Module #1A: SPSS

Shakeeb Tahir

500837388

Dr. Brian Ceh

SA8903 – Applied Spatial Statistics

Regression Model Changes:

Variable	Unrated	Rated	Rated + Log10
R^2	0.621	0.638	0.644
Adjusted R^2	0.602	0.620	0.627
Durbin Watson	1.799	1.765	1.747

Table 1: Model Comparison

Final Model: Rated + Log10

Descriptives

Descriptives

Descriptive Statistics

	N Statistic	Minimum Statistic	Maximum Statistic	Mean Statistic	Std. Deviation Statistic	Skewness Statistic
AverageFamilyIncomeLog	106	4.77	5.26	4.9430	.09544	.979
Unemployment rate	106	3.6	10.5	6.547	1.4528	.412
UniversityDegreeOrCertificateMastersDegreeRatedAndLog	106	.84	2.04	1.4001	.29348	.243
FinanceAndInsuranceRatedAndLog	106	.91	1.76	1.3358	.24220	.107
ArtsEntertainmentAndRecreationRatedAndLog	106	.66	1.66	1.0238	.17823	.645
AgricultureForestFishingAndHuntingRatedAndLog	106	-.41	1.82	.5934	.61715	.279
Valid N (listwise)	106					

Descriptive Statistics

	Skewness	Kurtosis	
	Std. Error	Statistic	Std. Error
AverageFamilyIncomeLog	.235	1.020	.465
Unemployment rate	.235	-.240	.465
UniversityDegreeOrCertificateMastersDegreeRatedAndLog	.235	-.996	.465
FinanceAndInsuranceRatedAndLog	.235	-1.268	.465
ArtsEntertainmentAndRecreationRatedAndLog	.235	.938	.465
AgricultureForestFishingAndHuntingRatedAndLog	.235	-1.206	.465
Valid N (listwise)			

Figure 1: Final Model (Rated + Log10) Descriptive Statistics

Through our descriptives tables, we can see that our final model has reached a moderately acceptable level of skewness and kurtosis values. All of the values for the skewness statistic are not less than -1 or above 1 and the kurtosis statistics are between +2 and -2. Previously, in the rated model we had skewness above +1 for all the variables except unemployment rate and finance & insurance.

For kurtosis, the rated model had all variables above +2 except for unemployment rate, university degree, and finance & insurance. This is why all variables except unemployment rate were logged, since we needed to normalize the data and solve for the skewness and kurtosis. The reason finance & insurance was logged as well and not unemployment rate is because finance & insurance was still moderately skewed as it had a skewness statistic of 0.683 in the rated model which is between 1 and 0.5 while unemployment rate had a skewness statistic of 0.412 which is between 0 and 0.5. This was also confirmed by testing all the variables by dividing all their skewness values by the standard error, and the only one which was less than 1.96 was the unemployment rate.

All of the variables in the final model as positive which indicate a right-skewed distribution. This is important since it signifies that most values are clustered around the left tail of the distribution and indicates that the mean, median, and mode of the values will be on the lower end. Finally, although we have kurtosis values still above 1 which indicate slight leptokurtic distribution, there is not much more we can do to correct the data as it has already been rated and logged.

For some context of why skewness and kurtosis is important, what the skewness value refers to is how skewed the data (left or right) while kurtosis measures whether the data is peaked or flat. The closer our skewness statistic is to 0, the more normally distributed the data is and the closer the kurtosis is to 0, the more of a mesokurtic distribution we have in our data which is what we what want as it the shape of a normal distribution.

Correlations

Correlations

		AverageFamilyIncomeLog	Unemployment rate	UniversityDegreeOrCertificateMastersDegreeRatedAndLog
Pearson Correlation	AverageFamilyIncomeLog	1.000	-.449	.617
	Unemployment rate	-.449	1.000	.031
	UniversityDegreeOrCertificateMastersDegreeRatedAndLog	.617	.031	1.000
	FinanceAndInsuranceRatedAndLog	.592	-.059	.706
	ArtsEntertainmentAndRecreationRatedAndLog	.260	-.225	.215
	AgricultureForestFishingAndHuntingRatedAndLog	-.275	-.382	-.715
Sig. (1-tailed)	AverageFamilyIncomeLog	.	<.001	<.001
	Unemployment rate	.000	.	.375
	UniversityDegreeOrCertificateMastersDegreeRatedAndLog	.000	.375	.
	FinanceAndInsuranceRatedAndLog	.000	.275	.000
	ArtsEntertainmentAndRecreationRatedAndLog	.004	.010	.013
	AgricultureForestFishingAndHuntingRatedAndLog	.002	.000	.000
N	AverageFamilyIncomeLog	106	106	106
	Unemployment rate	106	106	106

