Practice 3 SPSS

Partially based on Notes from the University of Reading: http://www.reading.ac.uk

Simple Linear Regression

A simple linear regression model is fitted when you want to investigate whether there is a linear relationship between two quantitative variables. It takes the form

$y = \alpha + \beta x + \varepsilon$

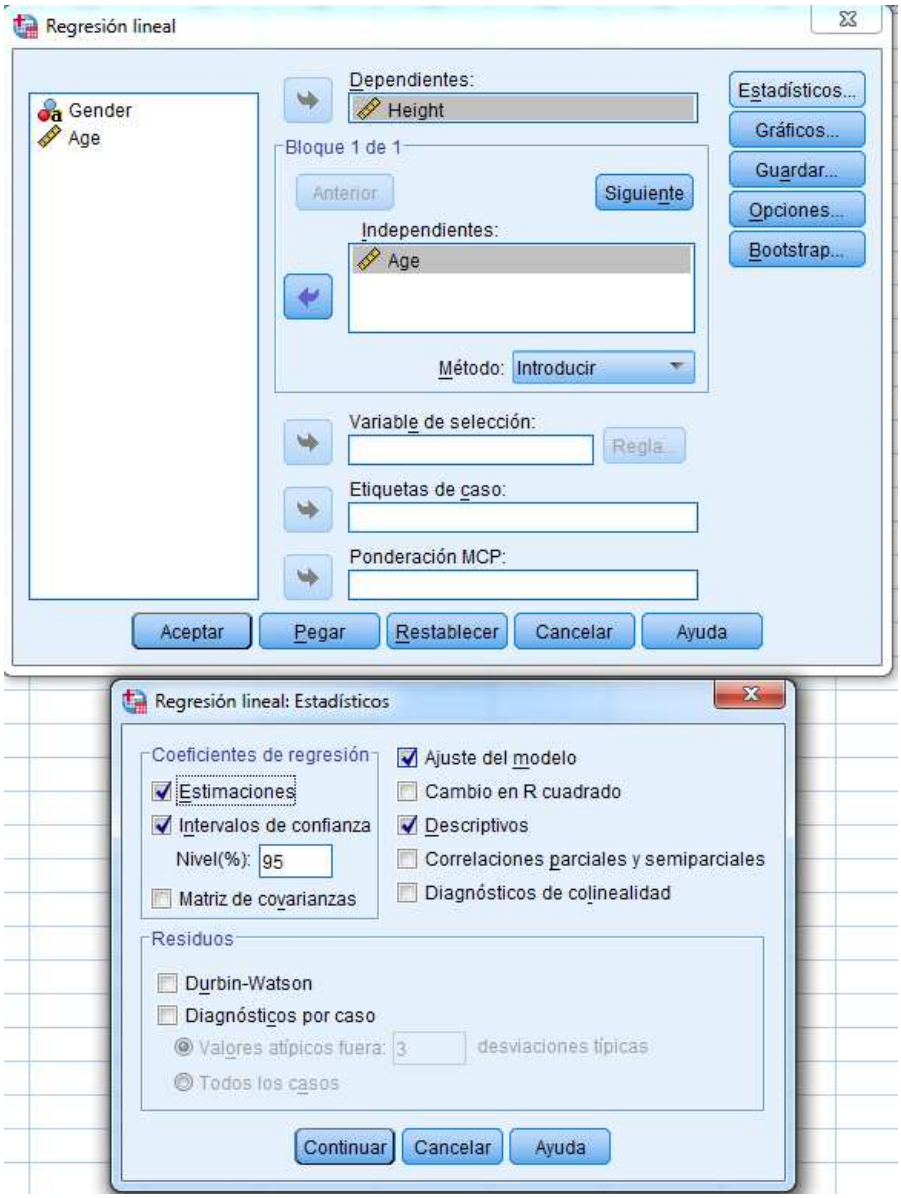
where y is the response (dependent) variable, in this next case height, x is the explanatory (independent) variable, which will be age, α is the intercept term of the model, β is the gradient of the linear model and ε is the error term.

To fit a Simple Linear Regression model to the data go to Analyze, followed by Regression, and then Linear. Drag drop *Height* in to the Dependent variable box and *Age* into the Independent variable box.

