

## Regression and Autocorrelation

Douglas R. White © 2006

Let's say I am an economics student and I want to check the theory that monetization comes out of urban exchange networks. Choosing variables from the Standard Cross-Cultural Sample (SCCS), my initial hypothesis is that urbanization (v152) → monetization (v155). I am aware however that the use of money requires neighboring societies with money.

In the ordering of SCCS societies in the datafile, neighboring societies are adjacent, so I can use correlations between neighbors to model the effects of having similar neighbors. To do that I switch to variable view and create three new variables in the last 3 columns of the SPSS file: z155off1, a copy of v155, and z155bak1. These will allow me to use two sets of *neighbors' levels of monetization* as predictors.

	z155off1	z155a	z155bak1
1	1.00	1	.
2	3.00	1	1.00
3	1.00	3	1.00
4	4.00	1	3.00
5	4.00	4	1.00
6	1.00	4	4.00
7	2.00	1	4.00
8	1.00	2	1.00
9	2.00	1	2.00
10	2.00	2	1.00
11	4.00	2	2.00
12	1.00	4	2.00
...	5.00	1	4.00

You do this by copying z155 into z115a in the second to last column in the file, copying cases 2-186 of z155a into the first column, cells 1-185 of z155off1, and copying 1-185 of z155 into cells 2-186 of the third column (for z155bak1). The values in successive cells form a diagonal. Once this is done you can use the SPSS menu options Analyze/Regression/Linear/Method: Stepwise. Enter v155 as the dependent variable, and v152, z155off1 and z155bak1 as the three independent variables. I click Options and choose a method of handling missing values (in this case it won't make much difference). Then I click 'Ok' to get my tables. I cut or copy and paste the following tables into the Word file mostly using (paste special, as picture).

Variables Entered/Removed(a)		(cut, paste, use word/table/autofit and then adjust row dividers)	
Model	Variables Entered	Variables Removed	Method
1	v152 Urbanization	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
2	z155bak1	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
3	z155off1	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).

a Dependent Variable: v155 Money

The second table to copy is the Model Summary

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.370 <sup>a</sup>	.137	.132	1.377
2	.431 <sup>b</sup>	.186	.177	1.341
3	.465 <sup>c</sup>	.216	.203	1.319

a. Predictors: (Constant), v152 Urbanization

b. Predictors: (Constant), v152 Urbanization, z155bak1

c. Predictors: (Constant), v152 Urbanization, z155bak1, z155off1

**(cut, paste special, as picture)**

What the Model Summary shows is that the Adjusted R square is .132 for Urbanization (Model 1), .177 including z155bak1 (Model 2), and .201 including also z155off1 (Model 3, all three predictors). That means that 20.3% of the variance in monetization is predicted by the urbanization and neighbors' monetization. Variance is sum of squared differences between actual and predicted divided by variance in the dependent variables.

The third table to copy is the Anova (*An\_alysis o\_f va\_riance*) result, which gives the names of the variables. The 'Sig.' column shows that the chances of these results occurring by chance (what is called "statistical significance") are less than  $p=.001$ ,

**ANOVA<sup>d</sup>**

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	54.721	1	54.721	28.855	.000 <sup>a</sup>
	Residual	345.144	182	1.896		
	Total	399.864	183			
2	Regression	74.349	2	37.175	20.671	.000 <sup>b</sup>
	Residual	325.515	181	1.798		
	Total	399.864	183			
3	Regression	86.529	3	28.843	16.569	.000 <sup>c</sup>
	Residual	313.335	180	1.741		
	Total	399.864	183			

a. Predictors: (Constant), v152 Urbanization

b. Predictors: (Constant), v152 Urbanization, z155bak1

c. Predictors: (Constant), v152 Urbanization, z155bak1, z155off1

d. Dependent Variable: v155 Money

**(cut, paste special, as picture)**

Comparing the Model Summary again, urbanization accounts for .132 or 13.2% of the variance in monetization, back-neighbor (z155bak1) accounts for an additional 4.5% (.177-.132), and forward-neighbor (z155off1) accounts for an additional 2.6% (.203-.177) variance, for a total of  $13.2+4.5+2.6 = 20.3\%$  variance total. The total variance accounted for by neighbors is only 7.1% compared to urbanization 13.2%.

