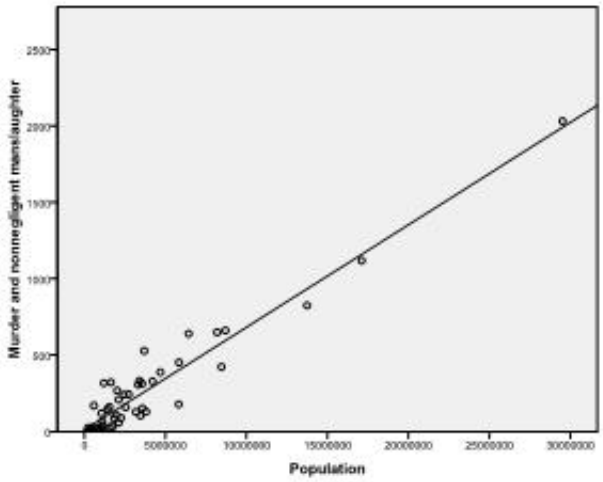
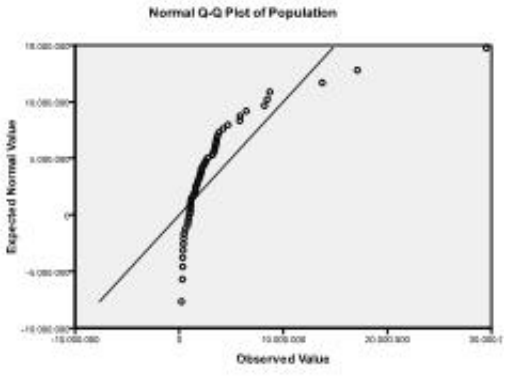
The Linear Regression Analysis in SPSS

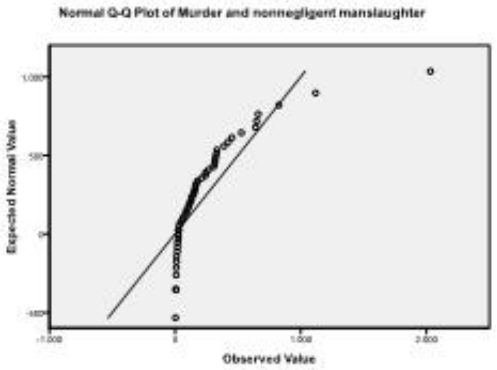
This example is based on the FBI's 2006 crime statistics. Particularly we are interested in the relationship between size of the state and the number of murders in the city.

First we need to check whether there is a linear relationship in the data. For that we check the scatterplot. The scatter plot indicates a good linear relationship, which allows us to conduct a linear regression analysis. We can also check the Pearson's Bivariate Correlation and find that both variables are highly correlated (r = .959 with p

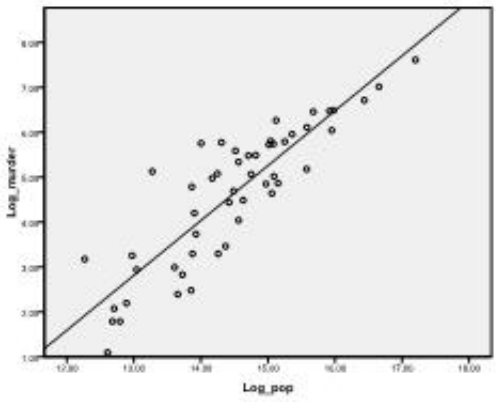


Secondly we need to check for multivariate normality. In our example we find that multivariate normality might not be present.