Analizar → Regresion → Lineales

It is obtained the *R-Square* value which is the amount of variation in the response that is explained by the model proposed. It is the square of Pearson's correlation coefficient. We would ideally like this percentage to be greater than 65%.

Coefficients of the fitted regression model along with their standard errors and p-values (Sig.) are also shown.



Note:

Assumptions of the model are Normality of the errors and a constant variance for all individuals.

Interpretation:

In this example the regression model that has been fitted is

Height = 100.43 + 0.35*Age

The t-statistics are testing whether $\alpha = 0$ or $\beta = 0$ and the corresponding p-values (*Sig.*) are given.

These p-values show us that both the intercept and gradient parameters are significantly different from zero as they are both significant. The result for β provides strong evidence of an association between height and age.

The Rsquare value is 0.459 which means that 45.9% of the variation in height is explained by age. The ANOVA table can be ignored as the hypothesis being tested for a simple linear regression model is again that β = 0, the results of which are already given in the T-test section.

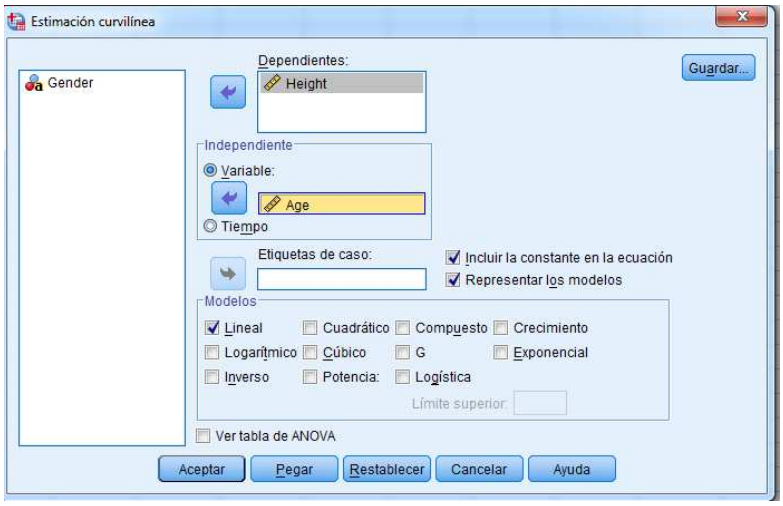
Creating and adding a regression line to a scatterplot

When creating a scatterplot you may want to fit a regression line to demonstrate graphically what your regression model is describing. In SPSS we need to create the scatterplot and add the fitted line all at once.

First click on Analyze, followed by Regression, and then Curve Estimation. Drag drop *Height* to the *Dependent variable* box and *Age* to the *Independent variable* box.

Under Models tick *Linear* is checked and leave the rest as the default settings.

Analizar → Regresion → Estimacion curvilinea



Interpretation:

From this graph it can be seen that there is a strong positive relationship between age and height. This means as age increases height will generally increase too. The graph explains what was seen earlier in the regression model output.

Multiple Linear Regression

Multiple regression involves fitting a model for a quantitative response (dependent) variable involving more than one explanatory variable, which is linear in its parameters. The data we will use for this example is delivery.txt. The file is about drinks delivery and contains three columns:

