TASK

Using the database of more than 1,000 respondents from two cities, study a particular characteristic of social activity. Use the wide-spread, informal definition of this characteristic to see how it can be defined in terms of variables in the data set. Split the residents by location and a particular legal status. Perform analysis separately in each of the resulting groups, whenever possible. Determine which factors influence the given characteristic of social activity. As factors consider demographics, family status as well as several indicators of income and intelligence.

SOLUTION

The analysis below is an abridged version of the solution provided to the client. In particular, client's version included all the SPSS output of the exploratory analysis, principal components, regression analysis and ANOVA. Here we display only the most relevant SPSS output.

ANALYSIS OF A PARTICULAR TYPE OF SOCIAL ACTIVITY IN TWO CITIES

Data Preparation

The following modifications were made to the data set.

- For each numerical (scale or ordinal) variable, if the response was "refuse" or "do not know" it was treated as a missing value. The reason was difficulty of putting such response anywhere on the scale. The only exception was answer "not sure/don't know" to one question that began as "How interested are you in..." This answer contained information about the strength of interest and could be placed between answers "not interested" and "somewhat interested".
- If removal of respondents answering "refuse" or "do not know" lead to the categorical (nominal) variable having only two possible values, we treated the variable as a binary numerical variable and used it in regression models.
- Categorical variables C1 and C2 did not have enough respondents in several categories for the purposes of ANOVA analysis. This effect would especially be pronounced if analyzing those variables separately for each location and each type of legal status L1. Therefore, categories "C1-A", "C1-C", "C1-D", "C1-E", "C1-G" and "C1-H" of C1 were merged into category "Other". Categories of "C2-E" and "C2-F" of C2 were merged. Still certain cells were not populated sufficiently for certain combinations of location and L1. For that reason, ANOVA analysis with factors C1 and C2 was performed separately for each legal status L1 but not location. Respondents from the two cities were pulled together.
- We distinguished three types of legal status L1: "L1-A", "L1–B" and "L1-C".

Social Activity Index Construction

- 11 questions shed light on the social activity characteristic of interest (**SA**). They correspond to 11 variables. Variables S1 – S7, S9 and S11 are defined for most respondents. Variables S8 and S10 are defined for categories "L1-A" and "L1-B" only.
- We were running the risk of obtaining false significance when analyzing too many variables and models. Therefore we combined all the social activity variables into a single index for each respondent. For those who were "L1-A" or "L1-B", all 11 variables were the constituents. For

respondents that were “L1-C” the constituents were all social activity variables except for S8 and S10.

- Much thought was given to how to combine the variables into a single index. The original assumption was that there was a single main factor driving all the variables and that factor was the strength of interest in certain aspects of the life of the society (not disclosed in this case study). However, the principal component analysis revealed several strong factors driving the 11 variables. The principal component analysis was run separately for group “L1-A” & “L1-B” and group “L1-C”. Please see the corresponding SPSS output. Among other things, the output displays the loadings of principal components and their relative importance. Because in each group there were several important factors, it was not clear which one should have been chosen as the index. It was somewhat expected that the main principal component was likely to represent a particular kind of social conscience or activeness of the respondent. However we needed a formal verification of that.
- We analyzed all 11 variables and saw that each of them contained important information about SA. That’s why we decided to allow them to play equal roles in the index. Mathematically that meant that, in group “L1-A” & “L1-B”, we standardized the 11 constituents and summed them up with equal weights to form index SA_A&B. In group “L1-C”, we standardized the 9 constituents and summed them up with equal weights to form index SA_C. Because the standardized version of any variable could go negative our indexes could take negative values as well.
- Whenever the values of a subset of constituents were missing we recorded the index as missing.
- At the end, we found it reassuring that the correlation of SA_A&B with the main principal component in group “L1-A” & “L1-B” was -0.951, while the correlation of SA_C with the main principal component in group “L1-C” was -0.838. That meant that our way of defining the index was not much different from the way using the main principal component (up to the change of sign).

Relation of Social Activity to Independent Variables

As the next step, we researched if any of the independent variables had predictive power for SA_A&B or SA_C. We had 8 independent variables in the data set. Six of them were numerical (ordinal, to be more exact): Q1, Q2, Q3, Q4, Q5 and Q6. Two independent variables were categorical: C1 and C2.

Let us remind you that the data set was small, especially considering all the missing observations and our objective to split the analysis by location and legal status L1, wherever possible. Therefore, analyzing the influence of each categorical variable value by value had limitations. We could not get a precise grasp of what each category meant for SA. Therefore we chose to run ANOVA to determine if the categorical variable altogether had any predictive power for SA. ANOVA analysis allowed us to answer the question whether the level of SA was different in different categories.

