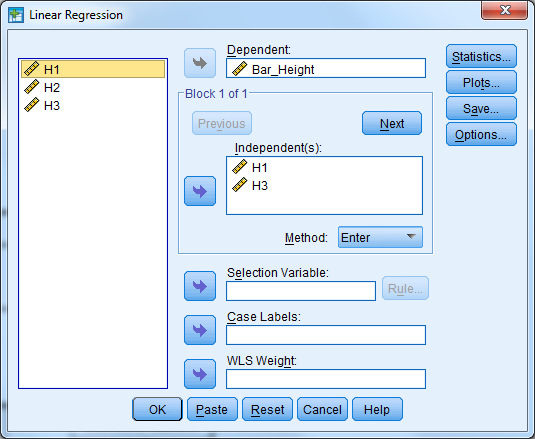


Figure 5 Linear regression dialogue box

Click on “Statistics” and select the options shown in Figure 6 and hit “Continue”

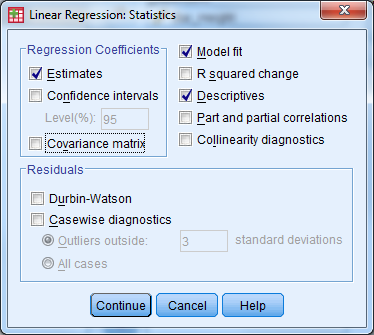


Figure Statistics options for linear regression

Click on “Plots” and select the options shown in Figure 7 and hit “Continue”



Figure Plot options for linear regression

Click on “Save” and check Mahalanobis. This produces a new column in your .sav file which allows you to check for outliers in your data. In linear regression an outlier is a data point whose actual value is very different from the predict value. It has a very large residual. Hit “Continue”.

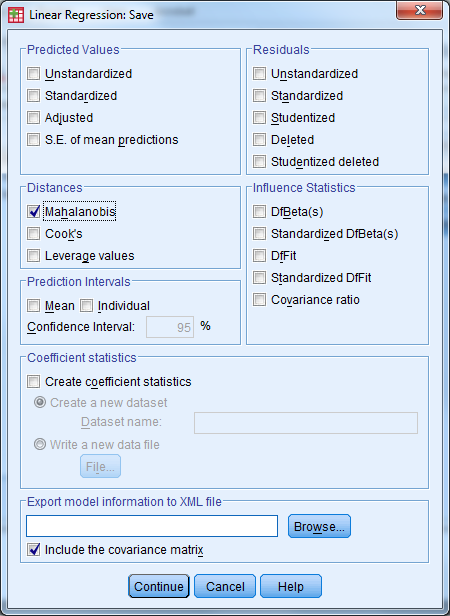


Figure Save options for linear regression

Back at the Linear Regression window (Figure 5), hit “OK” to perform the regression.

The following tables are the output generated from the above procedure. The first table of importance displays the Pearson correlations among the variables. Note that H1 and Bar Height are significantly correlated (r = .529, p < .001), which suggests that H1 was a good variable to include as a predictor. H3 is not significantly correlated with Bar Height (r = .132, p = .181), which suggests it may not be a good predictor. However, the fact that H1 is not strongly correlated with H3 (r = .011, p = .469) indicates there will not likely be any [multicollinearity](http://en.wikipedia.org/wiki/Multicollinearity) issues if they are both entered into the prediction model.

| **Correlations** | | | | |
| --- | --- | --- | --- | --- |
|  | | Bar\_Height | H1 | H3 |
| Pearson Correlation | Bar\_Height | 1.000 | .529 | .132 |
| H1 | .529 | 1.000 | .011 |
| H3 | .132 | .011 | 1.000 |
| Sig. (1-tailed) | Bar\_Height | . | .000 | .181 |
| H1 | .000 | . | .469 |
| H3 | .181 | .469 | . |
| N | Bar\_Height | 50 | 50 | 50 |
| H1 | 50 | 50 | 50 |
| H3 | 50 | 50 | 50 |

The following table indicates which variables were entered into the model and how they were entered.