- Column 1 is the delivery time in minutes

- Column 2 is the number of cases being delivered

- Column 3 is the distance walked by the delivery man in feet

A multiple regression equation for this problem takes the form

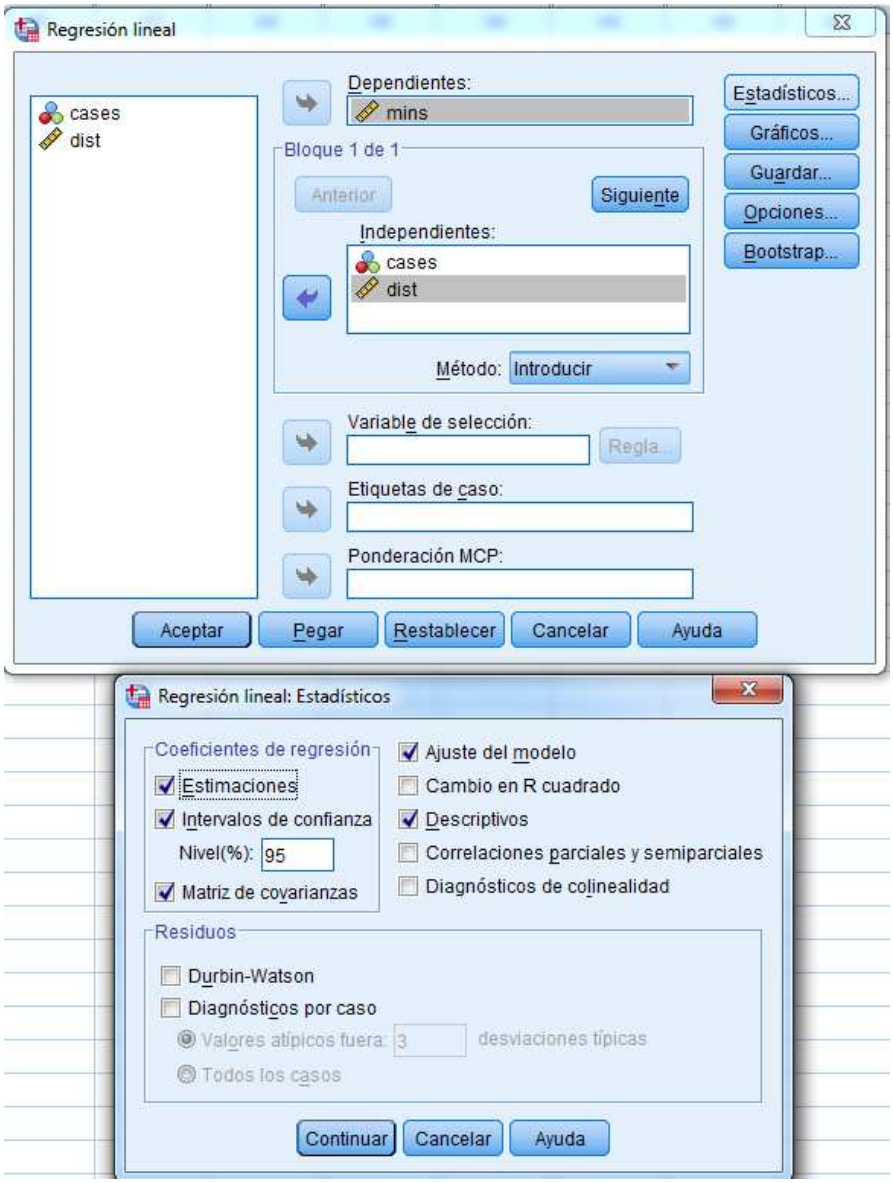
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

where the y variable is the response (delivery time), and the *x*'s are the explanatory variables (number of cases delivered and distance walked) and ε is the error term.

To fit the regression model select Analyze, then Regression, followed by Linear.

Analizar → Regresion → Lineales

It is shown the calculated coefficients for the equation (under the *unstandardized coefficients*, B column). Std Error is the standard error for the estimated coefficient. The T-statistics (t) are testing to see whether $\beta_i = 0$ and the corresponding p-values are given (*Sig.*).



It is shown the estimated standard deviation (Std Error of the Estimate) for the error in the model. The R Square value is the amount of variation in the response that is explained by the model. It should be as high as possible.

The ANOVA table is used to see if any of the variables are significant, unlike the T-tests which looks more specifically at the significance of individual variables.

Note:

Unless there is a good reason to keep a variable in the model when it is not significant then the variable should be removed and the model fitted again. This avoids unnecessary increased variability in predictions arising from the model.

Interpretation:

Looking at the p-values of the coefficients we can see that the constant term in the model is only just significant at the 5% level, but we still include it. The p-values for the other terms in the model are highly significant (being less than 0.001) and so we reject the null hypothesis that each parameter is equal to zero, given the other terms are in the model. Therefore each variable influences the response.

The Rsquare value for this model is 96%, meaning that 96% of the variation in delivery time is explained by the regression model with cases and distance.

From the ANOVA table we can see that the model is significant with $p<0.001$. This means that there is strong evidence to suggest that at least one of the parameters is non-zero, which we have already concluded from the earlier tests.