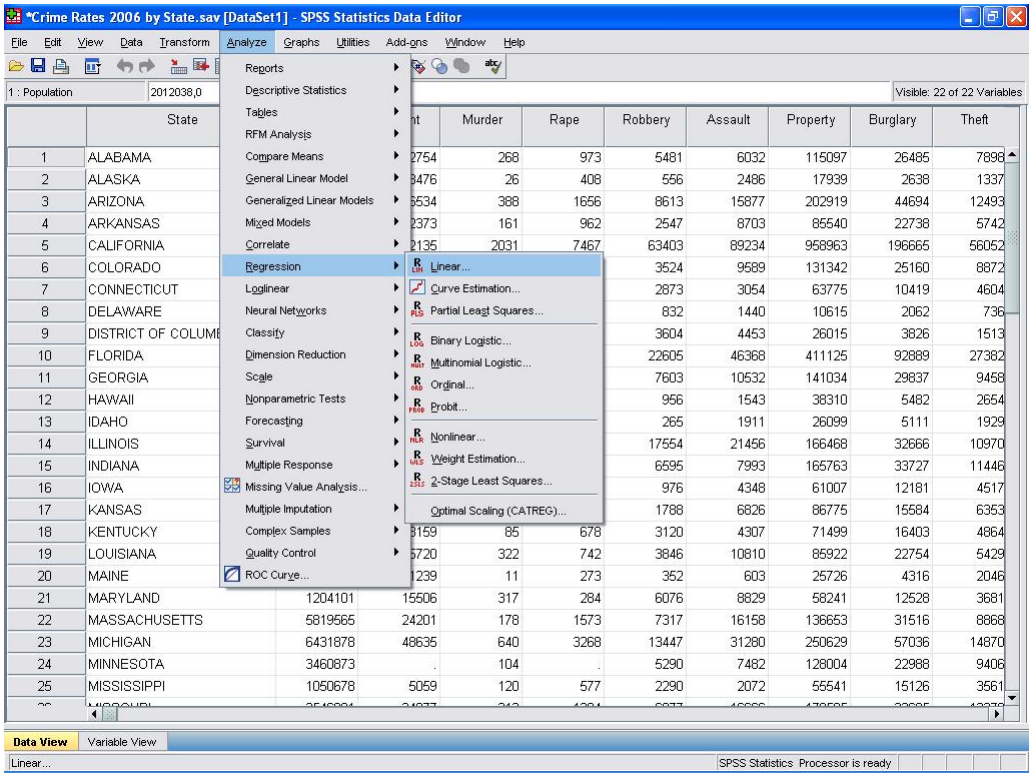




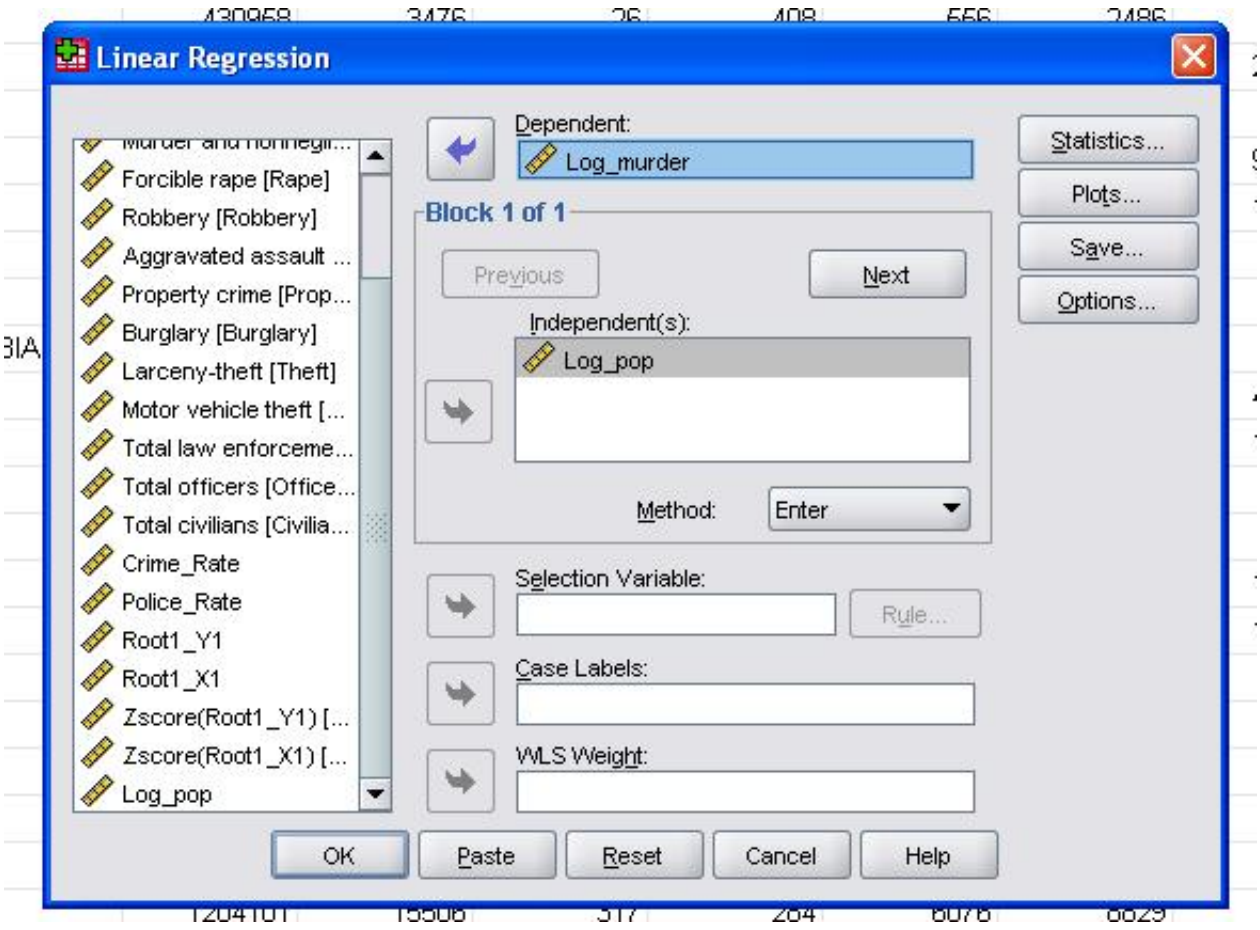
The Kolmogorov-Smirnov test confirms this suspicion (p = 0.002 and p = 0.006). Conducting a ln-transformation on the two variables fixes the problem and establishes multivariate normality (K-S test p = .991 and p = .543).



