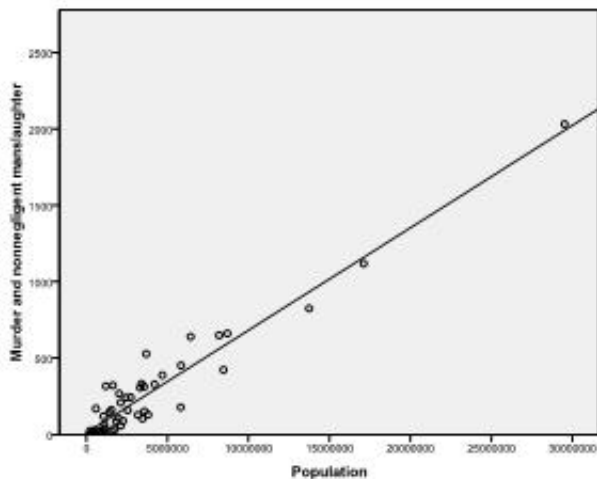


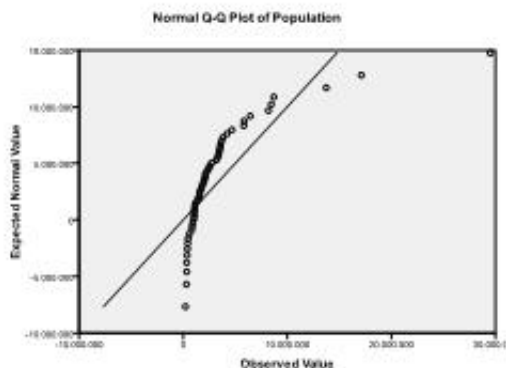
The Linear Regression Analysis in SPSS

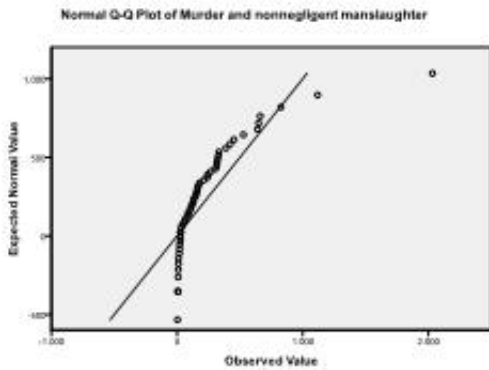
This example is based on the FBI's 2006 crime statistics. Particularly we are interested in the relationship between size of the state and the number of murders in the city.

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

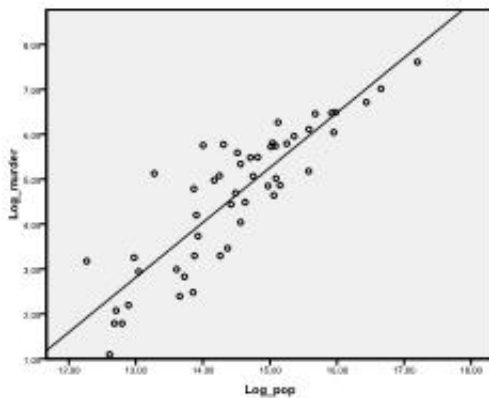


Secondly we need to check for multivariate normality. In our example we find that multivariate normality might not be present.





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 $p = .543$).