Model Checking

The main reason for model checking is to see if the underlying assumptions of a multiple regression hold. The main assumptions are as follows:

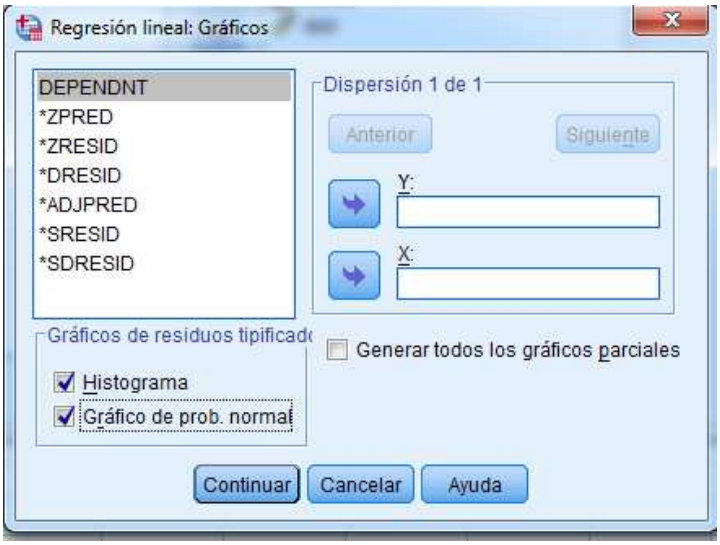
- The errors are Normally distributed (which implies that the response is as well, if all explanatory variables are quantitative).

- There is a linear relationship between the response and explanatory variable.

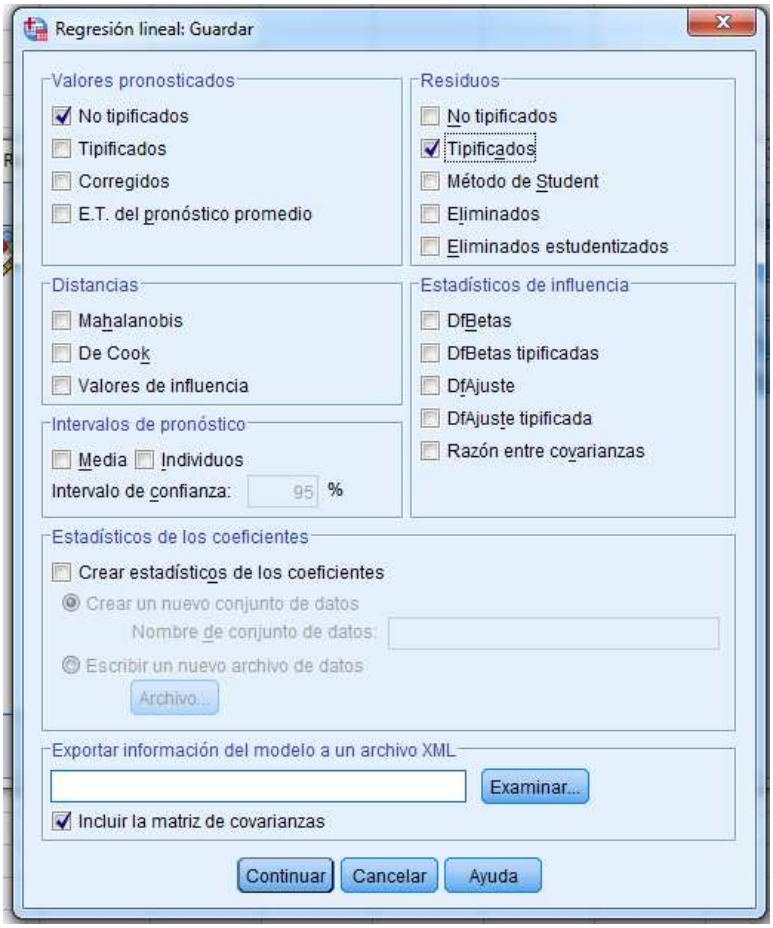
- The variance of the errors is equal for all observations.

When checking the model assumptions select Analyze, Regression, followed by Linear, The same window will appear as above. Click Plots.