The conclusions:

- C1 has no predictive power for SA for all types of L1. The ANOVA F-test has p-values 0.41, 0.47 and 0.75 for groups “L1-A”, “L1-B” and “L1-C” respectively. In all three groups we accept the null hypotheses about the same level of SA in all C1 categories.
- For groups “L1-A” and “L1-B, variable C2 has strong predictive power for SA (p-values 0.003 and 0.01 respectively). Bonferroni post hoc tests reveal that there is significant difference in SA between categories “C2-A” and “C2-D”.
- On the contrary, for those in group “L1-C”, characteristic SA is uniform over the categories of C2. The p-value of the F-test is 0.68. This means that the F-statistic is not significant and we have to accept the null hypothesis.

Finally, we wanted to sense how strong the influence of numerical predictors on SA was. We could not run any ordered logistic or probit regression because the separate constituents were ordinal but the indexes were not. Therefore we approached the task within the framework of multiple linear regression.

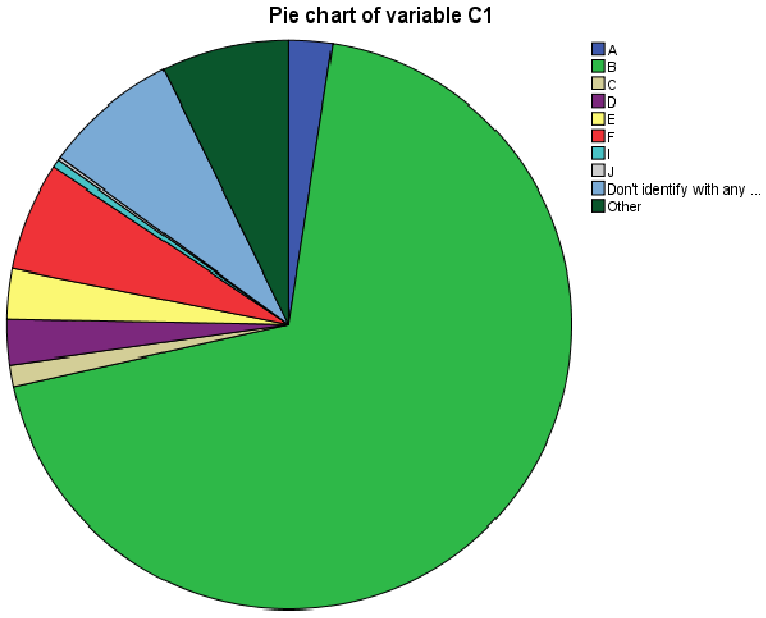
We could not indulge in more complicated, non-linear models because of the scarcity of the data. The analysis was run separately for each location and each category of L1. Each time we started with building the full main effects model. Then we pruned it and grew until we obtained the best linear model with all significant predictors. The SPSS output with the best linear models is placed at the end of this document.

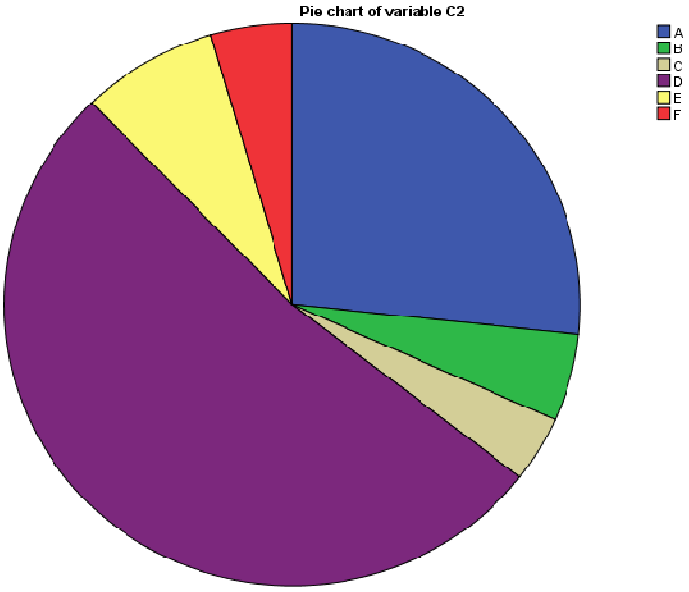
The conclusions:

- In each study (for each location and category of L1), Q6 is a strong and significant predictor of SA. Always the higher level of Q6 implies the higher level of SA. This is true in each city and each L1 category.
- In selected cases other predictors are statistically significant as well. Q4 is statistically significant for those a) having status “L1-A” and living in city A, b) having status “L1-B” and living in city B... People with higher Q4 tend to be more socially active (according to the studied measure).
- Q1 is statistically significant for those a) having status “L1-A” and living in city A, b) having status “L1-B” and living in city A, c) having status “L1-B” and living in city B... People with higher Q1 tend to be more socially active.
- Q5 is statistically significant for those having status “L1-A” and living in city B... More Q5-succesfull people tend to be more socially active.
- Q3 is statistically significant for a) having status “L1-B” and living in city B, b) having status “L1-C” and living in city B... More Q3-fit people tend to be more socially active.
- Oftentimes, the effect of a predictor on SA could not be identified accurately because of the scarcity of the data in the study. With more data, more predictors could come out to be statistically significant. This is our suspicion.
- In each study (for each location and category of L1), the “optimal” model satisfies the distributional assumptions of linear regression. The residuals are normally distributed, as seen from the Kolmogorov-Smirnov tests. In other words, the “optimal” model fits the data relatively well.
- Even though some of the independent variables are highly significant, they cannot predict SA accurately. Even when they act together, as a group, the R^2 of the regression is quite low. Usually it is in the zip code of 5-20%. This tells us that there are other very influential variables which are not included in our analysis.

