# 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 <u>linear</u> regression analysis. We can also check the <u>Pearson's Bivariate Correlation</u> 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 Intransformation 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...

### Statistics Solutions Advancement Through Clarity https://www.statisticssolutions.com

e <u>E</u> dit ∖	<u>∕iew D</u> ata <u>T</u> ransform	Analyze Graphs Utilities	Add-	- <u>o</u> ns 1	<u>Mindow</u> Help									
	📴 🦘 🏞 揓 📴 Reports				• 🗟 🕲 🤷									
Population 2012038,0		Descriptive Statistics			Visible: 22 of 22 Variabl									
	State	State         Tables           RFM Analysis         REMANA           LABAMA         Compare Means           LASKA         General Linear Model           RIZONA         Generalized Linear Models		nt	Murder	Rape	Robbery	Assault	Property	Burglary	Theft			
1	ALABAMA			2754	268	973	5481	6032	115097	26485	7898			
2	ALASKA			8476	26	408	556	2486	17939	2638	1337			
3	ARIZONA			6534	388	1656	8613	15877	202919	44694	12493			
4	ARKANSAS	Mi <u>×</u> ed Models	•	2373	161	962	2547	8703	85540	22738	5742			
5	CALIFORNIA	<u>C</u> orrelate	•	2135	2031	7467	63403	89234	958963	196665	56052			
6	COLORADO	Regression	•		ear		3524	9589	131342	25160	8872			
7	CONNECTICUT	Loglinear	•	Cur	ve Estimation		2873	3054	63775	10419	4604			
8	DELAWARE	Neural Net <u>w</u> orks 1[ Classi <u>f</u> y		Neural Net <u>w</u> orks		R Partial Least Squares		832	1440	10615	2062	738		
9	DISTRICT OF COLUME			R Binary Logistic			3604	4453	26015	3826	1513			
10	FLORIDA Dimension Reduction		R. Multinomial Logistic			22605	46368	411125	92889	27382				
11							7603	10532	141034	29837	9458			
12	HAWAII	Nonparametric Tests	•	R. Probit.		956	1543	38310	5482	2654				
13	IDAHO	Forecasting	•				265	1911	26099	5111	1929			
14	ILLINOIS	<u>S</u> urvival	•	NLR NOT	linear		17554	21456	166468	32666	10970			
15	INDIANA	Multiple Response	Weight Estimation     Sts 2-Stage Least Squares			6595	7993	165763	33727	11446				
16	IOWA	Missing Value Anal <u>y</u> sis				es	976	4348	61007	12181	4513			
17	KANSAS	Multiple Imputation	Optimal Sc		imal Scaling (CAT	REG)	1788	6826	86775	15584	6353			
18	KENTUCKY	Complex Samples	<b>ب</b>	8159	85	678	3120	4307	71499	16403	4864			
19	LOUISIANA	Quality Control	•	5720	322	742	3846	10810	85922	22754	5429			
20	MAINE	🖉 ROC Cur <u>v</u> e		1239	11	273	352	603	25726	4316	2046			
21	MARYLAND	1204101	1	5506	317	284	6076	8829	58241	12528	3681			
22	MASSACHUSETTS	5819565	2	4201	178	1573	7317	16158	136653	31516	8868			
23	MICHIGAN	6431878	4	8635	640	3268	13447	31280	250629	57036	14870			
24	MINNESOTA	3460873		-	104	1	5290	7482	128004	22988	9408			
25	MISSISSIPPI	1050678		5059	120	577	2290	2072	55541	15126	3561			
~~		2545004	~	1077	242	4004	FEOD	40000	470505	20005	4001			

In this simple case we need to just add the variables log\_pop and log\_murder to the model as dependent and independent variables.

## Statistics Solutions

Advancement Through Clarity https://www.statisticssolutions.com

A INIC	arder and nonnegit		der		Statistics
Fo	rcible rape [Rape]	-Block 1 of 1			Plo <u>t</u> s
Ar RO	oppery [Roppery]	DIOCK TOT T			Save
	operty crime (Prop	Pre <u>v</u> ious		Next	Ontions
🖉 Bu	irglary [Burglary]		(\$):		
🔗 La	rceny-theft [Theft]	Cog_pop			
Mc Mc	otor vehicle theft [	₩			
	tal law enforceme				
To	tal civilians [Civilia	M	ethod: Enter	•	
🔗 Cri	ime_Rate	Selection Ver	iahla:		
Po	lice_Rate			Rule	
Ro	ot1_Y1	Case Labels:			
Ro	ot1_X1				
V 23	core(Root1_11)[	WISWeight			
75	COPERSION ATEL				

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

### Statistics Solutions Advancement Through Clarity

https://www.statisticssolutions.com

Ford Ford Rob Prop E Larc	uer and nor cible rape [F bery [Robbi gravated as: perty crime glary [Burg] ceny-theft [	Linear Regression Regression Coeffi Estimates Confidence interv Level(%): 95	on: Statistics icient: ☑ Mod □ R so vals □ Des □ □ Part	el fit quared change criptives and partial correla	ations	Statistics Plots Save Options
Motion Motion	or venicle tr al law enfor al officers [0 ne_Rate ce_Rate t1_Y1 t1_X1 ore(Root1	Covariance matri Residuals Durbin-Watson Casewise diagn Outliers outsic	× Colli iostics ie: 3 Cancel	nearity diagnostic: ] standard deviati	ons	

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.

D	ependent: (	
<sup>tate</sup> [S 🚰 Linear Regression: Pl	lots	
iolent i DEPENDNT urder *ZPRED orcible *ZRESID obber: *DRESID ggrav: *ADJPRED ropert: *SRESID urglar: *SDRESID	Scatter 1 of 1          Previous       Next         Y:       *ZRESID         X:       *ZPRED	Save
otor vi otal lav otal of otal civ rime_F olice_F oot1_v	Plots Produce all partial plots Cancel Help	

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<sup>2</sup> of our model is 0.756 with the R<sup>2</sup> = .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 <sup>b</sup>								
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin- Watson				
1	,872ª	,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	5
-------	---

Mode	el	Sum of Squares	df	Mean Square	F	Sig.
1	Regression	90,988	1	90,988	156,202	,000ª
	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 In 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

3								
		Unstandardize	d Coefficients	Standardized Coefficients			Collinearity Statistics	
Model		В	Std. Error	Beta	t	Sig.	Tolerance	ME
1	(Constant)	-13,067	1,420		-9,203	,000		
	Log_pop	1,222	,098	,872	12,498	,000	1,000	1,000

Coefficients\*

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.

# Dependent Variable: Log\_murder

### Histogram

Scatterplot

