

Advanced Regression Project

## How Much Do Your Neighbors Earn?

Analyzing the Variations of Household Median Income in the Dallas

Collin County Area

Sheheryar Banuri (Data Analysis)

Shaofei Chen (GIS Analysis)

Tianxiao Yang (Documentation and Analysis)

## **Introduction**

We are interested in the factors influencing income in the Dallas-Collin county area in 2005. We choose explanatory and control variables considering demographic, social, economic, and housing profiles of the census tracts. We test these variables and examine the potential problems of the regression. We also compare the regression outcomes in Dallas and Collin counties and analyze residuals with GIS mapping.

## **Section 1. Variable Selections and Expectations**

**Dependent Variable:** T\_HHMEDINC: median household income. Since it is positively skewed, we will perform a log transformation on it.

### **Explanatory Variables:**

- I. demographic profile:
  - i-iv. T\_AMINDIAN, T\_AP\_BLACK, T\_AP\_HISP, T\_AP\_ASIAN: we expect to see negative relationships between T\_HHMEDINC and the first three variables respectively, because minorities tend to have less income than Caucasians. Meanwhile, we expect to see a positive relationship between T\_HHMEDINC and T\_AP\_ASIAN because Asians tend to make more money (in general).
  - v. T\_MED\_AGE: Older individuals tend to have higher incomes on average, so we expect to see positive relationship here. However, the literature suggests an individual's income to peak in their forties, so the relationship might not be linear. An age-squared variable will be tested in our regression.
- II. Social profile:
  - i. T-ED-SCHPE: Married couples with children are willing to pay a higher price for houses with a better school district. We hypothesize that better school districts attract higher income families, so we expect to see a positive relationship between median income and the percentage of school-going children in a census tract.
  - ii. T\_ED\_PSDEG: Since higher degrees are commonly related to higher wages, a higher percentage of individuals with a post-secondary degree would impact median income in a positive fashion.
- III. Economic profile
  - i. T\_LF\_PART: A higher labor force participation rate would imply a higher level of the median income in a given census tract, so we expect to see a positive relationship here as well as a control for unemployment.
  - ii. T\_BPL\_FAM: We expect a negative relationship between the number of families below the poverty line and median income. We will need to test for correlations between this variable and the labor force participation rate in order to avoid the multicollinearity issue.
  - iii. T\_FAM\_AVGS: Large families, especially families with more children tend to have lower incomes. We expect to see negative relationship between size of family and median income. This may be related to the variable for the number of school-going children in a census tract. We would test for possible multicollinearity here.
- IV. Housing profile
  - i. T\_HU\_VALUE: Higher income households are more likely to purchase expensive

houses, so the median owned home value would have a positive relationship with median income.

- ii. T\_HU\_BLT80: Newer houses tend to be more expensive. We suspect wealthier households tend to purchase newer houses: we expect a positive relationship between T\_HU\_BLT80 and median income.

**Control Variable:** T\_MS\_MARRIED: Married couples may be dual-income households. This increases income even when other profiles are similar between two households. We should control for this when running the regression.

**GIS Variable:** DISTANCE: This variable is feet from the centroid of Highland Park to that of each census tract. We hypothesize a distance decay effect in the Dallas-Collin county area, so we expect a negative relationship between distance from Highland Park and median household income.

**Additional Variable:** totalpop: Total population. We assume poor and rich people are evenly distributed. Since median income is positively skewed, rich people take up a smaller percentage of total population compared to poor people. Therefore, the more population in a census tract, the more likely it is that the median income is dragged downwards. We thus expect a negative relationship here.

## Section 2. Variable Adjustments and Regression Models

### I. Variable Transformation and Preliminary Regression Model:

Since T\_HHMEDINC is positively skewed, we log-transform it and name the transformed variable LOG\_INCOME. We run a Shapiro Test on LOG\_INCOME, and the results ( $W = 0.9905$ ,  $p\text{-value} = 0.001218$ ) show that at 0.01 significant level, we can reject the null hypothesis that LOG\_INCOME is not normally distributed. We also perform log-transformation on T\_HU\_VALUE (renamed as LOG\_HomeValue) for the same reason. This should lead to a better regression model.

We decide to drop observations where the median home income or house value equals to zero as exceptions. These observations are 78, 85, 167, 214, 248, 360, 384, 392, 400, 405, 422, 454, 483, 497, and 508.