Figure 2: Pearson Correlation Table

The Pearson correlation tells us the strength and direction of the relationship between the variables in the regression model. In figure 2 we can see that there is a moderate negative correlation of -0.449 between unemployment rate and average family income and a slightly positive correlation of 0.617 between university degree and average family income. These both are below 0.7 and greater than -0.7 which indicates there is currently no multicollinearity between any variables.

Correlations

		FinanceAndInsuranceRatedAndLog	ArtsEntertainmentAndRecreationRatedAndLog	AgricultureForestFishingAndHuntingRatedAndLog
Pearson Correlation	AverageFamilyIncomeLog	.592	.260	-.275
	Unemployment rate	-.059	-.225	-.382
	UniversityDegreeOrCertificateMastersDegreeRatedAndLog	.706	.215	-.715
	FinanceAndInsuranceRatedAndLog	1.000	-.054	-.711
	ArtsEntertainmentAndRecreationRatedAndLog	-.054	1.000	.094
	AgricultureForestFishingAndHuntingRatedAndLog	-.711	.094	1.000
Sig. (1-tailed)	AverageFamilyIncomeLog	<.001	.004	.002
	Unemployment rate	.275	.010	.000
	UniversityDegreeOrCertificateMastersDegreeRatedAndLog	.000	.013	.000
	FinanceAndInsuranceRatedAndLog	.	.291	.000
	ArtsEntertainmentAndRecreationRatedAndLog	.291	.	.169
	AgricultureForestFishingAndHuntingRatedAndLog	.000	.169	.
N	AverageFamilyIncomeLog	106	106	106
	Unemployment rate	106	106	106

Figure 3: Pearson Correlation Table Part 2

In figure 3 we can see that there is a moderate positive correlation of 0.592 between finance & insurance and average family income, a very slight positive correlation between arts & recreation and average family income, and a very slight negative correlation of -0.275 between agriculture/forestry/etc and average family income. The Pearson correlation is below 0.7 and greater than -0.7 for all the relationships which indicates there is currently no multicollinearity in the model so far.

Correlations

		AverageFamilyIncomeLog	Unemployment rate	UniversityDegreeOrCertificateMastersDegreeRatedAndLog
	UniversityDegreeOrCertificateMastersDegreeRatedAndLog	106	106	106
	FinanceAndInsuranceRatedAndLog	106	106	106
	ArtsEntertainmentAndRecreationRatedAndLog	106	106	106
	AgricultureForestFishingAndHuntingRatedAndLog	106	106	106

Correlations

		FinanceAndInsuranceRatedAndLog	ArtsEntertainmentAndRecreationRatedAndLog	AgricultureForestFishingAndHuntingRatedAndLog
	UniversityDegreeOrCertificateMastersDegreeRatedAndLog	106	106	106
	FinanceAndInsuranceRatedAndLog	106	106	106
	ArtsEntertainmentAndRecreationRatedAndLog	106	106	106
	AgricultureForestFishingAndHuntingRatedAndLog	106	106	106

Figure 4: Pearson Correlation Table Part 3

Figure 4 shows the number of data points in each relationship.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	4.614	.098		47.148	<.001
	Unemployment rate	-.023	.005	-.349	-4.492	<.001
	UniversityDegreeOrCertificateMastersDegreeRatedAndLog	.161	.034	.494	4.733	<.001
	FinanceAndInsuranceRatedAndLog	.145	.041	.368	3.514	<.001
	ArtsEntertainmentAndRecreationRatedAndLog	.041	.036	.076	1.131	.261
	AgricultureForestFishingAndHuntingRatedAndLog	.031	.019	.200	1.651	.102

Coefficients[a]