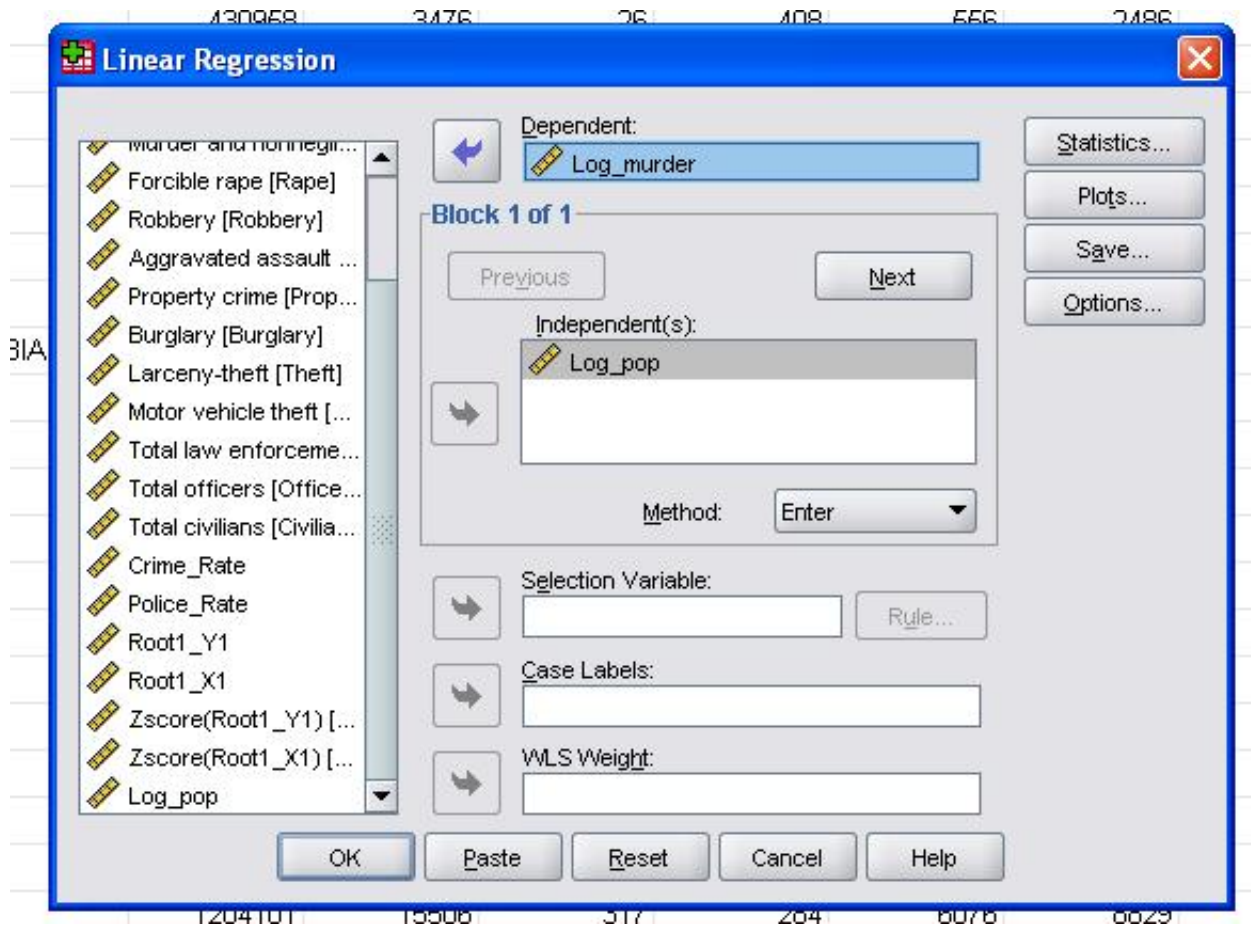


We now can conduct the linear regression analysis. Linear regression is found in SPSS in Analyze/Regression/Linear...

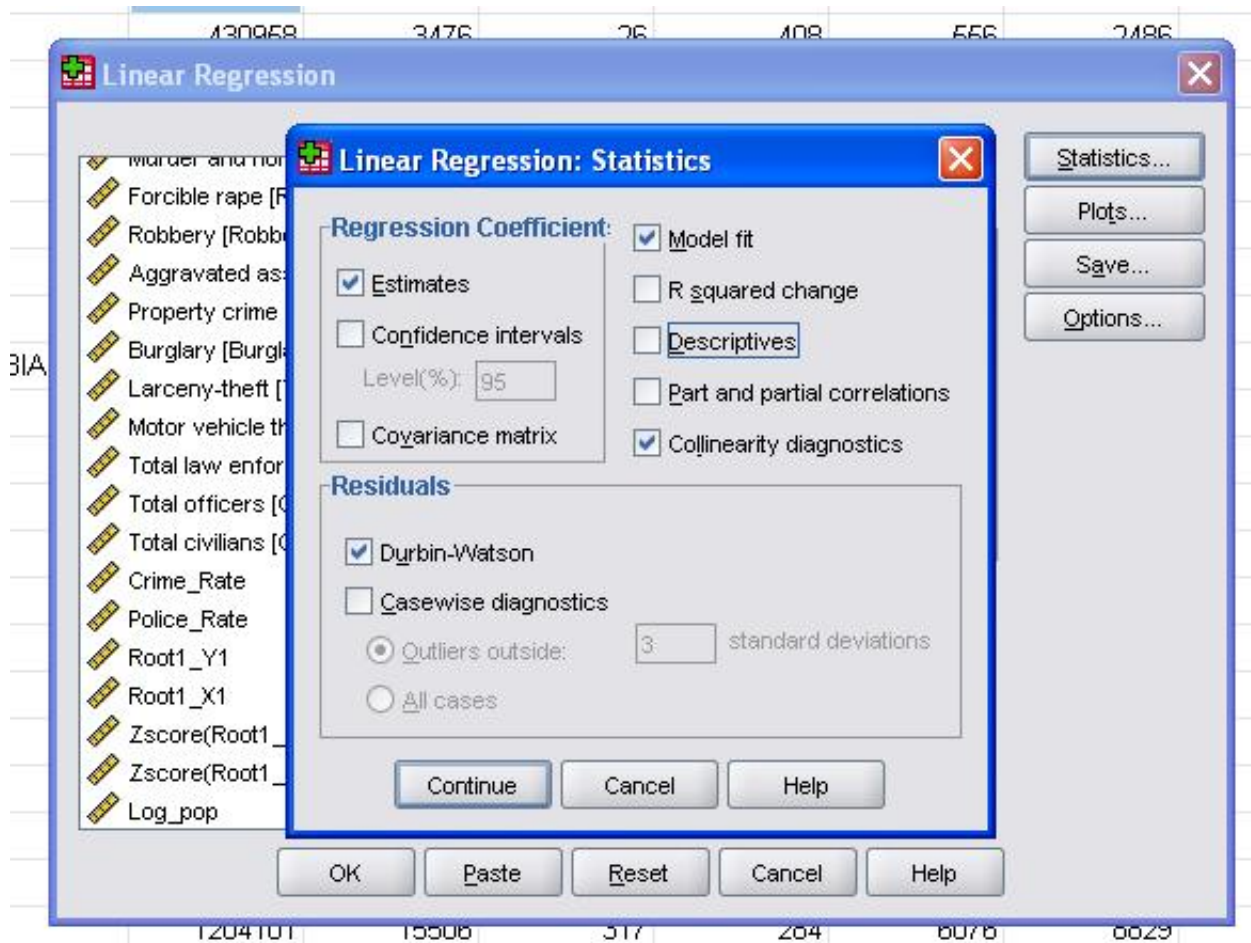
The screenshot shows the SPSS Statistics Data Editor window with the following data table visible:

