Multiple Linear Regression

Spatial Application II: Village Accessibility, 1940-2000

Equations taken from Zar, 1984.

$\hat{y} = a + b_1 x_1 + b_2 x_2 \ldots + b_n x_n$ **where n is the number of variables**

Example: The data table to the right contains three measures of accessibility for 40 villages and towns in Michoacán, Mexico. Valued accessibility is a measure of the sum of the shortest linear distance it takes to connect one location to all other locations. Degree of circuity is a measure of how well placed a location is on a transportation network, higher values meaning that a location is accessible by circuitous routes. Circuity is measured as the difference between the actual route and a straight line. The variables are valued accessibility in 2000 (**L-2000**), valued accessibility in 1940 (**L-1940**), and degree of circuity in 1940 (**DC-1940**). This area has seen improvements in its road system since 1940. Our hypothesis is that places that were highly accessible in 2000 would have been highly accessible in 1940. Places that are not well predicted by our model will be of particular interest since those places would have experienced either an increase or decrease in accessibility between 1940 and 2000.

In multiple regression there is another assumption that we must take into account: that of multicollinearity. When the independent variables are themselves related, it is termed multicollinearity. Remember that in regression we partition out the explanatory power of one variable while holding the others constant. Think of multicollinearity as the amount of correlation between the independent variables. This correlation is the overlap in explanatory power and makes it impossible to determine which of the independent variables is "explaining" the dependent variable. In other words, we can not hold the variables constant since they are associated with each other.

We will use the 1940 accessibility measures to predict the 2000 accessibility measure using multiple linear regression. Since this process is very similar to that of bivariate linear regression, we will let SPSS do the calculations.

Village	L-2000	L-1940	DC-1940
Ahuiran	2321.65	2067.61	4.57
Ajuno	2495.25	2277.47	3.05
Angahuan	3332.90	3117.88	15.11
Arantepacua	2203.89	2017.39	5.97
Aranza	2158.40	1931.49	4.33
Capacuaro	2386.20	2204.71	6.03
Charapan	3371.97	3295.21	19.41
Cheran	2017.79	1754.53	2.61
Cheranatzicurin	2361.68	1975.46	4.41
Cocucho	3010.60	3417.54	15.98
Comachuen	2208.99	2141.57	7.87
Corupo	3139.34	3129.34	17.85
Erongaricuaro	2326.42	2045.13	2.37
Ihuatzio	3344.74	4579.66	47.71
Jaracuaro	2596.82	2344.09	3.3
La Mojonera	2314.84	2406.9	8.8
Nahuatzen	1898.23	1786.39	4.07
Nurio	2567.76	2394.04	5.9
Paracho	2208.18	1953.99	4.38
Patzcuaro	2813.23	2748.86	3.58
Pichataro	2103.43	1866.97	2.4
Pomacuaran	2465.71	2375.97	7.11
Puacuaro	2576.92	2338.28	4.05
Quinceo	2283.61	2119.09	6.62
Quiroga	3478.46	3780.33	16.42
San Andres Tzirondaro	2904.81	3342.38	15.65
San Felipe	2855.60	2508.67	5.29
San Isidro-Nahuatzen	1986.06	2038.65	5.19
San Jeronimo Purenchecuaro	3061.24	3442.83	15.74
San Juan Tumbio	2255.13	2183.99	3.05
San Lorenzo	2736.30	2726.71	16.9
Santa Fe de la Laguna	3368.59	3702.54	17.32
Sevina	1883.44	1838.26	5.25
Tingambato	2916.13	2307.76	4.98
Turicuaro	2317.47	2070.96	6.64
Tzintzuntzan	3361.44	3915.81	26.76
Urapicho	2715.20	2585.5	8.65
Uricho	2387.98	2093.18	2.45
Uruapan	2984.07	2930.24	11.61
Zurumutaro	2935.64	4021.03	34.3