After above adjustments, the scatterplot matrix (Figure 1) shows that T\_HHMEDIC is approximately normally distributed. We construct regression model 1 using all explanatory and control variables.

Regression outcomes for all models are summarized in table 1. Since regression coefficients of T\_AMINDIAN, T\_AP\_BLACK, and T\_AP\_HISP are not significant at 0.05 significance level, we decide to drop them in regression model 2. In the new model, the coefficient of T\_AP\_ASIAN is negative and significant at 0.001 significance level; adjusted R-squared remains the same. We also run an F-test to test whether T\_AMINDIAN, T\_AP\_BLACK, and T\_AP\_HISP are jointly significant. According to F-test outcome ( $Pr(>F)=0.5013$ ), we cannot reject the null hypothesis that T\_AMINDIAN, T\_AP\_BLACK and T\_AP\_HISP are jointly insignificant at 0.1 significance level. Therefore, we can justifiably drop these variables.

### II. Total population, T\_HU\_BLT80, and AGE\_SQUARED

In regression model 3, we add in the total population variable. The regression outcome shows that it has a negative relationship with the median income as expected and it is

significant at 0.1 significance level.

Since regression coefficients of T\_HU\_BLT80 are insignificant at the 0.1 significance level in the first three regression models, we decide to drop it in regression model 4. The Adjusted R-Squared remains the same. Except for slight changes in the significance levels of T\_LF\_PART and T\_AP\_ASIAN, all other regression coefficients remain significant at their original significance levels. Therefore, dropping T\_HU\_BLT80 makes the new model simpler yet explanatory.

Since we suspect a non-linear relationship between age and median income, we add an “age\_squared” variable in regression model 5. The coefficient of age\_squared is not significant at 0.1 significance level and the Adjusted R-Squared does not increase. The age\_squared variable does not capture the relationship between age and median income properly, and should not be included. The scatterplot between age and median income, and that between age and house value (Figure 2) show that their relationships are nearly linear.

### **III. DISTANCE Variable and Comparison between Dallas Collin Counties**

Next, we add DISTANCE variable in regression model 6. Neither is its coefficient significant at 0.1 significance level nor does the Adjusted R-squared increase. This means median income in the Dallas-Collin county area is not distributed in a “ring-structure”. Therefore, we should drop the DISTANCE variable.

We decide to use regression model 4 from the above experiments. We now compare median income distribution in the Dallas and Collin counties. First, we map the income distribution in Figure 3. In Dallas County, median income in the south is much lower than that in the north. The median income range is wide. On the contrary, the median income in Collin County is more even, although the south-western part is richer than the rest. Then, we compare the regression outcomes in the two counties. In Table 2, adjusted R-squared of Dallas County is much lower than that of Collin County. This may be caused by the wide median income range in Dallas County, the complexity of which is less explained. More evenly distributed income in Collin County may also explain the higher R-squared statistic. Regression models for the two counties have different significant variables. Demographic disparities between them may result in different significant explanatory variables.

## **Section 3. Criticism**

### **I. Multicollinearity**

First, we calculate variance influencing factor (Table 3) and the correlation between independent variables (Table 4) of regression model 4. We combine the two as the criteria to identify highly correlated variables. T\_ED\_PSDEG has the highest VIF (6.971421) value and it has a high correlation of 0.88488685 with Log\_HomeValue. Therefore, we drop T\_ED\_PSDEG in regression model 7. In the regression results, Adjusted R-squared decreases from 0.9374 to 0.9068; coefficients of T\_AP\_ASIAN, T\_FAM\_AVGS, and totalpop become insignificant at 0.1 significance level. We also conduct an F-test between regression model 4 and 7. The F-test result ( $Pr(>F) < 2.2e-16$ ) shows that at near 0 significance level, we can reject the null hypothesis that the coefficient of T\_ED\_PSDEG is zero. According to above results, we suggest leaving T\_ED\_PSDEG in the regression model.

In the correlation matrix, PSDEG variable is highly correlated with Log\_HomeValue, and VIF of Log\_HomeValue (4.892465) is the second highest. Therefore, we drop

Log\_homeValue in regression model 8. In the new model, coefficient of T\_AP\_ASIAN is significant at near 0 significance level. Coefficients of other variables remain significant at original significance levels. Adjusted R-squared decreases only slightly (from 0.9374 to 0.9338). We conduct an F-test between regression model 4 and 8. The result ( $P_r(>F) = 1.864e-08$ ) suggests that we can reject the null hypothesis that the coefficient of Log\_HomeValue equals to zero. However, to reduce potential multicollinearity problem, we need to drop either T\_ED\_PSDEG or Log\_HomeValue. From above analyses, we decide to drop Log\_HomeValue.