Model		Correlations			Collinearity Statistics	
		Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)					
	Unemployment rate	-.449	-.410	-.268	.589	1.697
	UniversityDegreeOrCertificateMastersDegreeRatedAndLog	.617	.428	.282	.326	3.065
	FinanceAndInsuranceRatedAndLog	.592	.332	.210	.324	3.089
	ArtsEntertainmentAndRecreationRatedAndLog	.260	.112	.067	.784	1.275
	AgricultureForestFishingAndHuntingRatedAndLog	-.275	.163	.098	.243	4.115

a. Dependent Variable: AverageFamilyIncomeLog

Figure 5: Significance, VIF & Tolerance Table

In figure 5 we can see that all of our variables except arts & entertainment and agriculture/forestry/etc. are significant since they are less than 0.05 while the other two are above 0.05. This is important because what it means is that there is a less than 5% probability that our linear regression between those variables and average family income was due to chance. The variables above 0.05 indicate that any relationship to average family income was likely due to chance rather than through a true relation to the variable. In the unrated model, only university degree and unemployment rate were significant, and in the rated model the significance for the variable was the same as the log model but was even higher for arts & entertainment (0.843) and agriculture/forestry/etc. (0.663).

For our VIF statistic, all the variables are below 10 and for the tolerance all variables are above 0.2. These thresholds are important since it means that there is no multicollinearity in our model. If the variables did not meet these thresholds, it would indicate that there is multicollinearity in the model and there are multiple variables that are likely measuring the same thing or are too similar in nature. However, since all the variables meet both the VIF and tolerance thresholds, it means there is no multicollinearity in the model. In both the unrated and rated models, the variables met thresholds as well. Since the pairwise correlation values were also less than 0.7 or greater than -0.7, we can now confirm that there is no MCL (multi collinearity) in our model.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics		
					R Square Change	F Change	df1
1	.803[a]	.644	.627	.05833	.644	36.229	5

Model Summary[b]

Model	Change Statistics		Durbin-Watson
	df2	Sig. F Change	
1	100	<.001	1.747

a. Predictors: (Constant), AgricultureForestFishingAndHuntingRatedAndLog, ArtsEntertainmentAndRecreationRatedAndLog, Unem
UniversityDegreeOrCertificateMastersDegreeRatedAndLog, FinanceAndInsuranceRatedAndLog

b. Dependent Variable: AverageFamilyIncomeLog

Figure 6: Model Summary

In figure 6, we can see that our R square value is 0.644 which indicates that our model is moderately accurate at predicting the variance in the logged average family income. About 64.4% of the variance is able to be predicted by our model which is decent and having too high of a value would likely be inaccurate and would indicate that we chose the wrong variables or that there is a problem in our data. This R^2 and adjusted R^2 values are higher than the unrated and rated models, which indicates that transforming our data to make it normalized and fixing any skewness/kurtosis problems improved the accuracy of our model.

The Durbin-Watson value is 1.747 which is between the threshold of 1.6 and 2.4. This means that the residuals are independent from one another and there is no autocorrelation present. If the value was below or above the threshold it would have indicated that there is

autocorrelation within our residuals which means that they depend on each other and the values are influenced by one another rather than being independent. This would be a problem since it would mean that our model would not be of much use since the values are based on related objects rather than through our model being able to predict the values.

Casewise Diagnostics[a]

Case Number	Std. Residual	AverageFamilyIncomeLog	Predicted Value	Residual
1	.741	5.00	4.9602	.04322
2	1.479	4.83	4.7421	.08624
3	-.018	5.06	5.0598	-.00104
4	.325	4.91	4.8942	.01896
5	-.933	4.96	5.0191	-.05440
6	-.149	4.91	4.9160	-.00870
7	.808	4.98	4.9288	.04712
8	-.075	4.93	4.9356	-.00440
9	.011	4.90	4.8947	.00062
10	.004	5.01	5.0138	.00025
11	.307	4.94	4.9191	.01789
12	.532	5.06	5.0245	.03103
13	.081	4.88	4.8705	.00474
14	.936	4.99	4.9319	.05457
15	-1.952	4.83	4.9438	-.11387
16	-1.410	4.88	4.9639	-.08224
17	3.702	5.26	5.0480	.21593
18	.962	5.02	4.9638	.05610
19	2.658	5.21	5.0568	.15502
20	-.214	4.91	4.9186	-.01248
21	.656	4.99	4.9504	.03824
22	1.187	5.07	5.0020	.06921
23	-.172	5.02	5.0349	-.01005
24	-.584	4.85	4.8866	-.03405
25	-.984	4.91	4.9652	-.05739
26	-1.154	4.84	4.9095	-.06728
27	-.738	4.95	4.9948	-.04307
28	.277	4.88	4.8671	.01616
29	-.123	4.87	4.8779	-.00720
30	-.019	5.08	5.0802	-.00111
31	-1.137	4.81	4.8720	-.06634
32	.040	4.87	4.8629	.00235
33	.195	4.90	4.8873	.01135
34	-.267	4.90	4.9178	-.01558
35	1.458	4.86	4.7712	.08505
36	-.389	4.92	4.9428	-.02270
37	-1.020	4.90	4.9590	-.05949