Analizar → Regresion → Lineales (Graficos)



Clicking the Guardar (Save) button will produce the following window

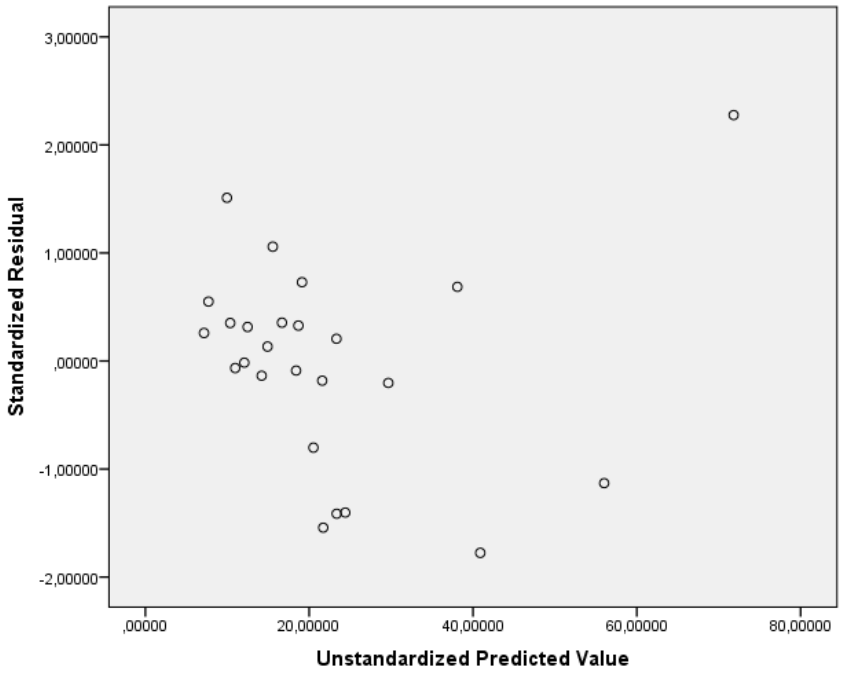


Note:

If the graph does not appear to be approximately Normally distributed then transformations, like *logarithms* or *inverses* will be needed.

Note:

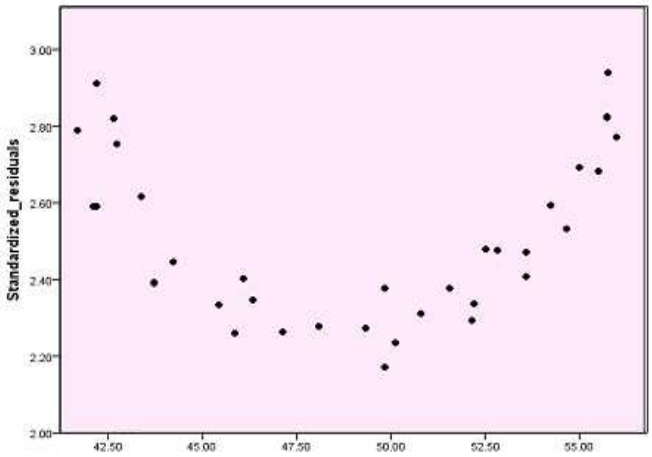
In the data view window you will now have two new columns, for residuals and predicted values. You will now need to create a scatterplot with your X variable as the unstandardized predicted value, and the Y variable as the standardized residual.



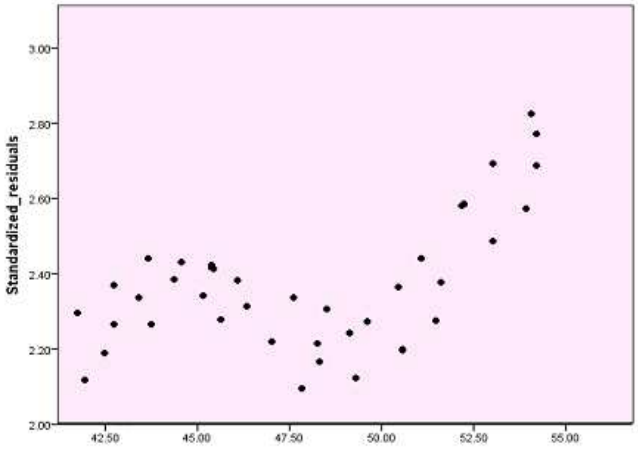
Note:

residual plots are a function of the difference between observed responses and those predicted by the model. If the standardised residual plots are not showing a random scatter of data, as being represented above, then patterns within them will indicate problems with the assumptions. Mild deviations from the ideal pattern are not too concerning. However, major deviations will suggest that the model is unreliable.