After dropping Log\_HomeValue, the highest VIF (Table 5) is that of T\_FAM\_AVGS (4.129740), and the highest correlation (Table 6) is -0.68672995 between T\_FAM\_AVGS and T\_ED\_PSDEG. They are not large enough to be risky. Correlation between T\_BPL\_FAM and LF\_PART (-0.5708), as well as that between SCHPE and average family size (0.6638) is acceptable.

## II. Heteroskedasticity

We plot studentized residuals against fitted values (Figure 4) in order to find potential heteroskedasticity problems. The dots do not show a specific pattern. We perform the Bruesh Pagan test whose results ( $BP = 106.3668$ ,  $df = 9$ ,  $p\text{-value} < 2.2e-16$ ) indicate that we can reject the null hypothesis that heteroskedasticity exists in our sample at near 0 significance level.

We also mapped the residuals and tested for spatial autocorrelation in the Dallas Collin County area (Figure 5). In the map, similar colored blocks (similar residuals) have a slight tendency to gather. The global moran's I value (0.1211) shows that there is a little spatial autocorrelation among the residuals. Therefore, besides independent variables, spatial factors also explain a little variation in median income.

## III. Beta Weights

To compare the influence of the independent variables on the dependent variable, we calculate beta weights. Beta weights indicate standard deviation change in the dependent variable, per 1-standard-deviation change in this independent variable.

From Table 7, we can see that the beta weight absolute value of T\_ED\_PSDEG (0.5739) is the largest, followed by that of T\_BPL\_FAM (0.3335). They have larger influence on the LOG\_INCOME than other independent variables in terms of standard deviations change per standard deviation. Therefore, 1-standard-deviation change in T\_ED\_PSDEG/ T\_BPL\_FAM leads to 0.5739/ 0.3335-standard deviation change in LOG\_INCOME (other things being equal).

## Section 4. Outlier Analysis

In the residual-quantile comparison plot (Figure 6), most studentized residuals are located within the confidence intervals. We identify census tracts 507, 480, 509, 227, and 485 as influential.

In the influence plot, among census tracts which are located beyond (-2, +2), 507, 506, 480, 227, 509, 485 are represented by larger circles and thus larger Cook distance indicating larger influence on the model as a whole.

Combining the results of both figures, we experiment to drop 485, 227, 480, 509, and 507 census tracts as outliers. We calculate the residuals, standardized residuals, studentized residuals, leverages, and Cook distances of these influential census tracts in Table 8. We

summarize these statistics sample-wide in Table 9. Studentized residuals test whether a case causes a significant shift in the regression intercept and so should be considered an outlier. In our sample, both minimum and maximum of the residual, standardized residual, and studentized residual are included in the dropped census tracts. Leverage (measured by hat value) reflects the potential for influence resulting from unusual X values. Hat values lower than 0.2 are safe. In our sample, all hat values are lower than 0.2 (Figure 8), but census tracts dropped have much higher leverage than sample mean (0.01795). Cook Distance measures a case's influence on the model as a whole. Size-adjusted cutoff is  $4/n$  which equals to 0.0072 in the sample. Dropped census tracts 227, 507, 509 have a higher Cook Distance indicating a high influence. Actually, census tract 507 has the maximal Cook Distance (0.8089484). Most census tracts have low Cook Distance though (Figure 7).

We calculate the DFBETAS of dropped census tracts in table 3.  $DFBETAS_{ik}$  measures the influence case  $i$  exerts on  $b_k$ . The size-adjusted cutoff of DFBETAS is  $\frac{2}{\sqrt{n}}$  for absolute value, which is 0.085 in our sample. The absolute value of DFBETAS of census tract 227 on T\_AP\_ASIAN, T\_ED\_PSDEG, T\_FAM\_AVGS, and T\_MS\_MARRI, as well as that of census tracts 480 on T\_AP\_ASIAN are higher than the standard. DFBETAS of other census tracts are smaller and indicate little case influence on regression coefficients.