Most Relevant SPSS output

1. Selected Descriptives



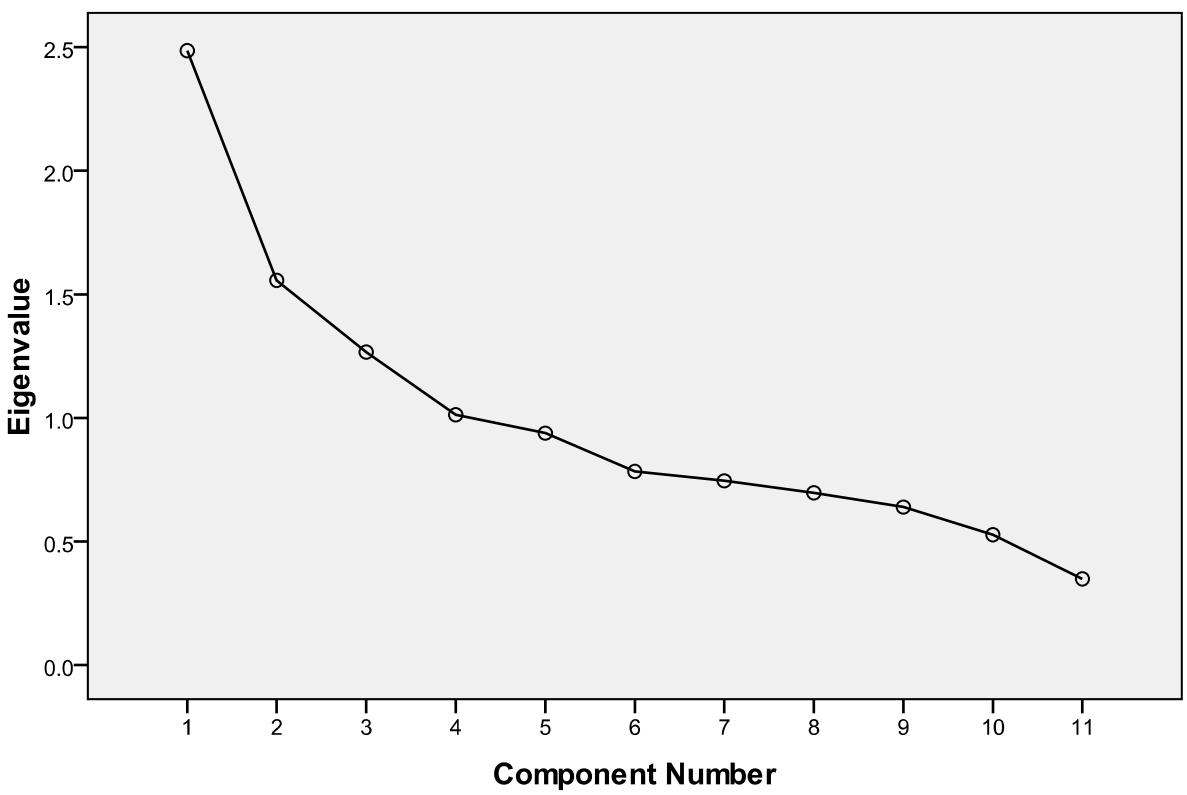


2. **Principal Components of Variables S1 – S11 for groups “L1-A” an “L1-B”**

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.485	22.595	22.595	2.485	22.595	22.595
2	1.556	14.149	36.743	1.556	14.149	36.743
3	1.266	11.512	48.255	1.266	11.512	48.255
4	1.013	9.205	57.461	1.013	9.205	57.461
5	.938	8.530	65.991			
6	.783	7.120	73.111			
7	.745	6.777	79.888			
8	.697	6.338	86.226			
9	.639	5.810	92.036			
10	.527	4.795	96.831			
11	.349	3.169	100.000			

Extraction Method: Principal Component Analysis.



