

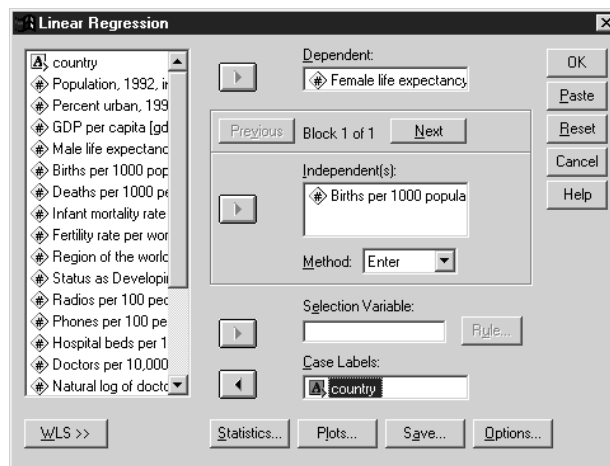
Chapter 4 – Regression Analysis SPSS

Linear regression analysis estimates the coefficients of a linear equation, involving one or more independent variables, that best predict the value of the dependent variable.

Data. The dependent and independent variables should be quantitative. Categorical variables, such as religion, major field of study, or region of residence, need to be recoded to binary (dummy) variables or other types of contrast variables.

Assumptions. For each value of the independent variable, the distribution of the dependent variable must be normal. The variance of the distribution of the dependent variable should be constant for all values of the independent variable. The relationship between the dependent variable and each independent variable should be linear, and all observations should be independent.

Procedure. To open the Linear Regression dialog box, from the menus choose:
Analyse → Regression → Linear.



Select more than one variable for the Independent(s) list, if you want to obtain a multiple linear regression. You can specify more than one list, or “block” of variables, using the *Next* and *Previous* buttons to display the different lists. Up to nine blocks can be specified.

‘Method:’ allows you to select a method for including / excluding independent variables in a model.

- **Enter.** All variables in the block are entered into the equation as a group.
- **Stepwise.** Selection of variables within the block proceeds by steps. At each step, variables already in the equation are evaluated according to the selection criteria for removal; then variables not in the equation are evaluated for entry. This process repeats until no variable in the block is eligible for entry or removal.
- **Remove.** All variables in the block that are already in the equation are removed as a group
- **Backward.** All variables in the block that are already in the equation are evaluated according to the selection criteria for removal. Those eligible are removed one at a time until no more are eligible.
- **Forward.** All variables in the block that are not in the equation are evaluated according to the selection criteria for entry. Those eligible are entered one at the time until no more are eligible.

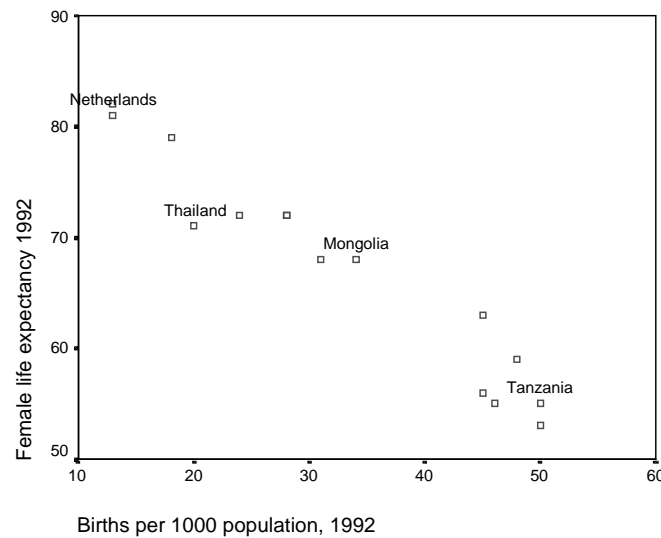
Click on ‘**Statistics...**’ button to request optional statistical output including regression coefficients, descriptives, model fit statistics, etc.

Click on **‘Plots...’** to request optional plots, including scatterplots, histograms, normal probability plots and outlier plots.

The **‘Save...’** button allows you to save predicted values, residuals, and related measures as new variables, which are added to the working data file. A table in the output shows the name of each new variable and its content.