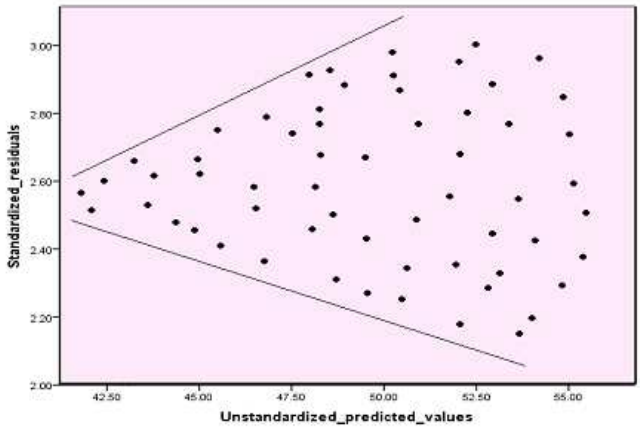
Examples of problems with residual plots



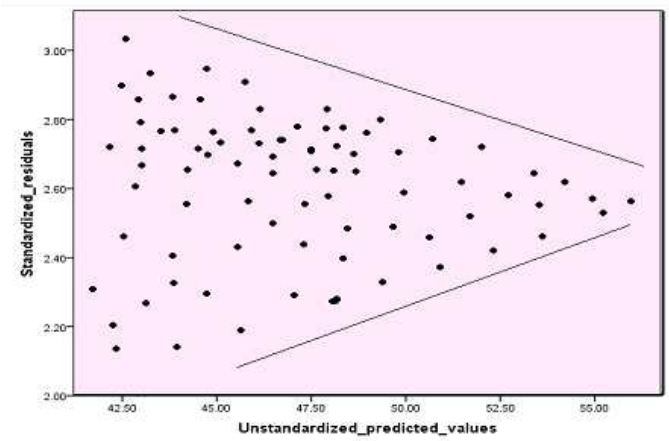
For residuals versus explanatory variable graph: If the following pattern is being displayed then a square term needs to be added to the model, i.e. x^2



For residuals versus explanatory variable graph: If the following pattern is being displayed then a cubic term needs to be added to the model, i.e. x^3



For residuals versus fit graph: If a funnel shape is shown like this, then the variances are not constant, and are in fact increasing with the fitted value.



For residuals versus fit graph: If a funnel shape is shown like this, then the variances are not constant, and are in fact decreasing with the fitted value.

Stepwise Selection

Stepwise selection is a method for building a model which contains only those variables which are *significant* (at a chosen level) in modelling the response (dependent) variable. It is particularly useful when there are many possible explanatory (independent) variables. Some of these variables may be highly correlated with each other and therefore will explain the same variation in the response and not be independently predictive. Some may also not influence the response in any meaningful way. It is advisable to construct a model with as few explanatory variables as possible to ensure an easy-to-interpret model which will be efficient for future prediction purposes.

Stepwise selection aims to construct a good model satisfying these aims by the dropping and adding of variables in a model according to their significance in the presence of the other variables.

For stepwise selection data would usually be in the form of one quantitative response variable, plus several quantitative explanatory variables. It is assumed normality for the response variable and it is build a multiple linear regression model.

Note: there is no guarantee that the final model produced is sensible. Standard model checking procedures should be applied to check that the assumptions of the final model are reasonable.