Component Matrix[a]

	Component			
	1	2	3	4
Z-score ofS1	-.528	.180	.316	-.403
Z-score ofS2	.424	.538	-.225	.225
Z-score ofS3	.493	-.324	-.234	.277
Z-score ofS4	-.211	-.329	-.158	.587
Z-score ofS5	-.574	.183	.210	.278
Z-score ofS6	.372	.479	-.288	-.169
Z-score ofS7	.478	-.419	-.321	-.205
Z-score ofS8	.597	-.008	.662	.151
Z-score ofS9	.405	.635	-.115	-.005
Z-score ofS10	.662	-.073	.593	.063
Z-score ofS11	.290	-.416	-.055	-.451

Extraction Method: Principal Component Analysis.

a. 4 components extracted.

Correlations

		SA_A&B	FAC1_1	FAC2_1	FAC3_1	FAC4_1
SA_A&B	Pearson Correlation	1	-.951**	.060	.150**	-.134**
	Sig. (2-tailed)		.000	.163	.000	.002
	N	539	539	539	539	539
FAC1_1	Pearson Correlation	-.951**	1	.000	.000	.000
	Sig. (2-tailed)	.000		1.000	1.000	1.000
	N	539	539	539	539	539
FAC2_1	Pearson Correlation	.060	.000	1	.000	.000
	Sig. (2-tailed)	.163	1.000		1.000	1.000
	N	539	539	539	539	539
FAC3_1	Pearson Correlation	.150**	.000	.000	1	.000
	Sig. (2-tailed)	.000	1.000	1.000		1.000
	N	539	539	539	539	539
FAC4_1	Pearson Correlation	-.134**	.000	.000	.000	1
	Sig. (2-tailed)	.002	1.000	1.000	1.000	
	N	539	539	539	539	539

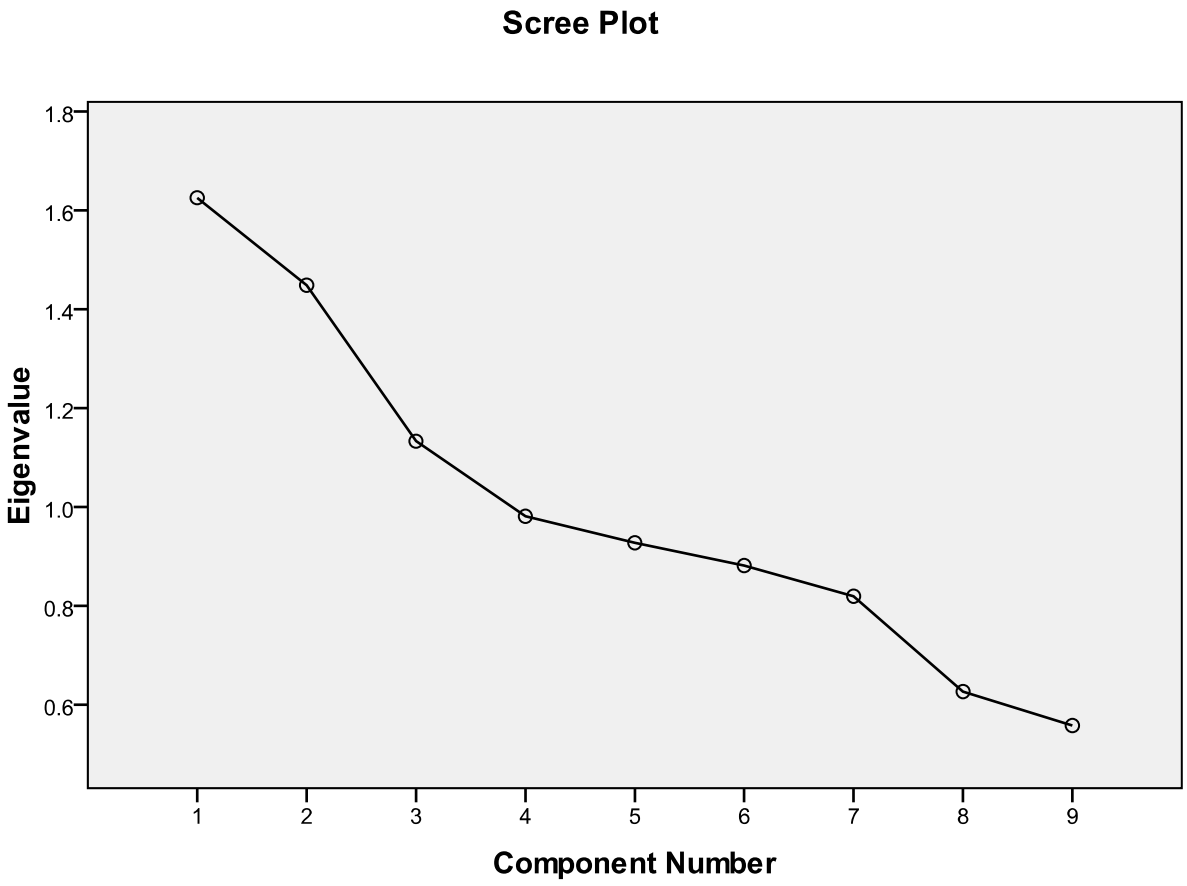
**. Correlation is significant at the 0.01 level (2-tailed).