The fourth table to copy shows the coefficients for each variable in the three different models and in the 'Sig.' column the chances of these results occurring by chance (what is called "statistical significance"). In model 3, probabilities by change are less than  $p=.001$ ,  $p=.006$ , and  $p=.009$  for the three variables.

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.512	.215		7.047	.000
	v152 Urbanization	.391	.073	.370	5.372	.000
	z155bak1					
	z155off1					
2	(Constant)	1.038	.253		4.097	.000
	v152 Urbanization	.355	.072	.336	4.955	.000
	z155bak1	.224	.068	.224	3.304	.001
	z155off1					
3	(Constant)	.740	.274		2.706	.007
	v152 Urbanization	.327	.071	.309	4.580	.000
	z155bak1	.191	.068	.191	2.809	.006
	z155off1	.180	.068	.180	2.645	.009

a. Dependent Variable: v155 Money

(cut, paste special, as picture)

## Discussion

One of the most important problems in empirical work that is nonexperimental is that the cases studied are nonindependent: neighbors interact and affect one another, or common effects propagate, diffuse or travel along channels of communication. When this is true the null hypothesis of statistical independence has to be modified. It is no longer valid to assume, like a coin tossing experiment, that the outcomes are independent. When there are common effects, the deviations for the independence assumption (or expectation of any given hypothesis, for that matter) are said to be *autocorrelated* along the travel-paths of the common effects. Since neighbors are the most common paths of common effects or local interactions, the easiest way to incorporate autocorrelation is to include in the model the attributes of neighbors as variables that produce common effects. That is what we have done in this example.

The problem of autocorrelation is especially prevalent but not restricted to the study of human culture: cultural features are free to travel, so to speak. Anthropologists actually discovered and gave a name to autocorrelation as *Galton's Problem* in reference to a 1896 article by Sir Francis Galton, commenting on correlational analysis by Sir E. B. Tylor.

For live links see:

Statistics for Cross-Cultural Research @ <http://eclectic.ss.uci.edu/~drwhite/xc!/XC-BK8.pdf>

The evolution of Human Diversity @ <http://www.ucl.ac.uk/heeg/culture.htm>

Further live link (the second requires uci VPN access with your UCInetID):

Andrey Korotayev and Victor de Munck. 2003. "Galton's Asset" and "Flower's Problem": Cultural Networks and Cultural Units in Cross-Cultural Research (Or, the Male Genital Mutilations and Polygyny in Cross-Cultural Perspective). *American Anthropologist* 105(2):353-358. <http://www.anthrosource.net/doi/abs/10.1525/aa.2003.105.2.353>  
<https://vpn.nacs.uci.edu/http/0/www.anthrosource.net/doi/abs/10.1525/aa.2003.105.2.353>

## *Revisiting our Initial Hypothesis*

Because 20% of the variance in monetization is left explained 80% is left unexplained. Is it possible that my hypothesis is too simple and that I had left out other factors that would predict monetization. For this purpose it would be relevant to examine the other complexity variables for the Standard Cross-Cultural Sample (SCCS) are listed below from <http://eclectic.ss.uci.edu/~drwhite/courses/SCCCodes.htm>.

Murdock, George P., and Caterina Provost. 1971. *ETHNOLOGY* 12:379-392.

Datafile: STDS06.DAT Vars. 149-158 cultural complexity

### 149. SCALE 1- WRITING AND RECORDS

73	1 = None
49	2 = Mnemonic devices
21	3 = Nonwritten records
12	4 = True writing; no records
31	5 = True writing; records

### 150. SCALE 2- FIXITY OF RESIDENCE

28	1 = Nomadic
21	2 = Seminomadic
20	3 = Semisedentary
15	4 = Sedentary; impermanent
102	5 = Sedentary

### 151. SCALE 3- AGRICULTURE

38	1 = None
17	2 = 10% food supply
11	3 = 10%; secondary
63	4 = Primary; not intensive
57	5 = Primary; intensive