| **Variables Entered/Removedb** | | | |
| --- | --- | --- | --- |
| Model | Variables Entered | Variables Removed | Method |
| 1 | H3, H1a | . | Enter | |
| a. All requested variables entered. | | | |
| b. Dependent Variable: Bar\_Height | | | |

The most important variable in the following table is Adjusted R Square, which indicates the percentage of variability in the dependent variable (Bar Height) that was predicted by the two independent variables (H1 and H3). In this case, our model accounts for 26.6 % of the variability in Bar Height.

| **Model Summaryb** | | | | | |
| --- | --- | --- | --- | --- | --- |
| Model | | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|  | 1 | .544a | .296 | .266 | 16.99311 |
| a. Predictors: (Constant), H3, H1 | | | | | |
| b. Dependent Variable: Bar\_Height | | | | | |

The following ANOVA table asses the overall significance of the model. Although the model only predicts 26.6% of the variance in bar height, it is significant (reliable).

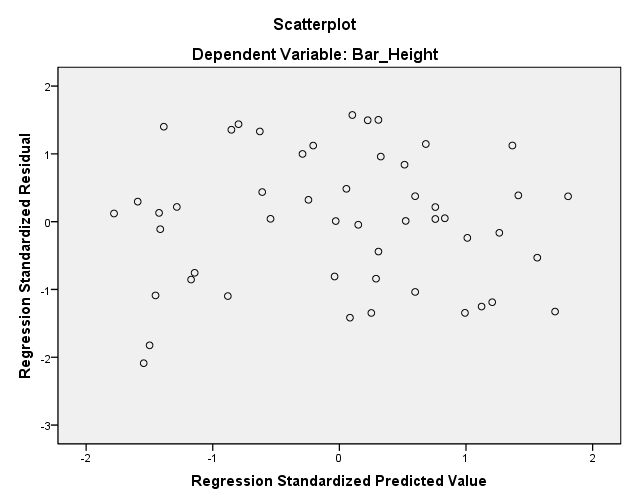
| **ANOVAb** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| Model | | Sum of Squares | df | Mean Square | F | Sig. |
| 1 | Regression | 5706.725 | 2 | 2853.362 | 9.881 | .00026a |
| Residual | 13571.995 | 47 | 288.766 |  |  |
| Total | 19278.720 | 49 |  |  |  |
| a. Predictors: (Constant), H3, H1 | | | | | | |
| b. Dependent Variable: Bar\_Height | | | | | | |

The Unstandardized Coefficients in the following table allow us to create a prediction equation: **Bar Height = 110.6 + 25.3\*H1 + 1.01\*H3**. The difference between the true Bar Height known from the .sav data set and the one estimated from this prediction equation is known as a residual. A residual graph will be shown later. At this point, the equation does not match the one shown on the first page because H2 has been left out of the model and SPSS has tweaked the conversion factors for H1 and H3 as well as added a constant (110.6) in order to minimize the size of the residuals for this particular set of data.

The following table also provides an indication of the relative importance of each predictor. This can be determined by examining the last three columns. The size of the Standardized Coefficients indicates its relative importance. In this analysis we can see that H1 (.528) is much more important than H3 (.126). This is confirmed by the final column which shows that H1 was significant, while H3 was not. H3 does not account for a significant amount of variability beyond what is accounted for by H1.

| **Coefficientsa** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
| B | Std. Error | Beta |
| 1 | (Constant) | 110.6 | 25.824 |  | 4.283 | 0.00009 |
| H1 | 25.3 | 5.857 | .528 | 4.313 | 0.00008 |
| H3 | 1.01 | .984 | .126 | 1.027 | .310 |
| a. Dependent Variable: Bar\_Height | | | | | | |