We construct regression model 9 after dropping above mentioned five census tracts. All regression coefficients remain significant at their original significance level. Adjusted R-squared decreases slightly. Dropping the census tracts does not influence the regression model a lot. Possible reasons are: 1. the hat values of the census tracts are lower than 0.2, so their leverage is not riskily large. 2. Only three dropped census tracts have Cook Distance higher than size-adjusted cutoff value. All have Cook Distance lower than absolute cutoff value—1. 3. Of fifty DFBETAS calculated (ten variables and five census tracts), only five of them are higher than size-adjusted cutoff value, others do not appear influential. Therefore, we decide to keep them in the regression model, and model 8 is our final decision.

## Conclusion

In regression model 8, coefficients of T\_AP\_ASIAN and T\_FAM\_AVGS have opposite to expected sign. Diligence of Asians may not be enough to lead to higher income. Our assumption on average family size is proved wrong. We drop coefficients of T\_AMINDIAN, T\_AP\_BLACK, T\_AP\_HISP, and HU\_BLT80 because they are not significant at 0.1 significance level. We dropped the log-transformed variable T\_HU\_VALUE because of the potential multicollinearity problem. All other selected variables influence median income in expected directions. Beta weights (Table 7) indicates that post secondary degree has the highest influence on median income, followed by family below poverty line. Beta weight of T\_AP\_ASIAN is very small (-0.049). This partly explains the opposite to expected signal of its coefficient.

## Appendix

### List of Figures:

- Figure 1. Log-Housevalue Scatterplot Matrix of T\_HHMEDINC and Ethnicity Variables
- Figure 2. Scatterplot Matrix between Log\_HomeValue, LOG\_INCOME, and T\_MED\_AGE
- Figure 3. Income Distribution in Dallas-Collin County Area
- Figure 4. Studentized Residuals against Fitted Values in Regression Model 8
- Figure 5. Residual Spatial Distribution
- Figure 6. Residual-Quantile Comparison
- Figure 7. Influence Plot
- Figure 8. Hat Value
- Figure 9. Cook Distance

### List of Tables:

- Table 1. Summary of Regression Models
- Table 2. Comparison between Dallas and Collin Counties
- Table 3. VIF of Regression Model 4
- Table 4. Correlation Matrix of Regression Model 4
- Table 5. VIF of Regression Model 8
- Table 6. Correlation Matrix of Regression Model 8
- Table 7. Beta Weights of Regression Model 8
- Table 8. Outlier Analysis
- Table 9. Influence Analysis Summary
- Table 10. DFBETAS of Influential Tracts