### 152. SCALE 4- URBANIZATION

56	1 = fewer than 100 persons
43	2 = 100-199 persons
33	3 = 200-399 persons
30	4 = 400-999 persons
24	5 = 1000 persons

### 153. SCALE 5- TECHNOLOGICAL SPECIALIZATION

39	1 = None
27	2 = Pottery only
31	3 = Loom weaving only
56	4 = Metalwork only
33	5 = Smiths, weavers, potters

### 154. SCALE 6- LAND TRANSPORT

108	1 = Human only
42	2 = Pack animals
14	3 = Draft animals
11	4 = Animal-drawn vehicles

11 5 = Automotive vehicles

155. SCALE 7- MONEY

77 1 = None  
14 2 = Domestically usable articles  
43 3 = Alien currency  
27 4 = Elementary forms  
25 5 = True money

156. SCALE 8- DENSITY OF POPULATION

58 1 = less than 1 person/square mile  
25 2 = 1-5 persons/square mile  
28 3 = 5.1-25 persons/square mile  
35 4 = 26-100 persons/square mile  
40 5 = 100 persons/square mile

157. SCALE 9- POLITICAL INTEGRATION

11 1 = None  
72 2 = Autonomous local communities  
46 3 = 1 level above community  
28 4 = 2 levels above community  
29 5 = 3 levels above community

158. SCALE 10- SOCIAL STRATIFICATION

65 1 = Egalitarian  
52 2 = Hereditary slavery  
19 3 = 2 social classes, no castes/slavery  
20 4 = 2 social classes, castes/slavery  
30 5 = 3 social classes or castes, with or without slavery

158.1. SUM OF CULTURAL COMPLEXITY (v149-v158)

The “Sum of cultural complexity” includes the variable we want to predict (v155), and so cannot be used as a predictor because that would create a circular argument. We can, however, consider all the other variables simultaneously (149-151, 153-154, and 156-158). If I include all these additional variables in the regression, I get the following results: four variables jointly predict 41.6% of the variance in monetization, as shown in the model summary. One of them is *Land Transport*, for which higher values entail good connections to neighbors. The neighbors’ levels of monetization are no longer predictive. But my initial variables and hypothesis about urbanization is also not predicting. Instead, I find that *Density of Population* is the major predictor (30.7%) that when supplemented by *Writing and Records* jointly predict 42.8%. This is a very strong result. *Land Transport* and *Social Neighborhood* predict small residual variances, and *Land Transport* squared is a slightly better predictor.

**Variables Entered/Removed <sup>a</sup>**

Model	Variables Entered	Variables Removed	Method
1	v156 Density of Population	.	Stepwise (Criteria: Probability-of-F-to-enter <= .099, Probability-of-F-to-remove >= .150).
2	v149 Writing and Records	.	Stepwise (Criteria: Probability-of-F-to-enter <= .099, Probability-of-F-to-remove >= .150).
3	v154 Land Transport	.	Stepwise (Criteria: Probability-of-F-to-enter <= .099, Probability-of-F-to-remove >= .150).
4	NghAverage	.	Stepwise (Criteria: Probability-of-F-to-enter <= .099, Probability-of-F-to-remove >= .150).

<sup>a</sup>. Dependent Variable: v155 Money

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.557 <sup>a</sup>	.310	.307	1.231
2	.659 <sup>b</sup>	.435	.428	1.118
3	.668 <sup>c</sup>	.446	.437	1.110
4	.674 <sup>d</sup>	.455	.443	1.104