State	Population	Murder	Rape	Robbery	Assault	Property	Burglary	Theft	
1 ALABAMA	2012038,0	2754	268	973	5481	6032	115097	26485	7898
2 ALASKA		3476	26	408	556	2486	17939	2638	1337
3 ARIZONA		6534	368	1656	8613	15877	202919	44694	12493
4 ARKANSAS		2373	161	962	2547	8703	85540	22738	5742
5 CALIFORNIA		2135	2031	7467	63403	89234	958963	196665	56052
6 COLORADO					3524	9589	131342	25160	8872
7 CONNECTICUT					2873	3054	63775	10419	4604
8 DELAWARE					832	1440	10615	2062	736
9 DISTRICT OF COLUMBIA					3604	4453	26015	3826	1513
10 FLORIDA					22605	46368	411125	92889	27382
11 GEORGIA					7603	10532	141034	29837	9458
12 HAWAII					956	1543	38310	5482	2654
13 IDAHO					265	1911	26099	5111	1929
14 ILLINOIS					17554	21456	166468	32666	10970
15 INDIANA					6595	7993	165763	33727	11446
16 IOWA					976	4348	61007	12181	4517
17 KANSAS					1788	6826	86775	15584	6353
18 KENTUCKY		8159	85	678	3120	4307	71499	16403	4864
19 LOUISIANA		5720	322	742	3846	10810	85922	22754	5429
20 MAINE		1239	11	273	352	603	25726	4316	2046
21 MARYLAND	1204101	15506	317	284	6076	8829	58241	12528	3681
22 MASSACHUSETTS	5819565	24201	178	1573	7317	16158	136653	31516	8868
23 MICHIGAN	6431878	48635	640	3268	13447	31280	250629	57036	14870
24 MINNESOTA	3460873		104		5290	7482	128004	22988	9406
25 MISSISSIPPI	1050678	5059	120	577	2290	2072	55541	15126	3561
26 MISSOURI	3545884	34973	343	1884	6937	15585	178585	38885	12372