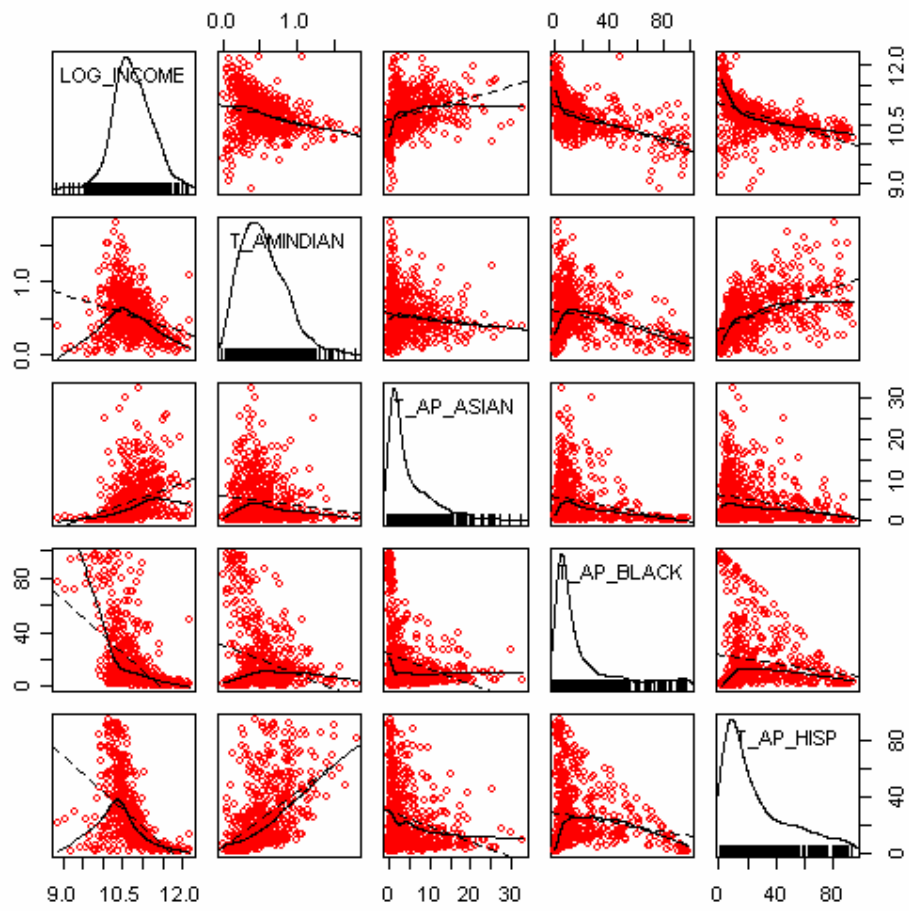


Figure 1. Scatterplot Matrix of T\_HHMEDINC and Ethnicity Variables

		Variations in Median Income in the Dallas/Collin County Area			
Variable		Model 1	Model 2	Model 3	Model 4
Intercept	(Intercept)	7.735***	7.918***	7.973***	7.986***
		(28.597)	(33.264)	(33.321)	(33.444)
Home Value	Log_HomeValue	0.1125***	0.105***	0.1034***	0.1032***
		(5.881)	(5.768)	(5.717)	(5.709)
American Indian (%)	T_AMINDIAN	0.02452			
		(1.082)			
Asian (%)	T_AP_ASIAN	-0.00363**	-0.0037**	-0.003905***	-0.00365**
		(-3.113)	(-3.237)	(-3.413)	(-3.285)
Black (%)	T_AP_BLACK	0.00041			
		(0.922)			
Hispanic (%)	T_AP_HISP	-0.00039			
		(-0.547)			
Poverty Line Families (%)	T_BPL_FAM	-0.0151***	-0.0151***	-0.01531***	-0.0153***
		(-15.172)	(-15.742)	(-15.89)	(-15.878)
Post Secondary Degree (%)	T_ED_PSDEG	0.0098***	0.0098***	0.009779***	0.009875***
		(14.600)	(16.092)	(16)	(16.394)
School Children (%)	T_ED_SCHPE	0.0085***	0.0097***	0.009508***	0.009808***
		(4.494)	(6.519)	(6.396)	(6.757)
Average Family Size (Average)	T_FAM_AVGS	0.1551***	0.13002***	0.1309***	0.1259***
		(4.031)	(5.163)	(5.207)	(5.126)
Labor Force Participation (%)	T_LF_PART	0.0022*	0.002.	0.001884.	0.002043*
		(2.122)	(1.943)	(1.922)	(2.116)
Median Age (Number)	T_MED_AGE	0.00639***	0.0068***	0.006182***	0.005778***
		(3.810)	(4.338)	(3.893)	(3.782)
Married Individuals (%)	T_MS_MARRI	0.01012***	0.0097***	0.009832***	0.009916***
		(11.992)	(12.994)	(13.141)	(13.354)
Homes Built after 1980 (%)	T_HU_BLT80	0.000018	0.0001	0.000226	
		(0.071)	(0.483)	(0.935)	
Tract Total Population (Number)	totalpop			-0.0000055.	-0.00000486.
				(-1.882)	(-1.703)
R-Squared		0.9385	0.9383	0.9387	0.9386
Adjusted R-Squared		0.9371	0.9371	0.9374	0.9374
F-Statistic		637.7	829.8	758.2	834.1
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

To be continued..