3. Principal Components of Variables S1 – S7, S9 and S11 for group "L1-C"

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.625	18.060	18.060	1.625	18.060	18.060
2	1.449	16.095	34.155	1.449	16.095	34.155
3	1.133	12.587	46.742	1.133	12.587	46.742
4	.981	10.900	57.643			
5	.927	10.304	67.947			
6	.881	9.793	77.740			
7	.819	9.103	86.843			
8	.626	6.960	93.804			
9	.558	6.196	100.000			

Extraction Method: Principal Component Analysis.



Component Matrix[a]

	Component		
	1	2	3
Z-score ofS1	-.141	.247	.452
Z-score ofS2	.741	.173	.211
Z-score ofS3	.154	-.189	-.578
Z-score ofS4	-.094	.344	-.401
Z-score ofS5	-.434	.549	.255
Z-score ofS6	.550	.286	-.270
Z-score ofS7	.225	-.638	.050
Z-score ofS9	.693	.311	.274
Z-score ofS11	.040	-.564	.417

Extraction Method: Principal Component Analysis.

a. 3 components extracted.

Correlations

		SA_C	FAC1_1	FAC2_1	FAC3_1
SA_C	Pearson Correlation	1	-.838**	.268**	.257**
	Sig. (2-tailed)		.000	.000	.000
	N	323	323	323	323
FAC1_1	Pearson Correlation	-.838**	1	.000	.000
	Sig. (2-tailed)	.000		1.000	1.000
	N	323	323	323	323
FAC2_1	Pearson Correlation	.268**	.000	1	.000
	Sig. (2-tailed)	.000	1.000		1.000
	N	323	323	323	323
FAC3_1	Pearson Correlation	.257**	.000	.000	1
	Sig. (2-tailed)	.000	1.000	1.000	
	N	323	323	323	323

**. Correlation is significant at the 0.01 level (2-tailed).

4. Selected Regression Analysis for City A and Category "L1-A"

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-1.043	.156		-6.685	.000	-1.353	-.734
	Q1	.010	.002	.361	4.633	.000	.006	.014
	Q4	.072	.021	.307	3.485	.001	.031	.113
	Q6	.104	.030	.307	3.511	.001	.045	.163

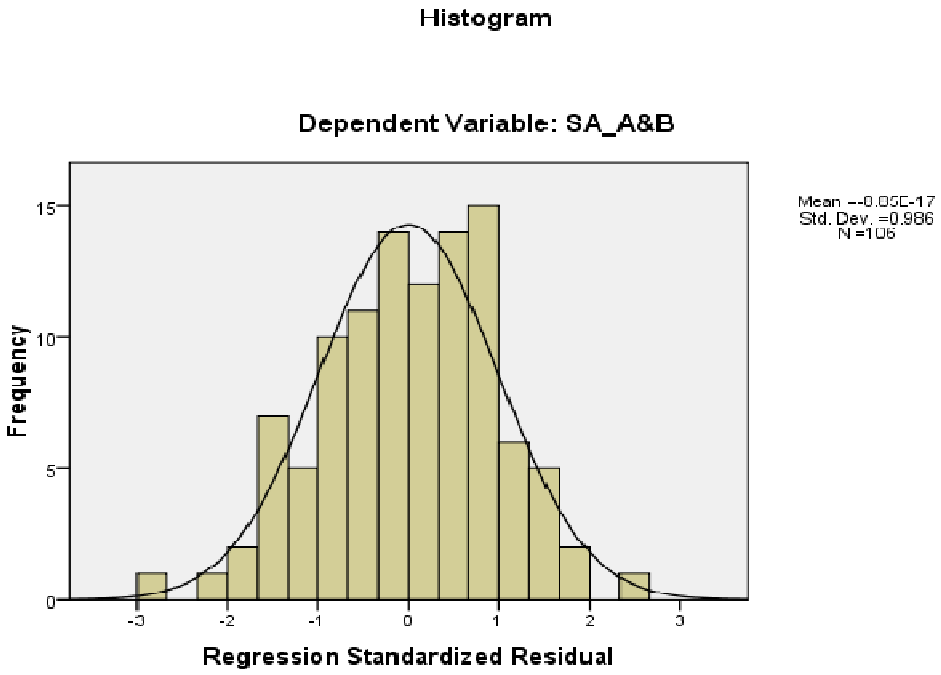
a. Dependent Variable: SA_A&B

One-Sample Kolmogorov-Smirnov Test

		SRE_1
N		106
Normal Parameters[a,,b]	Mean	.0005523
	Std. Deviation	1.00422805
Most Extreme Differences	Absolute	.076
	Positive	.040
	Negative	-.076
Kolmogorov-Smirnov Z		.787
Asymp. Sig. (2-tailed)		.566

a. Test distribution is Normal.

b. Calculated from data.