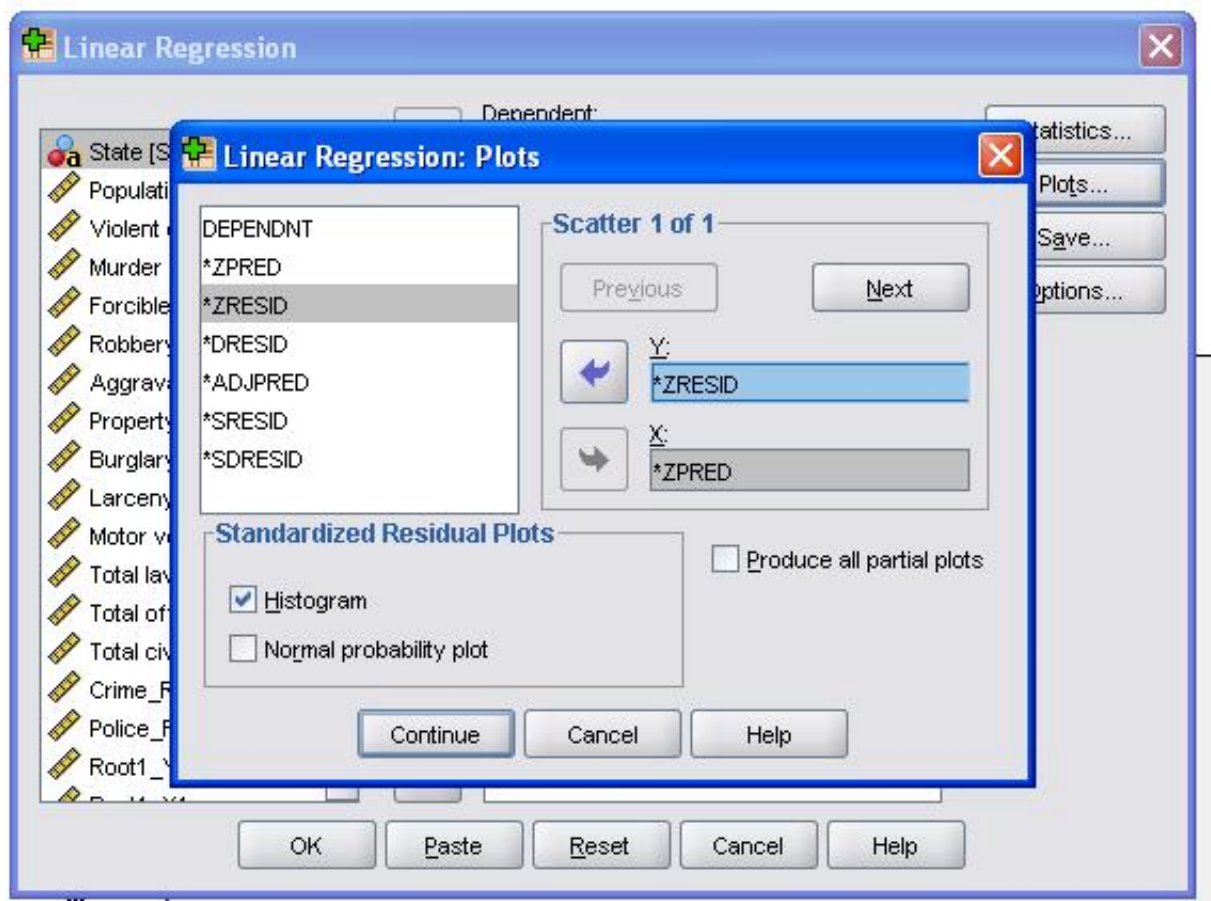
In this simple case we need to just add the variables log_pop and log_murder to the model as dependent and independent variables.



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 $y = 11.85 + 6.7 \cdot 10^{-5}$ 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

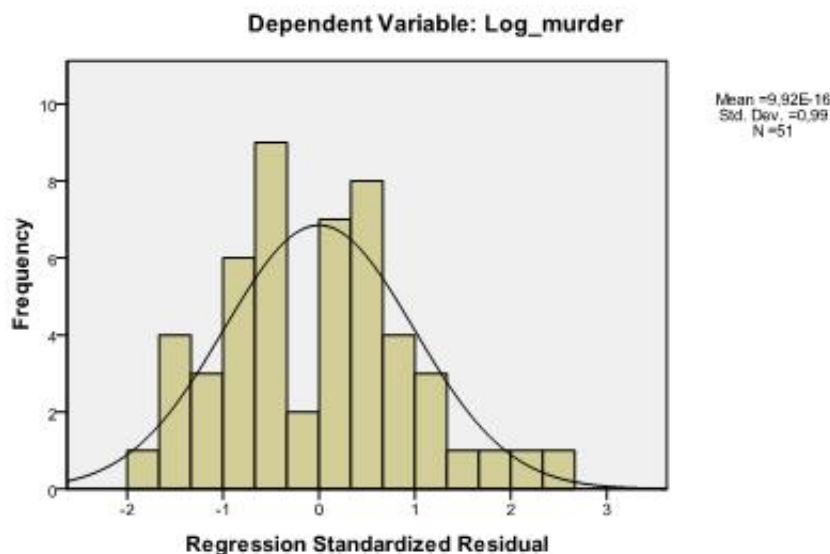
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-13,067	1,420		-9,203	,000		
	Log_pop	1,222	,098	,872	12,498	,000	1,000	1,000

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.

Histogram



Scatterplot

Dependent Variable: Log_murder