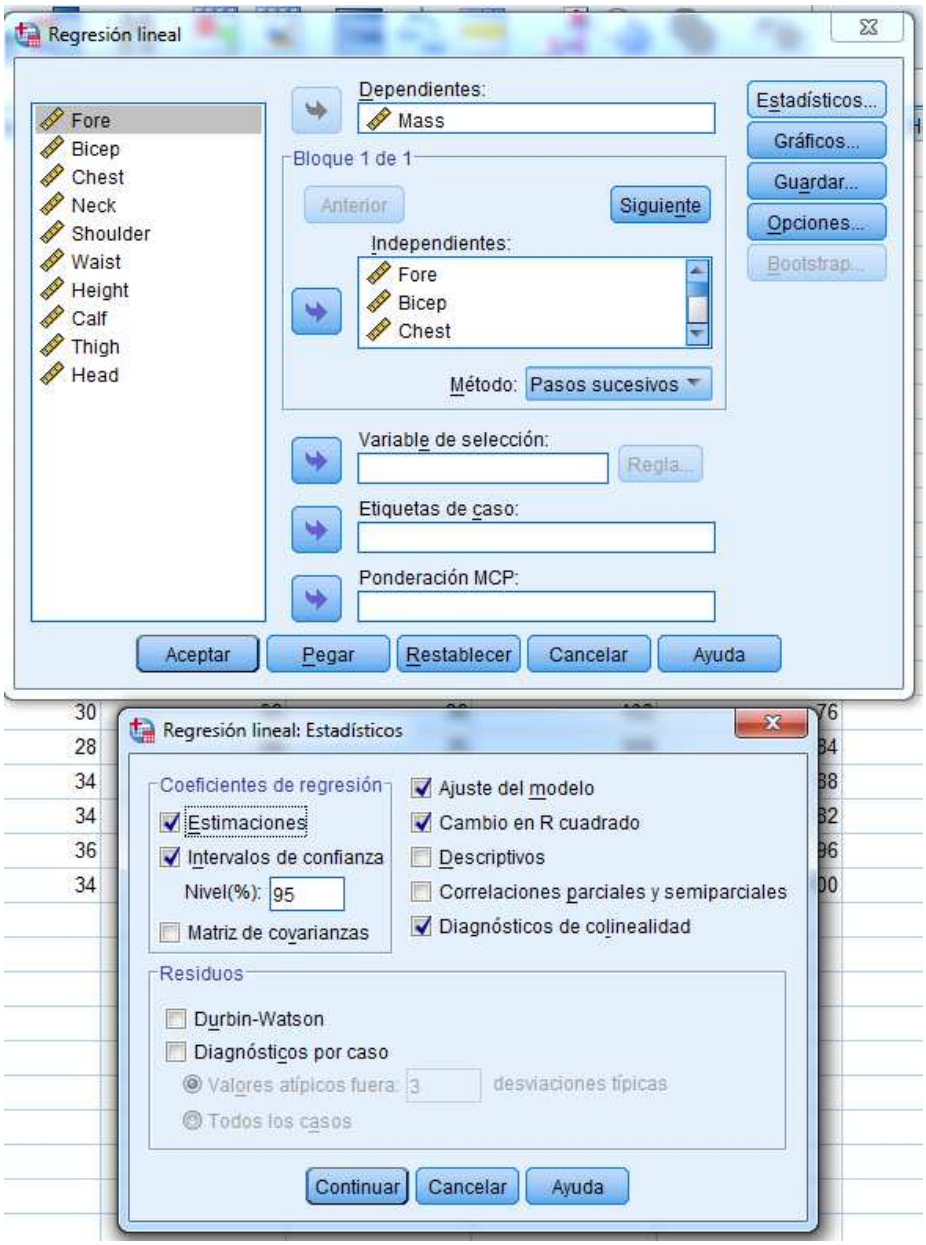
For this example the data that will be used is physical.xlsx. This dataset consists of the mass (weight in kg) plus various physical measurements for 22 healthy young males.

To perform stepwise selection select Analyze, followed by Regression then select Linear.

Analizar → Regresion → Lineales

Add the response variable to the Dependent box and add the explanatory variables into the Independent(s) box.

Select *Stepwise* (Pasos sucesivos) from the *Method* (Método) drop down menu.



Interpretation:

The variables entered/removed box shows the number of steps carried out in the analysis and which variables are added and removed at each stage.

The model summary section gives the summary statistics of each model in the analysis.

Note: of most interest is the fourth model as it is the final one.

R-Square is the amount of variation in the response that is explained by the model; *Adjusted R-Square* is the adjusted value that takes into account the number of variable in the model.

The ANOVA table is the final row of results which are relevant to the final model.

In the Coefficients section we can again see that there were only four steps carried out in this analysis to reach the final model, and we can see there are only four variable added from the original ten (all of the four variables are significant).

The estimates of the parameters for each of these variables are given in the *Unstandardized coefficients (B) column.*

The final model has an R-Square value of 96.6%, so the majority of variation in the response (mass) is explained by this model.

The final fitted model using this selection process includes variables:

Waist, Fore, Height and Thigh.