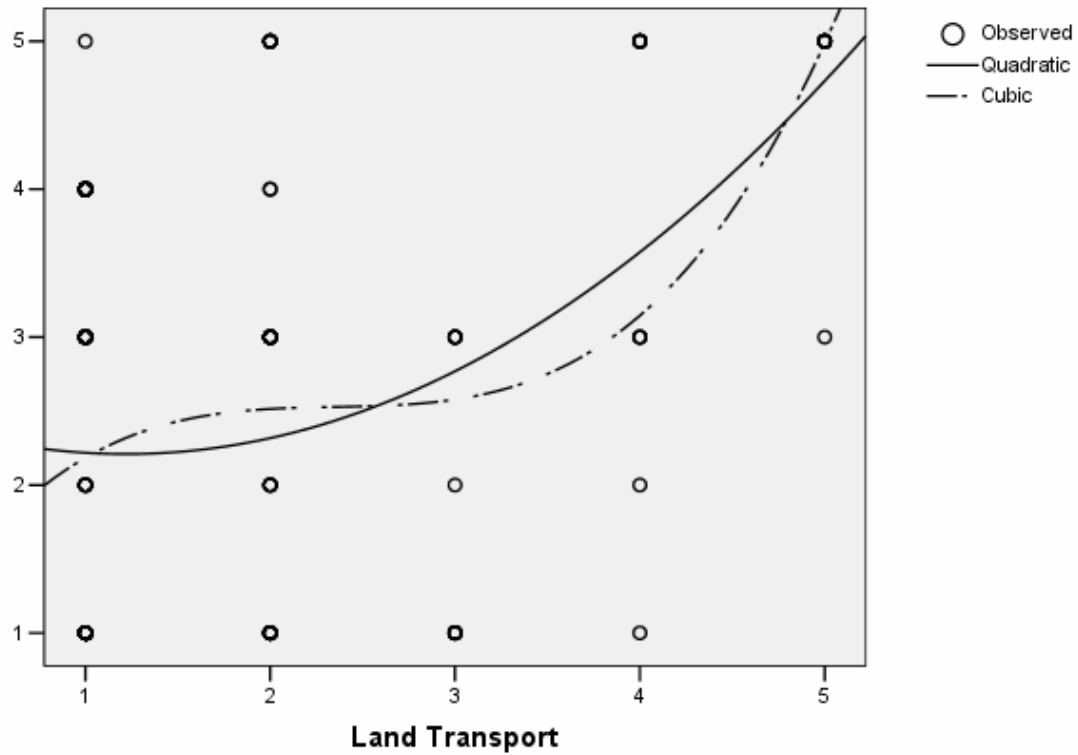
<sup>a</sup>. Predictors: (Constant), v156 Density of Population

<sup>b</sup>. Predictors: (Constant), v156 Density of Population, v149 Writing and Records

<sup>c</sup>. Predictors: (Constant), v156 Density of Population, v149 Writing and Records, v154 Land Transport

<sup>d</sup>. Predictors: (Constant), v156 Density of Population, v149 Writing and Records, v154 Land Transport, NghAverage

## Money



### Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.557 <sup>a</sup>	.310	.307	1.231
2	.659 <sup>b</sup>	.435	.428	1.118
3	.671 <sup>c</sup>	.450	.441	1.105
4	.678 <sup>d</sup>	.460	.448	1.098

a. Predictors: (Constant), v156 Density of Population

b. Predictors: (Constant), v156 Density of Population, v149 Writing and Records

c. Predictors: (Constant), v156 Density of Population, v149 Writing and Records, TransprtSquared

d. Predictors: (Constant), v156 Density of Population, v149 Writing and Records, TransprtSquared, z155off1

**ANOVA<sup>e</sup>**

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	124.213	1	124.213	81.941	.000 <sup>a</sup>
	Residual	275.893	182	1.516		
	Total	400.106	183			
2	Regression	173.864	2	86.932	69.548	.000 <sup>b</sup>
	Residual	226.242	181	1.250		
	Total	400.106	183			
3	Regression	178.477	3	59.492	48.318	.000 <sup>c</sup>
	Residual	221.628	180	1.231		
	Total	400.106	183			
4	Regression	181.989	4	45.497	37.338	.000 <sup>d</sup>
	Residual	218.117	179	1.219		
	Total	400.106	183			

a. Predictors: (Constant), v156 Density of Population

b. Predictors: (Constant), v156 Density of Population, v149 Writing and Records

c. Predictors: (Constant), v156 Density of Population, v149 Writing and Records, v154 Land Transport

d. Predictors: (Constant), v156 Density of Population, v149 Writing and Records, v154 Land Transport, NghAverage

e. Dependent Variable: v155 Money

The coefficients table shows that the first two predictors are both significant at  $p < .001$ , the third at  $p < .05$ , and the fourth at  $p < .09$ , which should not be considered significant given that we used so many variables (nine in all) to try to make a prediction (on average, a random expectation would be  $p \sim .09$ ).