The **‘Options...’** button allows you to control the criteria by which variables are chosen for entry or removal from the regression model; to suppress the constant term; and to control the handling of missing values.

Example 1. Can we predict life expectancy for females from a given birth rate of a country?



Looking at a scatterplot, we can see that as the birth rate increases, life expectancy decreases and vice versa. The relationship between the 2 variables appear to be linear in nature, the individual points (countries) tend to group around a straight line. This line is defined by a y-intercept and a slope, which are presented under the column B-coefficients in the following table (produced as an output of a linear regression analysis):

Coefficients						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	89.985	1.765		50.995	.000
	Births per 1000 population, 1992	-.697	.050	-.968	-13.988	.000

a Dependent Variable: Female life expectancy 1992

Now you can write the regression line as follows:

$$\text{Female life expectancy} = 89,985 - 0.697 * \text{Birth rate}$$

The beta coefficient tells you how strongly is the independent variable associated with the dependent variable. It is equal to the correlation coefficient between the 2 variables.

From the following table you can find how well the model fits the data.

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.968	.938	.933	2.54

a Predictors: (Constant), Births per 1000 population, 1992

b Dependent Variable: Female life expectancy 1992

This table displays R, R squared, adjusted R squared, and the standard error. R is the correlation between the observed and predicted values of the dependent variable. The values of R range from -1 to 1. The sign of R indicates the direction of the relationship (positive or negative). The absolute value of R indicates the strength, with larger absolute values indicating stronger relationships.

R squared is the proportion of variation in the dependent variable explained by the regression model. The values of R squared range from 0 to 1. Small values indicate that the model does not fit the data well. The sample R squared tends to optimistically estimate how well the models fits the population. Adjusted R squared attempts to correct R squared to more closely reflect the goodness of fit of the model in the population. Use R Squared to help you determine which model is best. Choose a model with a high value of R squared that does not contain too many variables. Models with too many variables are often over fit and hard to interpret.

Example 2. We will try to predict females' life expectancy from the following independent variables:

- Urban – percentage of the population living in urban areas
- Docs – number of doctors per 10,000 people
- Beds – number of hospital beds per 10,000 people
- GDP – Per capita gross domestic product in dollars
- Radios – Radios per 100 people

Besides R-squared (as presented above in Example 1), we can use ANOVA (Analysis of variance) to check how well the model fits the data. Look at the following table:

ANOVA						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	11834.950	5	2366.990	104.840	.000
	Residual	2483.490	110	22.577		
	Total	14318.440	115			

a Predictors: (Constant), GDP, RADIO, BEDS, URBAN, DOCS

b Dependent Variable: LIFEEXPF

The F statistic is the regression mean square (MSR) divided by the residual mean square (MSE). If the significance value of the F statistic is small (smaller than say 0.05) then the independent variables do a good job explaining the variation in the dependent variable. If the significance value of F is larger than say 0.05 then the independent variables do not explain the variation in the dependent variable, and the null hypothesis that all the population values for the regression coefficients are 0 is accepted.

After checking for the model fit, we might want to know the relative importance of each IV in predicting DV. The unstandardized (B) coefficients are the coefficients of the estimated regression model. In our example ...

Coefficients					
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	40.779	3.201		12.739	.000
URBAN	-0.007	.033	-.015	-.201	.841
BEDS	1.174	.750	.097	1.565	.120
RADIO	1.541	.703	.130	2.191	.031
DOCS	3.965	.568	.555	6.982	.000
GDP	1.626	.619	.225	2.629	.010

The regression equation can be written as follows:

$$\text{LIFEEXPF} = 40.8 - 0.007 \cdot \text{URBAN} + 1.2 \cdot \text{BEDS} + 1.5 \cdot \text{RADIO} + 3.97 \cdot \text{DOCS} + 1.6 \cdot \text{GDP}$$

Often the independent variables are measures in different units. The standardized coefficients or *betas* are an attempt to make the regression coefficients more comparable. If you transformed the data to z scores prior to your regression analysis, you would get the beta coefficients as your unstandardized coefficients. The t statistics can help you determine the relative importance of each variable in the model. As a guide regarding useful predictors, look for t values well below -2 or above +2.