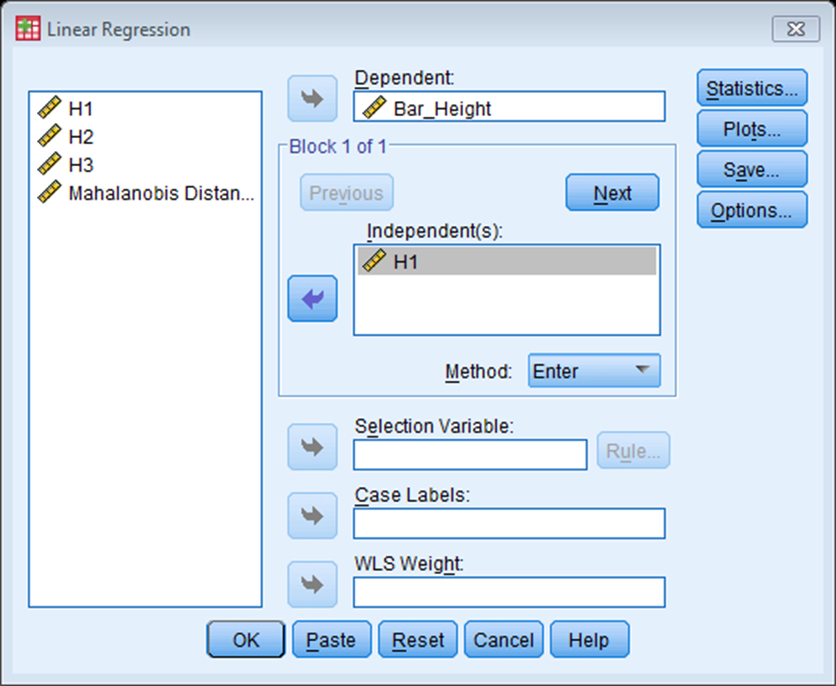
In the following plot, the predicted values have been converted to z-scores and plotted on the X axis, with their associated residual on the Y axis. This is a classic residual plot that suggests there are no issues with the data (no assumptions have been violated). There is no pattern in the scatterplot.



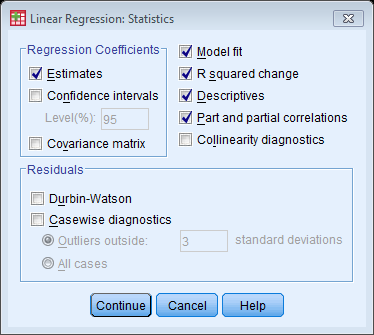
While the residual plot suggests that there are no outliers (very large residuals), this can be formally checked using the Mahalanobis distances, which should now be visible as an additional column in the .sav file. The larger the residual, the larger the Mahalanobis distance will be. A Chi-Square Distribution Table (p. 434 in our book) must be used to determine the cut-off value for outliers. The degrees of freedom is equal to the number of predictors (in this case 2), and the probability should be set at p < .001 (Tabachnick and Fidell, p.93). Using this two values (df = 2, p = .001) and referring to p.434 in the book, the critical value is 13.82, which is much larger than any value in our data. There are no outliers.

We will now re-run the analysis and modify some of the selections. The order in which variables are entered into a multiple regression can affect the interpretation of results. Ideally, a researcher will have formed hypotheses regarding the relationships between the predictors and criterion. These hypotheses should dictate how the variables are entered. In the first analysis, H1 and H3 were entered simultaneously. The analysis was not based on theory. In this analysis, we will assume that H1 is the most important variable, followed by H2, and lastly, by H3. As such we will use what is termed sequential or hierarchical regression.

To perform the regression, you need to click on the Analyze menu, select Regression and then Linear. Remove H3 from the Independent(s) list, so that only H1 is remaining and click “Next”. A new and blank Independent(s) list should appear; bring over H2 and click “Next”. Repeat for H3. You have just dictated the order in which the variables will be assessed.