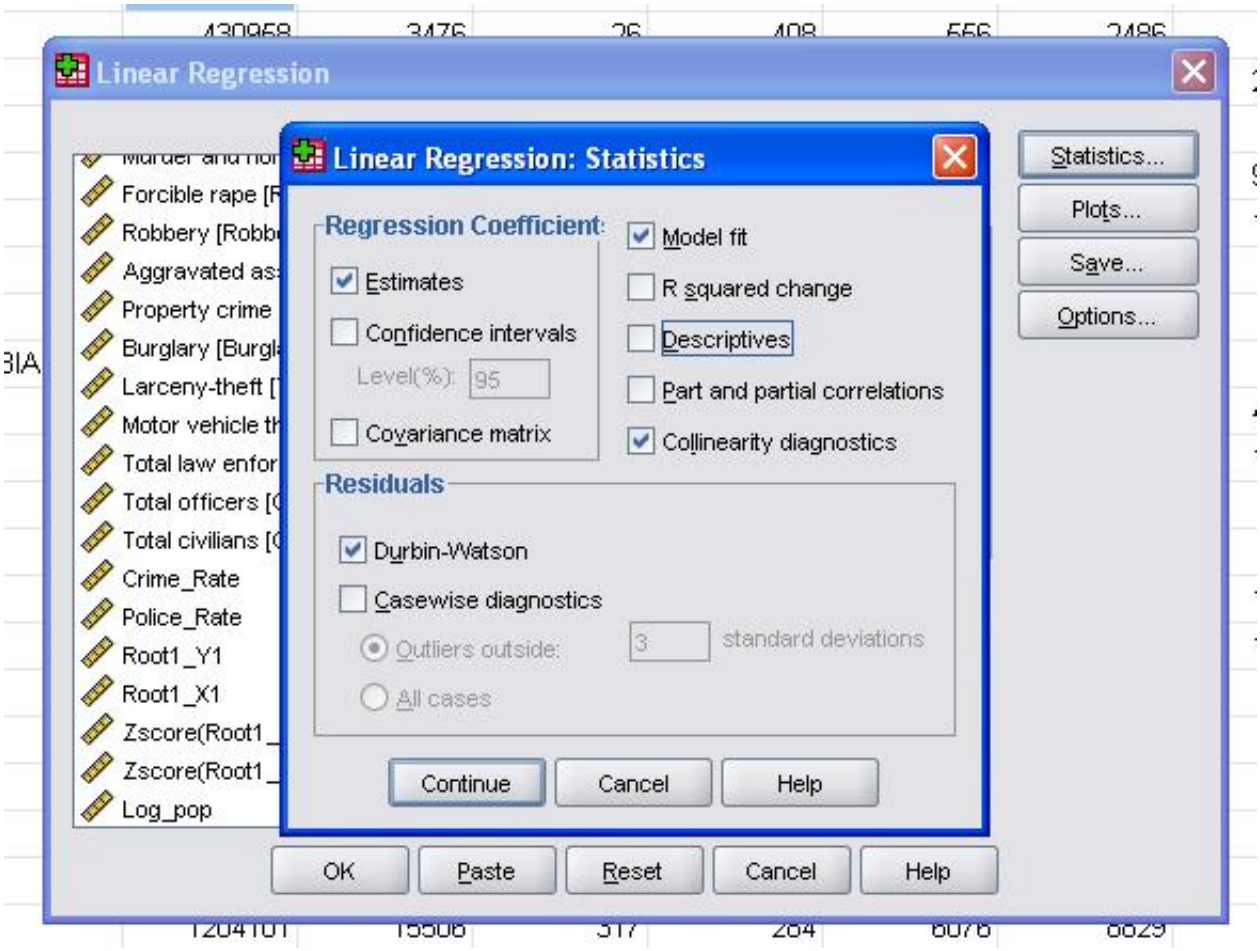
We now can conduct the linear regression analysis. Linear regression is found in SPSS in Analyze/Regression/Linear…