Figure 7: Casewise Diagnostics Part 1

Casewise Diagnostics[a]

Case Number	Std. Residual	AverageFamilyIncomeLog	Predicted Value	Residual
38	-.550	4.97	5.0057	-.03207
39	-.713	5.01	5.0488	-.04158
40	-.518	4.87	4.9040	-.03019
41	-.678	4.87	4.9112	-.03954
42	-.604	4.86	4.8941	-.03523
43	-.204	4.97	4.9804	-.01189
44	-1.103	4.94	5.0054	-.06435
45	.267	5.00	4.9836	.01554
46	-.830	4.89	4.9409	-.04842
47	-.375	4.96	4.9851	-.02189
48	-1.844	4.85	4.9594	-.10754
49	.307	5.02	5.0004	.01789
50	.854	5.08	5.0253	.04979
51	-.204	5.00	5.0167	-.01191
52	.080	5.04	5.0378	.00466
53	1.030	5.07	5.0120	.06005
54	-1.227	4.87	4.9438	-.07157
55	-.301	4.97	4.9893	-.01758
56	1.185	4.89	4.8216	.06910
57	.181	4.85	4.8392	.01057
58	.035	4.87	4.8677	.00204
59	.054	5.07	5.0678	.00317
60	2.012	5.17	5.0526	.11735
61	1.188	4.90	4.8257	.06931
62	-.142	5.02	5.0291	-.00829
63	.234	5.03	5.0189	.01366
64	-.092	4.95	4.9561	-.00539
65	-.058	4.97	4.9693	-.00339
66	-.086	4.94	4.9423	-.00502
67	-.646	4.90	4.9342	-.03770
68	-.270	5.02	5.0386	-.01577
69	-.434	4.87	4.8916	-.02529
70	-1.584	4.89	4.9845	-.09238
71	-.260	4.88	4.8908	-.01514
72	.686	5.02	4.9829	.04000
73	-1.002	4.84	4.8956	-.05844
74	-.083	4.84	4.8459	-.00485

Figure 8: Casewise Diagnostics Part 2

Casewise Diagnostics[a]

Case Number	Std. Residual	AverageFamilyIncomeLog	Predicted Value	Residual
75	-.413	5.02	5.0449	-.02407
76	-1.053	4.88	4.9369	-.06143
77	2.284	5.22	5.0890	.13323
78	.683	4.93	4.8948	.03986
79	.716	4.87	4.8296	.04174
80	-1.303	4.84	4.9150	-.07598
81	-1.291	4.82	4.8947	-.07529
82	-.610	4.83	4.8691	-.03557
83	-.683	4.85	4.8918	-.03985
84	-.915	4.86	4.9112	-.05339
85	.683	4.95	4.9144	.03984
86	-.382	4.88	4.8985	-.02228
87	-.742	4.84	4.8808	-.04326
88	1.271	4.91	4.8333	.07415
89	.894	5.08	5.0317	.05213
90	.482	4.88	4.8534	.02810
91	.834	4.90	4.8545	.04866
92	1.788	4.86	4.7579	.10426
93	1.700	5.16	5.0622	.09915
94	-1.436	4.92	5.0083	-.08378
95	-1.526	5.01	5.1004	-.08898
96	.944	5.05	4.9921	.05508
97	-.944	4.87	4.9258	-.05508
98	.548	5.04	5.0114	.03195
99	.665	5.02	4.9770	.03877
100	-1.095	4.97	5.0381	-.06386
101	.807	4.92	4.8738	.04708
102	.619	4.89	4.8522	.03612
103	-.685	4.91	4.9467	-.03993
104	.419	4.95	4.9276	.02444
105	-.977	4.79	4.8509	-.05698
106	-.603	4.77	4.8045	-.03519

a. Dependent Variable: AverageFamilyIncomeLog

Figure 9: Casewise Diagnostics Part 3

We can see that there are outliers in the data such as case number 15 at -1.95 or case number 17 at 3.70. Further inspection in the map would need to be done to find the reason why.