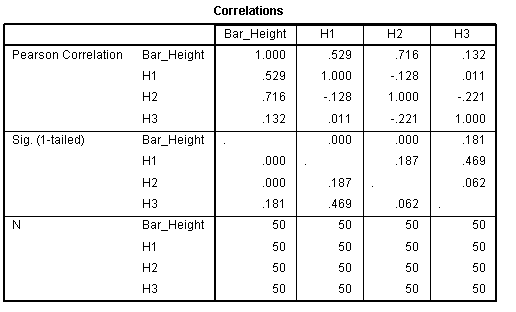


Click on “Statistics” and select the options shown below and hit “Continue”



Your other selections will remain the same as before. Back at the Linear Regression window, hit “OK” to perform the regression.

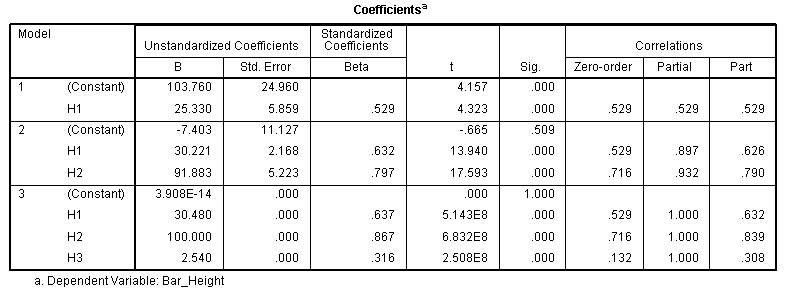
The Correlation table now contains H2, which has a strong correlation with Bar Height ( r = .716, p < .001).



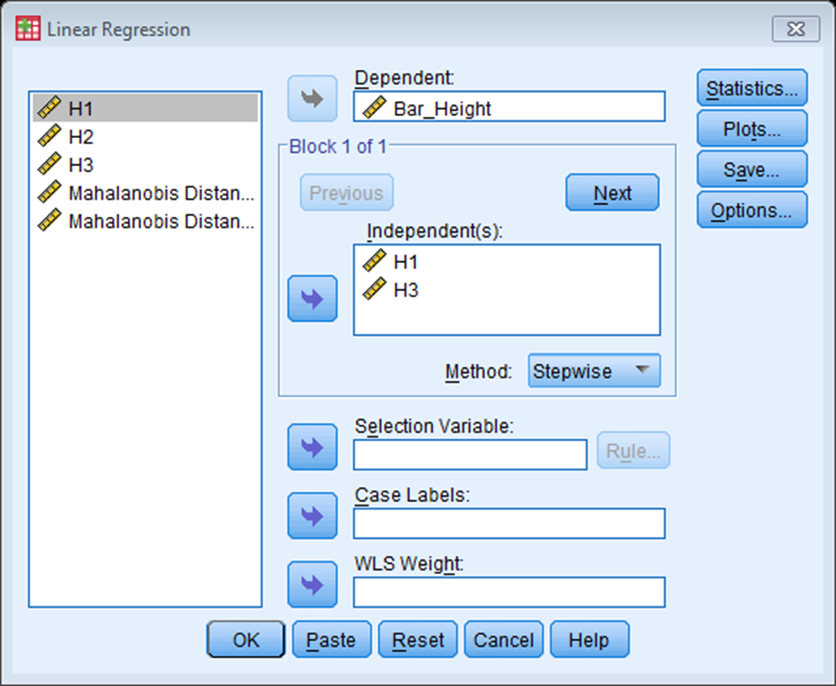
The Model Summary table now contains three separate models (dictated by how we entered the variables into the regression) and there is new information that should be considered. The Adjusted R Square column is still important and demonstrates how much variability in the dependent variable is predicted by the model. We can see that H1 alone predicts 26.5% which jumps to 90.1% when H2 is added to model. As expected, 100% of the variability is predicted, when H3 is added to the equation. Collectively, the change statistics reveal how much the model improved when each additional predictor was added. R Square Change shows how much additional variance was predicted by the inclusion of each variable, while the final column reveals whether this contribution is statistically significant.