5. Selected Regression Analysis for City A and Category "L1-B"

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-.510	.144		-3.540	.001	-.795	-.225
	Q1	.008	.002	.316	3.426	.001	.003	.012
	Q6	.062	.020	.279	3.022	.003	.021	.102

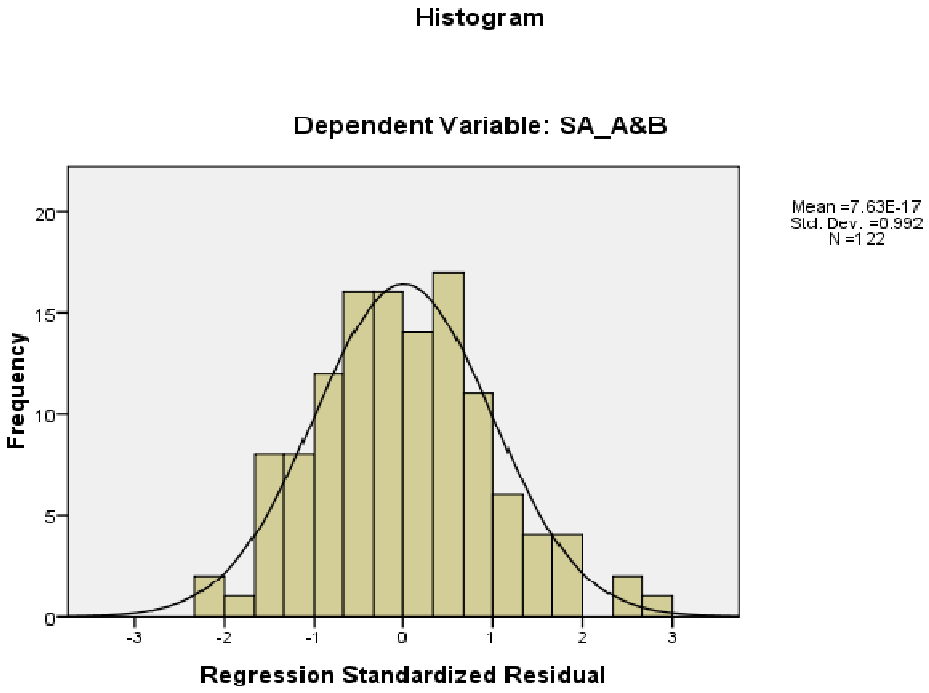
a. Dependent Variable: SA_A&B

One-Sample Kolmogorov-Smirnov Test

		SRE_1
N		122
Normal Parameters[a,,b]	Mean	-.0007170
	Std. Deviation	1.00415262
Most Extreme Differences	Absolute	.043
	Positive	.043
	Negative	-.027
Kolmogorov-Smirnov Z		.471
Asymp. Sig. (2-tailed)		.980

a. Test distribution is Normal.

b. Calculated from data.



6. Selected Regression Analysis for City A and Category "L1-C"

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-.232	.044		-5.263	.000	-.319	-.145
	Q6	.031	.014	.155	2.305	.022	.005	.058