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	4.7421	5.1004	4.9430	.07661	106
Residual	-.11387	.21593	.00000	.05692	106
Std. Predicted Value	-2.622	2.054	.000	1.000	106
Std. Residual	-1.952	3.702	.000	.976	106

a. Dependent Variable: AverageFamilyIncomeLog

Charts

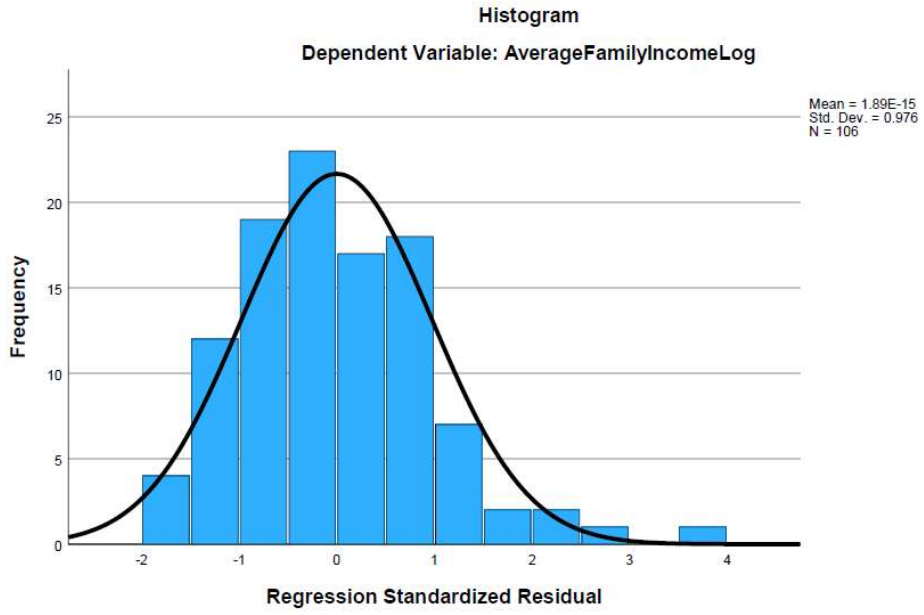

Figure 10: Residual Statistics & Histogram

Our residuals are right skewed which is why we can see that the mean is 1.89E-15. When the chart is right-skewed the mean, median, and mode will be more towards the lower values as they are to the left.

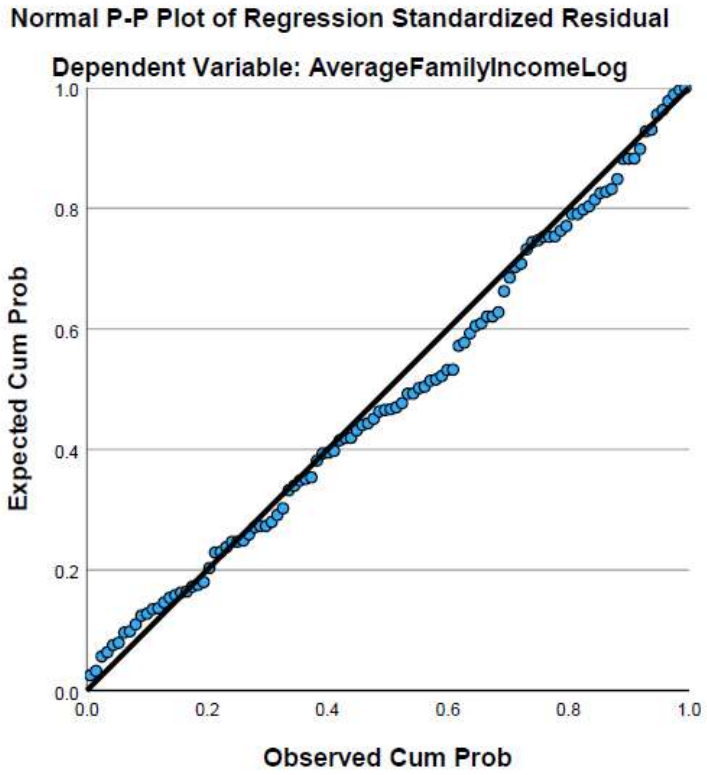

Figure 11: Residual Plot

We can see that most of our residuals are negative since our data was right-skewed which means that the model is underpredicting average family income in most census divisions.