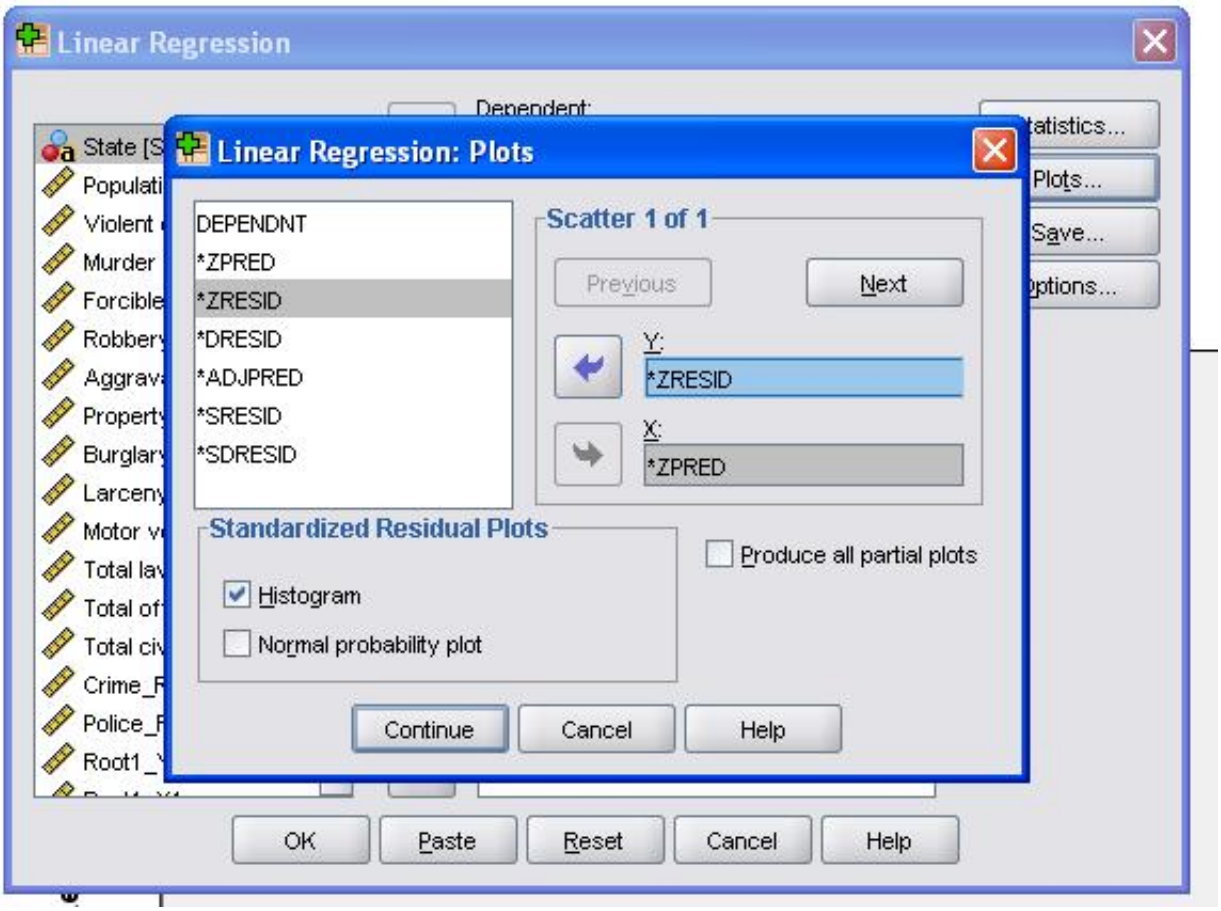
In this simple case we need to just add the variables log_pop and log_murder to the model as dependent and independent variables.