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.997	.190		5.241	.000
	v156 Density of Population	.529	.058	.557	9.052	.000
	v149 Writing and Records					
	v154 Land Transport					
	NghAverage					
2	(Constant)	.474	.192		2.474	.014
	v156 Density of Population	.399	.057	.420	7.008	.000
	v149 Writing and Records	.381	.060	.378	6.303	.000
	v154 Land Transport					
	NghAverage					
3	(Constant)	.364	.199		1.835	.068
	v156 Density of Population	.402	.057	.423	7.103	.000
	v149 Writing and Records	.293	.075	.290	3.887	.000
	v154 Land Transport	.173	.089	.138	1.936	.054
	NghAverage					
4	(Constant)	.160	.231		.691	.491
	v156 Density of Population	.376	.058	.396	6.469	.000
	v149 Writing and Records	.269	.076	.267	3.531	.001
	v154 Land Transport	.172	.089	.137	1.939	.054
	NghAverage	.066	.039	.103	1.698	.091

a. Dependent Variable: v155 Money

So *Population Density* is my biggest predictor. The final table uses a crosstabulation with correlations and significance, using option Analyze/Descriptive Statistics/Crosstabs with

Row variable 155, Column variable 156, and Layers variable 200, which is Region. The Cramer's V correlations for v155 with v156 are all positive and range between .421 and .729. That is a very strong replication.

Symmetric Measures					
v200 Region			Value	Asymp. Std. Error <sup>a</sup> Approx. T <sup>b</sup> Approx. Sig.	
1 Africa	Nominal by	Phi	.877		.159
	Nominal	Cramer's V	.438		.159
	N of Valid Cases		28		
2 Circum-Mediterranean	Nominal by	Phi	.730		.246
	Nominal	Cramer's V	.421		.246
	N of Valid Cases		28		
3 East Eurasia	Nominal by	Phi	1.028		.003
	Nominal	Cramer's V	.514		.003
	N of Valid Cases		34		
4 Insular Pacific	Nominal by	Phi	.844		.140
	Nominal	Cramer's V	.422		.140
	N of Valid Cases		31		
5 North America	Nominal by	Phi	.595		.472
	Nominal	Cramer's V	.343		.472
	N of Valid Cases		33		
6 South America	Nominal by	Phi	1.031		.000
	Nominal	Cramer's V	.729		.000
	N of Valid Cases		32		

<sup>a</sup>. Not assuming the null hypothesis.

<sup>b</sup>. Using the asymptotic standard error assuming the null hypothesis.

We have now come full circle: a well specified model that has only a tiny residual Galton's Problem effect because there little effect of neighbors taking into account *Population Density* and *Writing and Records*. Accounting for 45% of the variance is a cross-cultural variable is a significant prediction.

**Model Summary and Parameter Estimates**

Dependent Variable: v155 Money

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.168	37.029	1	184	.00000001	1.591	.514		
Logarithmic	.137	29.126	1	184	.00000002	2.091	1.007		
Inverse	.109	22.551	1	184	.00000041	3.682	-1.572		
Quadratic	.194	22.019	2	183	.000000003	2.471	-.430	.176	
Cubic	.206	15.772	3	182	.000000004	.817	2.142	-.903	.128
Compound	.126	26.524	1	184	.000001	1.453	1.216		
Power	.106	21.921	1	184	.000006	1.753	.390		
S	.088	17.757	1	184	.00004	1.185	-.619		
Growth	.126	26.524	1	184	.000001	.374	.195		
Exponential	.126	26.524	1	184	.000001	1.453	.195		
Logistic	.126	26.524	1	184	.000001	.688	.822		

The independent variable is v154 Land Transport.