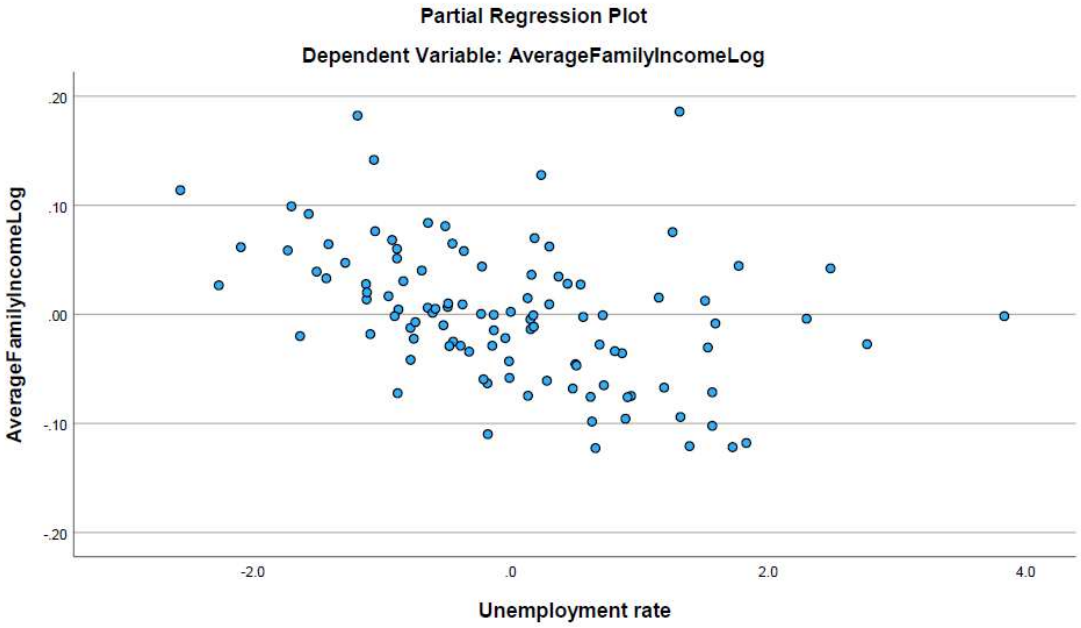

Figure 12: Unemployment Rate Partial Regression Plot

Negative correlation with a few outliers. Homoscedasticity.