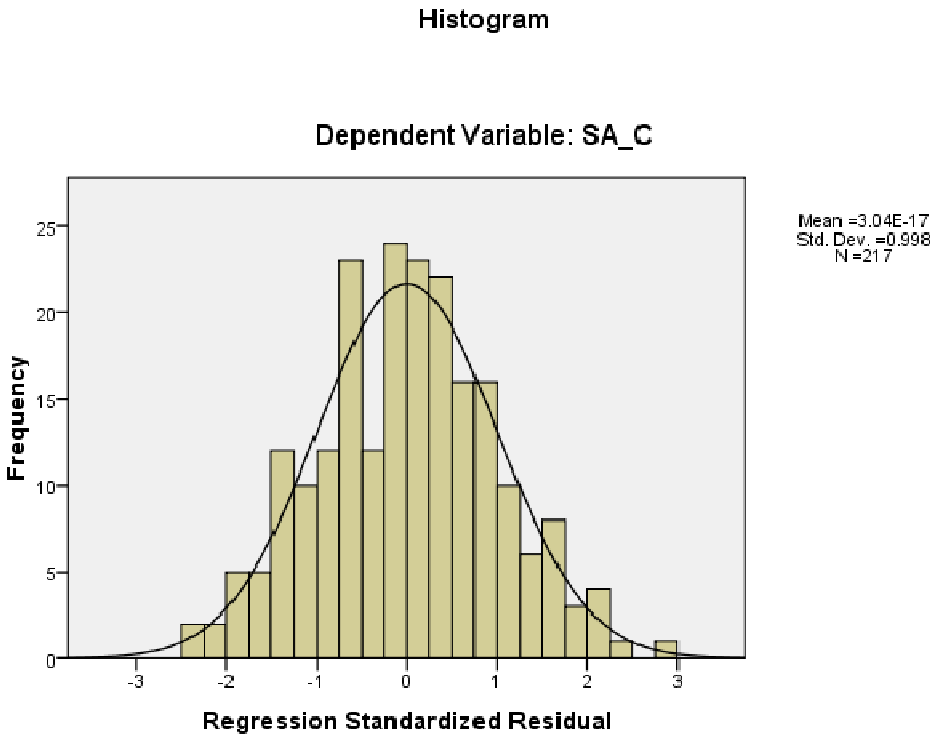
a. Dependent Variable: SA_C

One-Sample Kolmogorov-Smirnov Test

		SRE_1
N		217
Normal Parameters[a,,b]	Mean	.0001145
	Std. Deviation	1.00227644
Most Extreme Differences	Absolute	.035
	Positive	.027
	Negative	-.035
Kolmogorov-Smirnov Z		.508
Asymp. Sig. (2-tailed)		.958

a. Test distribution is Normal.

b. Calculated from data.



7. Selected Regression Analysis for City B and Category "L1-A"

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-.607	.189		-3.206	.002	-.985	-.229
	Q5	.237	.113	.233	2.094	.040	.011	.463
	Q6	.106	.035	.335	3.010	.004	.036	.177

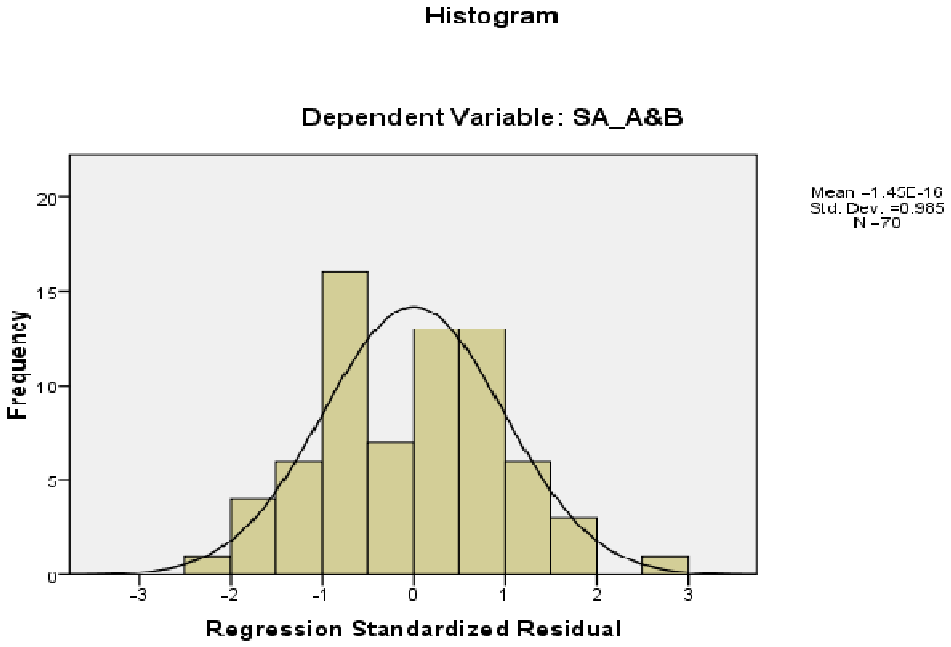
a. Dependent Variable: SA_A&B

One-Sample Kolmogorov-Smirnov Test

		SRE_1
N		70
Normal Parameters[a,,b]	Mean	.0013124
	Std. Deviation	1.00594894
Most Extreme Differences	Absolute	.080
	Positive	.080
	Negative	-.077
Kolmogorov-Smirnov Z		.670
Asymp. Sig. (2-tailed)		.760

a. Test distribution is Normal.

b. Calculated from data.



8. Selected Regression Analysis for City B and Category "L1-B"

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-.874	.182		-4.797	.000	-1.234	-.514
	Q1	.009	.002	.342	4.013	.000	.004	.013
	Q3	.102	.043	.223	2.344	.020	.016	.187
	Q4	.044	.019	.217	2.280	.024	.006	.082
	Q6	.044	.020	.194	2.173	.031	.004	.084