The field statistics allows us to include additional statistics that we need to assess the validity of our linear regression analysis.



It is advisable to additionally include the collinearity diagnostics and the Durbin-Watson test for auto-correlation. To test the assumption of homoscedasticity of residuals we also include a special plot in the Plots menu.



The SPSS Syntax for the linear regression analysis is

```
REGRESSION
/MISSING LISTWISE
/STATISTICS COEFF OUTS R ANOVA COLLIN TOL
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT Log_murder
/METHOD=ENTER Log_pop
/SCATTERPLOT=(*ZRESID ,*ZPRED)
/RESIDUALS DURBIN HIST(ZRESID).
```

The output's first table shows the model summary and overall fit statistics. We find that the adjusted R^2 of our model is 0.756 with the R^2 = .761 that means that the linear regression explains 76.1% of the variance in the data. The Durbin-Watson d = 2.323, which is between the two critical values of 1.5

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,872[a]	,761	,756	,76322	2,323

a. Predictors: (Constant), Log_pop

b. Dependent Variable: Log_murder

The next table is the F-test, the linear regression's F-test has the null hypothesis that there is no linear relationship between the two variables (in other words $R^2=0$). With F = 156.2 and 50 degrees of freedom the test is highly significant, thus we can assume that there is a linear relationship between the variables in our model.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	90,988	1	90,988	156,202	,000[a]
	Residual	28,543	49	,583		
	Total	119,531	50			

a. Predictors: (Constant), Log_pop

b. Dependent Variable: Log_murder

The next table shows the regression coefficients, the intercept and the significance of all coefficients and the intercept in the model. We find that our linear regression analysis estimates the linear regression function to be y = -13.067 + 1.222
* x. Please note that this does not translate in there is 1.2 additional murders for every 1000 additional inhabitants because we ln transformed the variables.

If we re-ran the linear regression analysis with the original variables we would end up with y = 11.85 + 6.7*10-5 which shows that for every 10,000 additional inhabitants we would expect to see 6.7 additional murders.