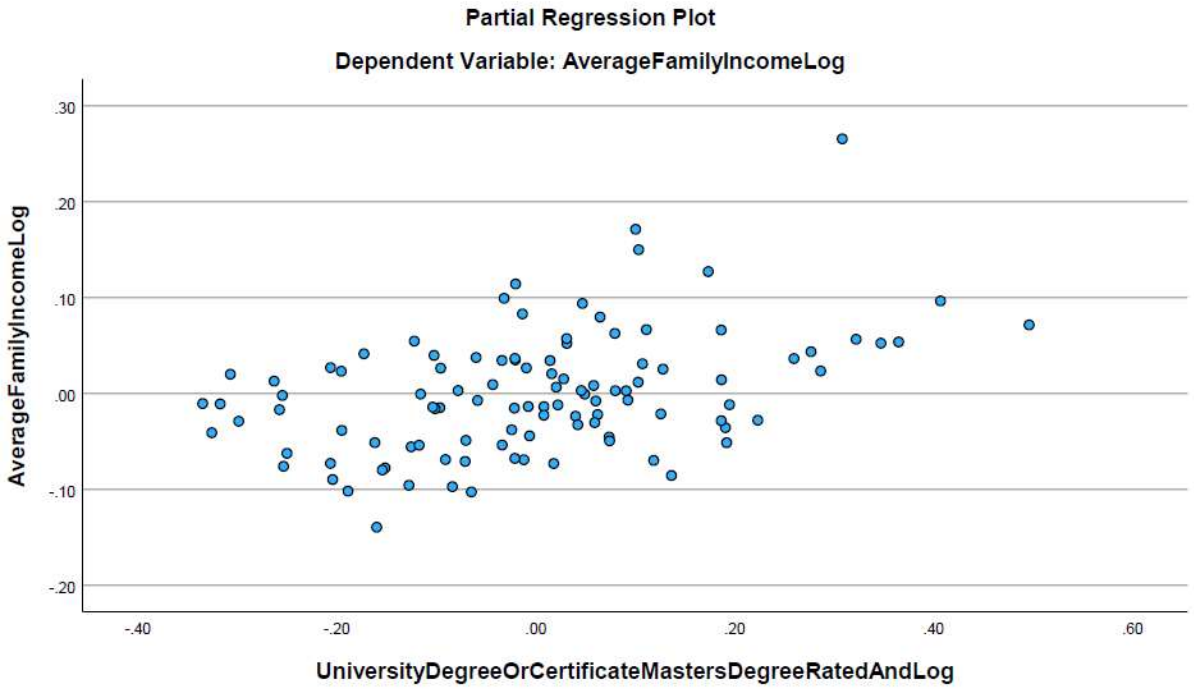

Figure 13: University Degree Partial Regression Plot

Slightly positive correlation with one or two outliers. Homoscedasticity.

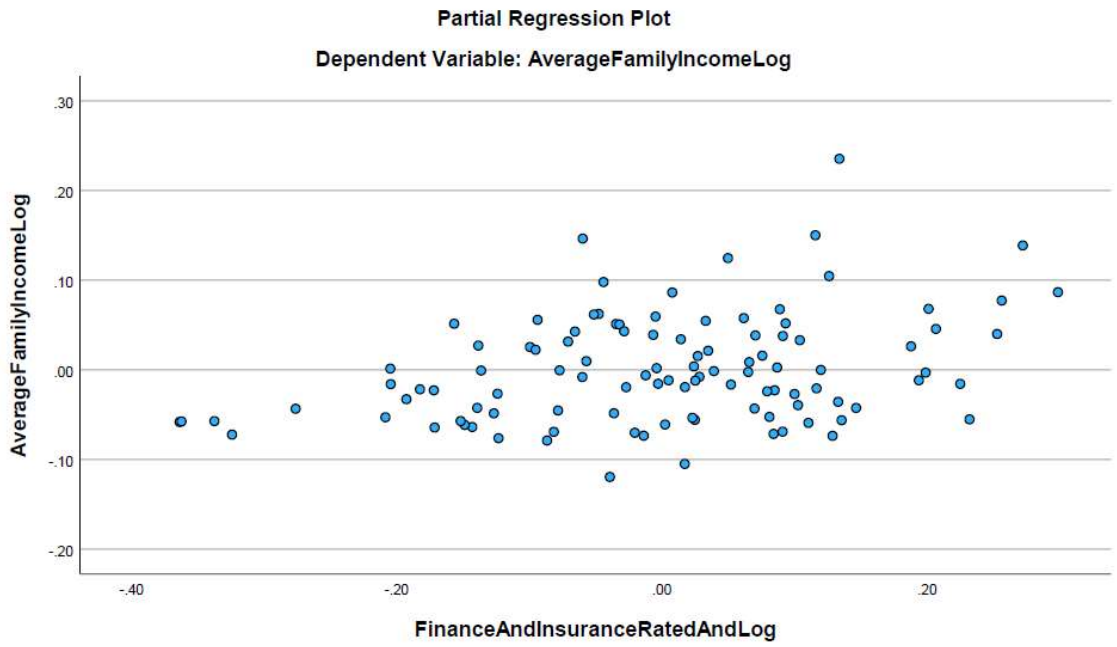

Figure 14: Finance & Insurance Partial Regression Plot

Very slightly positive, almost close to flat-line with a few outliers. Heteroscedasticity.