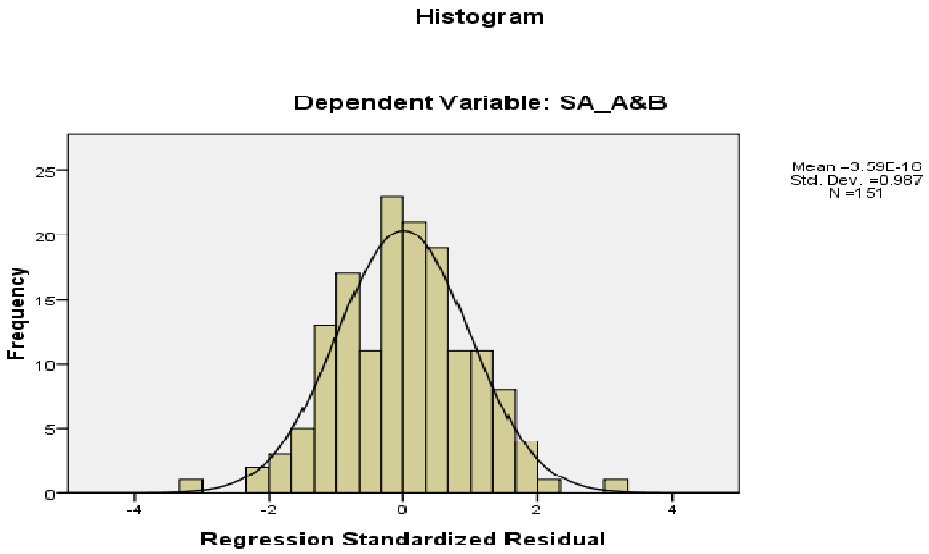
a. Dependent Variable: SA_A&B

One-Sample Kolmogorov-Smirnov Test

		SRE_1
N		151
Normal Parameters[a,,b]	Mean	-.0001946
	Std. Deviation	1.00319570
Most Extreme Differences	Absolute	.043
	Positive	.043
	Negative	-.040
Kolmogorov-Smirnov Z		.533
Asymp. Sig. (2-tailed)		.939

a. Test distribution is Normal.

b. Calculated from data.



9. Selected Regression Analysis for City B and Category "L1-C"

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-.305	.086		-3.528	.001	-.476	-.133
	Q6	.056	.019	.280	2.976	.004	.019	.093

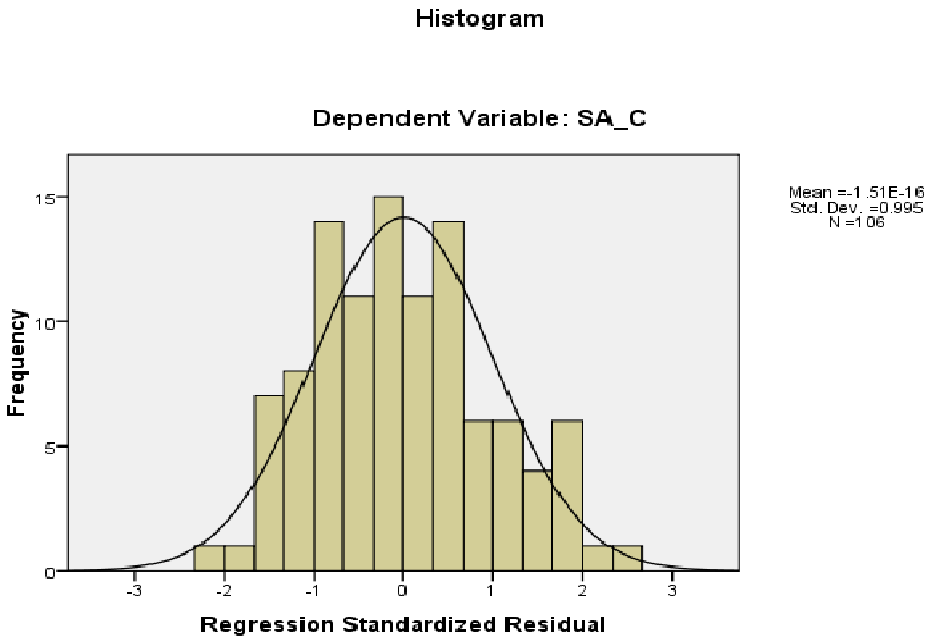
a. Dependent Variable: SA_C

One-Sample Kolmogorov-Smirnov Test

		SRE_1
N		106
Normal Parameters[a,,b]	Mean	.0003392
	Std. Deviation	1.00490017
Most Extreme Differences	Absolute	.057
	Positive	.057
	Negative	-.051
Kolmogorov-Smirnov Z		.585
Asymp. Sig. (2-tailed)		.884

a. Test distribution is Normal.

b. Calculated from data.



10. One-way ANOVA for Category “L1-A” when C1 is the factor

ANOVA

SA_A&B

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.681	3	.227	.971	.407
Within Groups	49.571	212	.234		
Total	50.253	215			

11. One-way ANOVA for Category “L1-B” when C2 is the factor

ANOVA

SA_A&B

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2.388	4	.597	3.362	.010
Within Groups	55.401	312	.178		
Total	57.788	316			

12. One-way ANOVA for Category “L1-C” when C2 is the factor

ANOVA

SA_C

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.312	4	.078	.581	.677
Within Groups	42.657	318	.134		
Total	42.969	322			