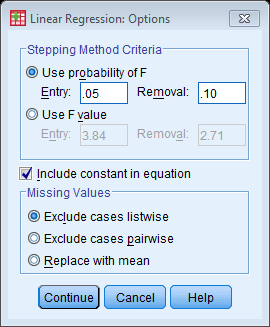


Focus on Model 3 in the Coefficients table below. You will notice, now that all 3 variables that determine Bar Height are included in the model, that the Unstandardized Coefficients would create an equation which is identical to the one shown on the first page of this document. Interestingly, in comparison to the previous analysis, you will notice, now that H2 has been included in the model, H3 is a significant predictor. Now examine the Correlation columns in the table below. The Zero-order correlations are the Pearson correlations. Perhaps the most important correlation is that labeled “Part”. When this value is squared, referred to as the squared semipartial correlation, it represents the unique contribution of that predictor to R2 in that set of predictors. To confirm this, square the semipartial correlation for H3 in Model 3 (.3082=.095) and you will notice that it is equal to the R Square Change value for Model 3 in the previous table.



We will now rerun the analysis for a third time. Let’s assume that we have no hypotheses regarding the relative importance of each variable and we are just looking for the best prediction equation possible. This is termed statistical or stepwise regression and the order of entry, as well as those variables retained in the final model, is based solely on statistical criteria determined through SPSS’s number crunching behind the scenes. We are going to repeat the first analysis including only H3 and H1, but this time we will choose Stepwise instead of the Enter method as shown in the image below.





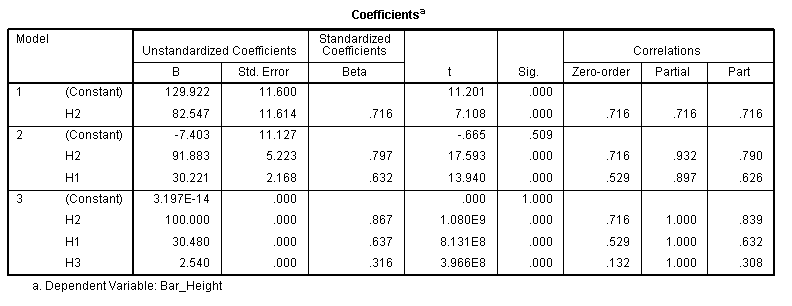
Select “Options” and the window shown on the right will appear. This window shows the criteria for adding and removing variables. A variable is entered into the model if the significance level of its F value is less than the Entry value and is removed if the significance level is greater than the Removal value. Entry must be less than Removal, and both values must be positive. To enter more variables into the model, increase the Entry value. To remove more variables from the model, lower the Removal value.

Hit “Continue”, and hit “OK” to run the regression.

You will notice that H3 is not included in the Model Summary, ANOVA, or Coefficients tables. It did not contribute significantly to the prediction beyond that of H1, so it was never entered. The reason is evident from the following table (p = .310).

| **Excluded Variablesb** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| Model | | Beta In | t | Sig. | Partial Correlation | Collinearity Statistics |
| Tolerance |
| 1 | H3 | .126a | 1.027 | .310 | .148 | 1.000 |
| a. Predictors in the Model: (Constant), H1 | | | | | | |
| b. Dependent Variable: Bar\_Height | | | | | | |

We will now rerun the analysis for a fourth and final time. This time include H1, H2, and H3 in the analysis and use Stepwise as the method. Compare the following output to the same table generated during the Enter method with all three variables. The most important thing to note is that H2 entered the model first as based on statistical criteria.



Assuming we performed a sequential regression with all three variables, the following would be satisfactory write-up in journal format.

A sequential multiple regression was performed between bar clearance height as the dependent variable and takeoff, flight, and clearance heights as the independent variables. Evaluation of the residual plot strongly suggested that the data met the assumptions of normality, linearity, and homoscedasticity. With the use of a p < .001 criterion for Mahalanobis distance no outliers were found among the 50 cases. All three of the independent variables contributed significantly to the prediction of bar height clearance. Flight height showed the largest amount of unique contribution (squared semipartial correlation = .70) followed by takeoff height (.40) and clearance height (.095). The overall fit of the model was perfect as collectively the three predictors account for all of the variability in bar clearance height.