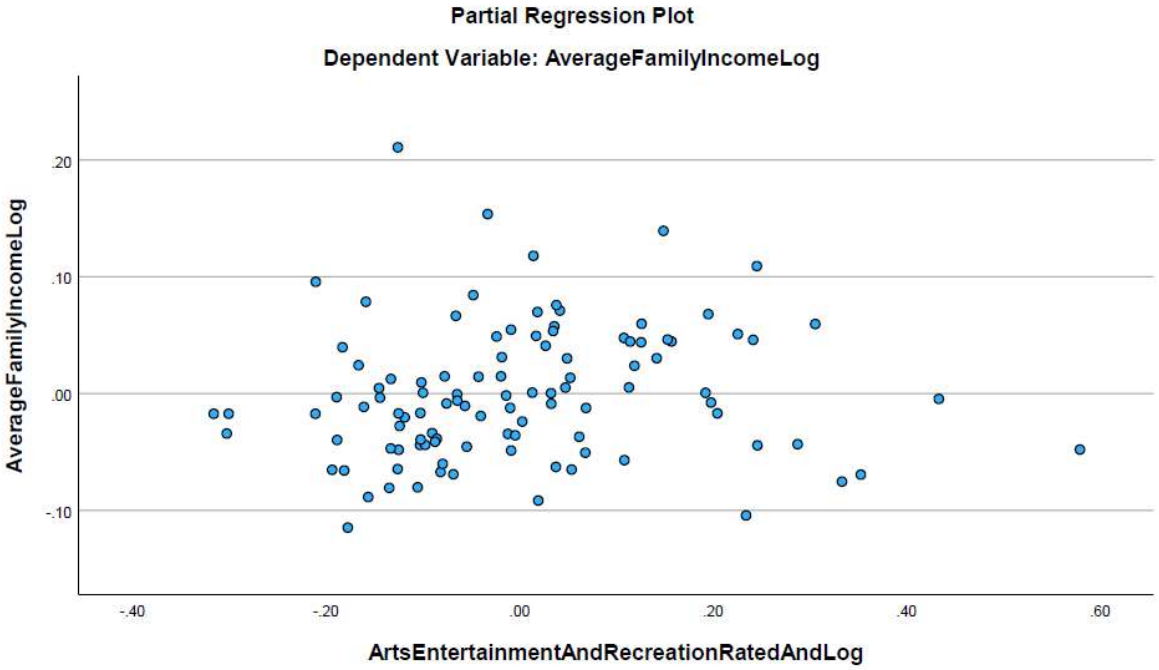

Figure 15: Arts & Entertainment Partial Regression Plot

Flat-Line with no correlation and many outliers. Heteroscedasticity.

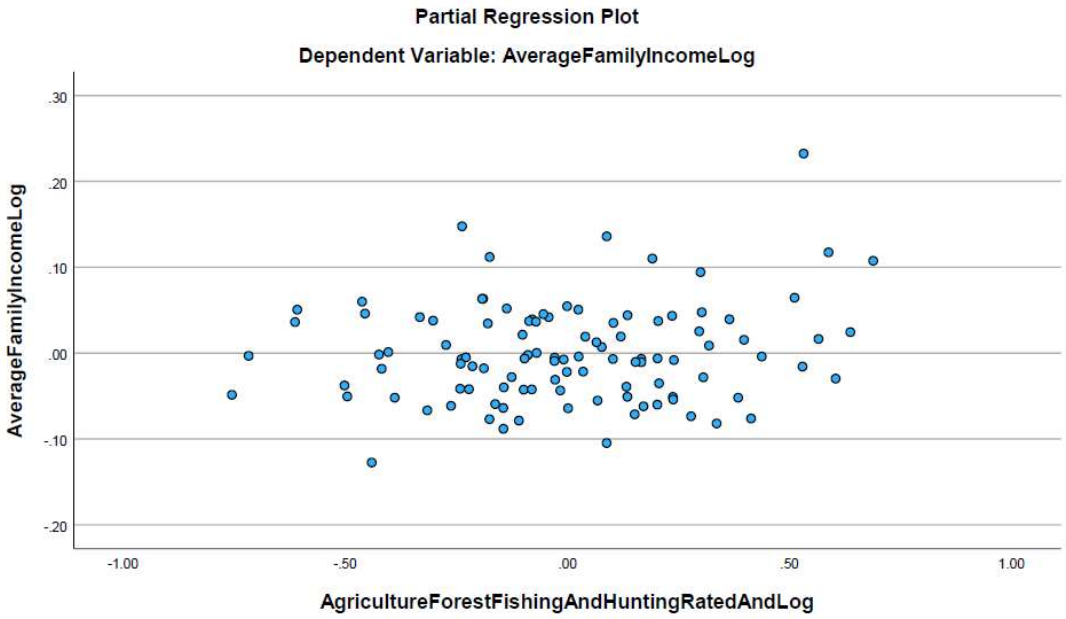

Figure 15: Agriculture/Forest/Fishing/Hunting Partial Regression Plot

Flat-line with no correlation and a few outliers. Heteroscedasticity.