SPSS Output for the Village Accessibility, 1940-2000

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.932[a]	.869	.862	170.31801	2.197

a. Predictors: (Constant), DC-1940, L-1940

b. Dependent Variable: L-2000

In our model valued accessibility in 2000 (L-2000) is the dependent variable, while valued accessibility in 1940 (L-1940) and degree of circuity in 1940 (DC-1940) are our independent variables. These two variables explain 86.2% of the variation in valued accessibility in 2000.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7151033	2	3575516.513	123.259	.000[a]
	Residual	1073304	37	29008.224		
	Total	8224337	39			

a. Predictors: (Constant), DC-1940, L-1940

b. Dependent Variable: L-2000

The F test is significant, meaning that there is some change in the $\boldsymbol{y}$ *values for changes in the* $\boldsymbol{x_1}$ *and* $\boldsymbol{x_2}$ *variables, and that the variation explained by the mode is not due to chance.*

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	532.596	172.506		3.087	.004
	L-1940	.915	.088	1.432	10.372	.000
	DC-1940	-28.733	6.692	-.593	-4.293	.000

a. Dependent Variable: L-2000

The regression parameters are listed in the above table resulting in the model:

$$\boldsymbol{L-2000 = 532.596 + 0.915(L-1940) - 28.733(DC-1940)}$$

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.932[a]	.869	.862	170.31801	2.197

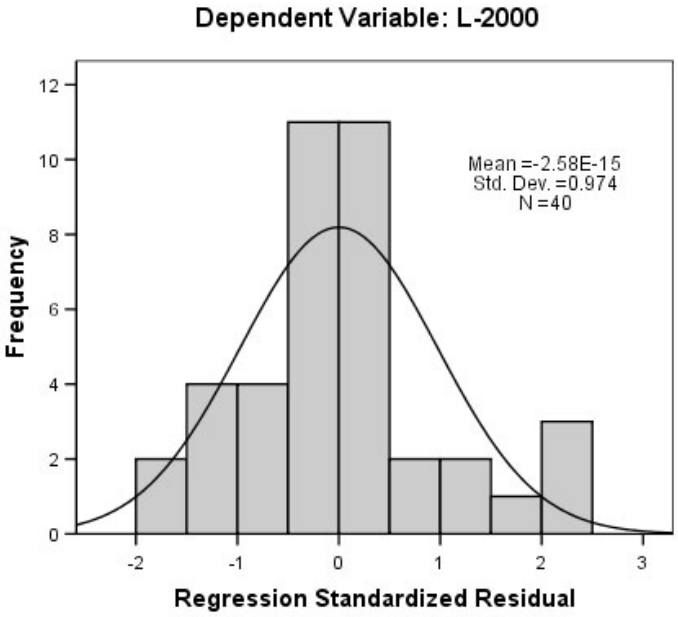
a. Predictors: (Constant), DC-1940, L-1940

b. Dependent Variable: L-2000

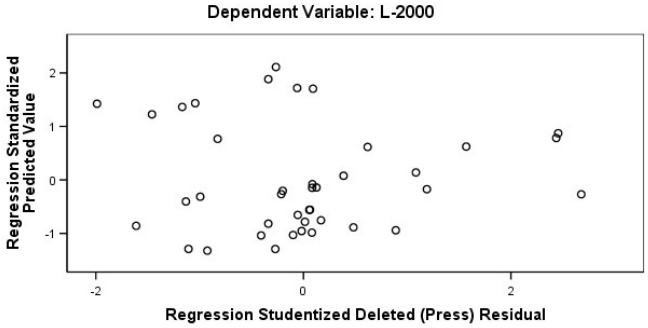
Descriptive Statistics

	Mean	Std. Deviation	N
L-2000	2616.4028	459.21721	40
L-1940	2594.4603	718.79555	40
DC-1940	10.0920	9.47487	40

Without prior knowledge of 1940 accessibility measures the best guess at the 2000 accessibility is 2616.4, with a standard deviation of 459.2. Note that the standard error of the estimate is only 170.3…further evidence that our model is useful at accurately predicting accessibility.



The residuals appear to be rather normally distributed (above) and there do not appear to be any outliers (below).