In our linear regression analysis the test tests the null hypothesis that the coefficient is 0. The t-test finds that both intercept and variable are highly significant (p

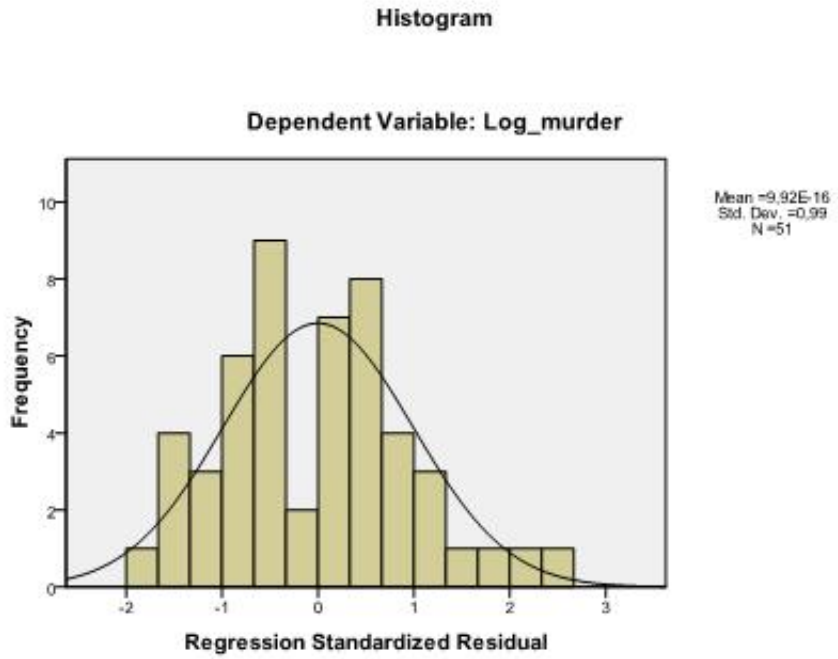
Coefficients[a]

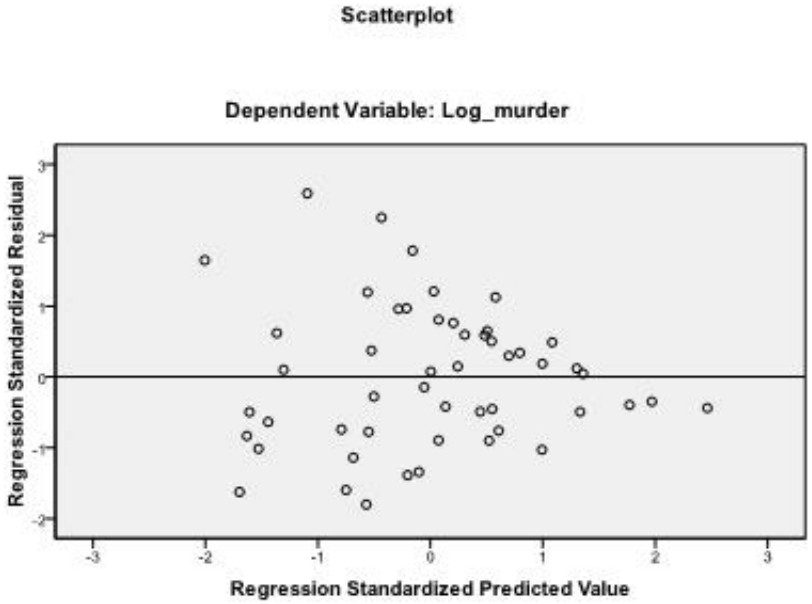
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-13,067	1,420		-9,203	,000		
	Log_pop	1,222	,098	,872	12,498	,000	1,000	1,000

a. Dependent Variable: Log_murder

This table also includes the Beta weights (which express the relative importance of independent variables) and the collinearity statistics. However, since we have only 1 independent variable in our analysis we do not pay attention to those values.

The last thing we need to check is the homoscedasticity and normality of residuals. The histogram indicates that the residuals approximate a normal distribution. The Q-Q-Plot of z*pred and z*presid shows us that in our linear regression analysis there is no tendency in the error terms.




Scatterplot
Dependent Variable: Log_murder
Regression Standardized Residual
Regression Standardized Predicted Value
3
2
1
0
-1
-2
-3