Correlations

		L-2000	L-1940	DC-1940
Pearson Correlation	L-2000	1.000	.897	.700
	L-1940	.897	1.000	.903
	DC-1940	.700	.903	1.000
Sig. (1-tailed)	L-2000	.	.000	.000
	L-1940	.000	.	.000
	DC-1940	.000	.000	.
N	L-2000	40	40	40
	L-1940	40	40	40
	DC-1940	40	40	40

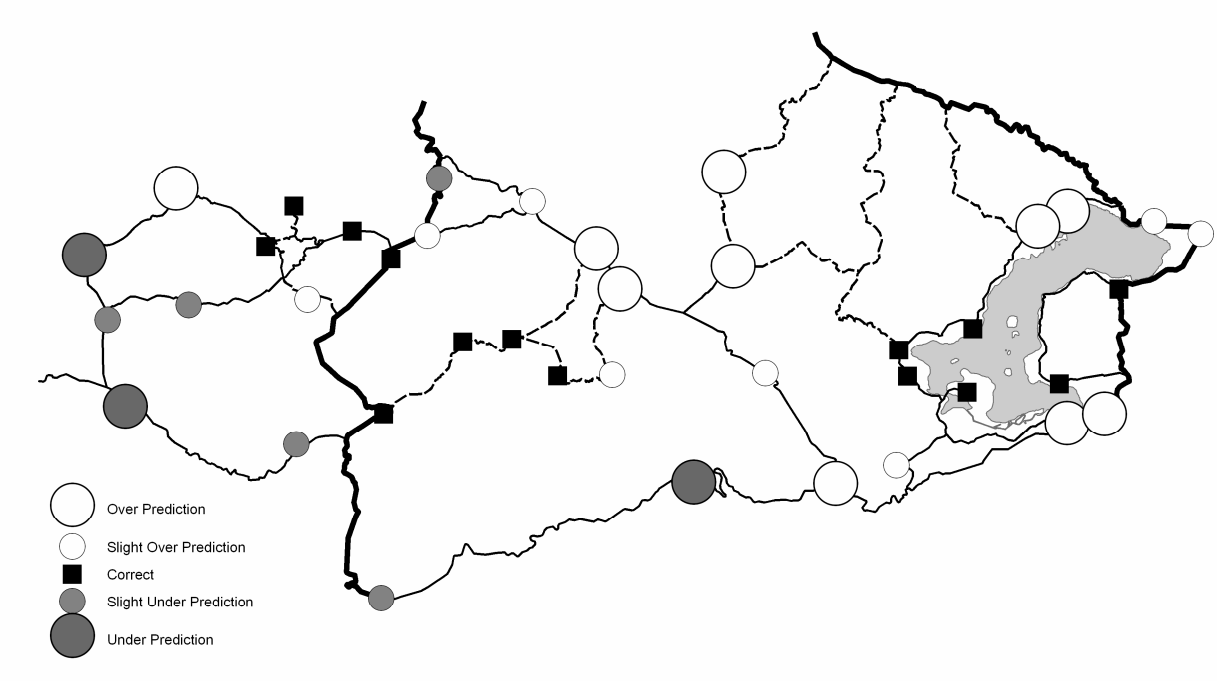
*However, look at the correlation matrix for the variables. L-1940 and DC-1940 appear to be highly correlated with each other (**0.903**).*

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Correlations			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	532.596	172.506		3.087	.004					
	L-1940	.915	.088	1.432	10.372	.000	.897	.863	.616	.185	5.405
	DC-1940	-28.733	6.692	-.593	-4.293	.000	.700	-.577	-.255	.185	5.405

a. Dependent Variable: L-2000

*You can access the collinearity assessment tools through **Analyze > Regression > Linear > Statistics** and then click on the **Collinearity diagnostics** radio button. Tolerance is the amount of the variance in a given independent variable what **can not** be explained by other independent variables. In this case **ONLY** 18.5% of the variance in one can not be explained by the other… meaning that 81.5% of the variance **IS** shared or collinear! Also, VIFs (variance inflation factors) higher than 2 are considered problematic and our VIFs are over 5. These independent variables are highly correlated with each other, and that "shared" variance will be partitioned into the ESS (residuals), lowering the accuracy of our predicted values.*



After mapping the residual values we can see that only a few towns accessibility levels were under predicted. Over prediction is a much more common occurrence. The towns that were over predicted saw a decrease in their accessibility relative to that of the other towns. Conversely, the towns that were under predicted saw an increase in their accessibility relative to that of the other towns. However, we can not be sure of the accuracy of our predicted values since the independent variables were not, in fact, independent. The next step of this process would be to drop or substitute one of the independent variables and see if the model improves.