	Variations in Median Income in the Dallas/Collin County Area				
Variable	Model 6	Model 5	Model 7	Model 8	Model 9
Intercept	7.947***	7.946***	5.936***	9.072***	9.085***
	(31.854)	(27.258)	(23.91)	(61.12)	(60.92)
Home Value	0.1052***	0.1039***	0.3159***		
	(5.707)	(5.681)	(20.547)		
Asian (%)	-0.003556**	-0.00367**	0.0007768	-0.0044***	-0.004225***
	(-3.163)	(-3.292)	(0.591)	(-3.878)	(-3.605)
Poverty Line Families (%)	-0.01523***	-0.01524***	-0.01426***	-0.01559***	-0.01551***
	(-15.676)	(-15.401)	(-12.152)	(-15.758)	(-15.614)
Post Secondary Degree (%)	0.009924***	0.009873***		0.01234***	0.01229***
	(16.293)	(16.374)		(28.596)	(28.099)
School Children (%)	0.00953***	0.00976***	0.009689***	0.008727***	0.008751***
	(6.2)	(6.658)	(5.47)	(5.896)	(5.882)
Average Family Size (Average)	0.1305***	0.1261***	0.003099	0.1359***	0.1322***
	(5.03)	(5.127)	(0.109)	(5.392)	(5.212)
Labor Force Participation (%)	0.002003*	0.002007*	0.004068***	0.001989*	0.001895.
	(2.068)	(2.054)	(3.481)	(2.003)	(1.9)
Median Age (Number)	0.005957***	0.007808	0.008323***	0.005623***	0.005623***
	(3.813)	(0.922)	(4.488)	(3.58)	(3.558)
Married Individuals (%)	0.009813***	0.009926***	0.01239***	0.01037***	0.01045***
	(12.813)	(13.335)	(13.958)	(13.661)	(13.587)
Tract Total Population (Number)	-0.000005055.	-0.000004884'	-0.00000554	-0.000005436.	-0.000005543.
	(-1.759)	(-1.711)	(-1.593)	(-1.856)	(-1.883)
Tract Distance from Highland Park (feet)	0.0000001331				
	(0.554)				
Median Age Squared (Number)		-0.00002885			
		(-0.244)			
R-Squared	0.9386	0.9386	0.9083	0.9349	0.9345
Adjusted R-Squared	0.9374	0.9373	0.9068	0.9338	0.9334
F-Statistic	757.3	757	602.2	872.7	859.4
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

Table 1. Summary of Regression Models (T-stat's in parentheses)

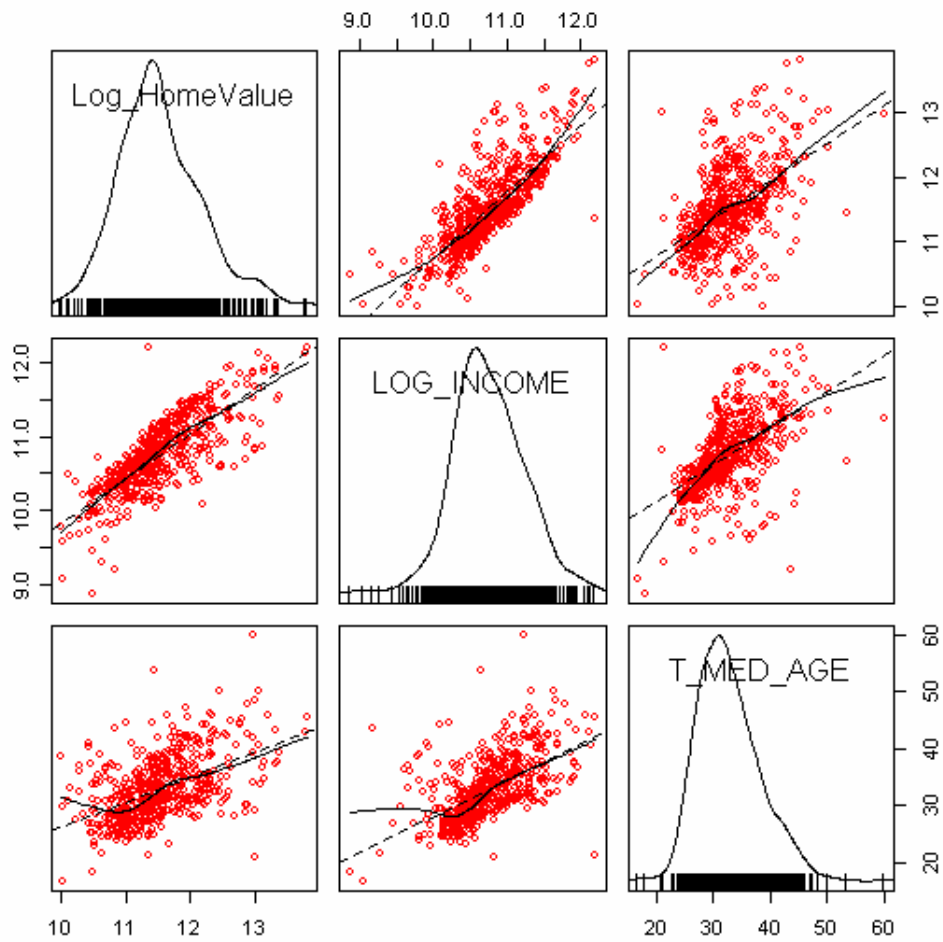


Figure 2. Scatterplot Matrix between Log\_HomeValue, LOG\_INCOME and T\_MED\_AGE

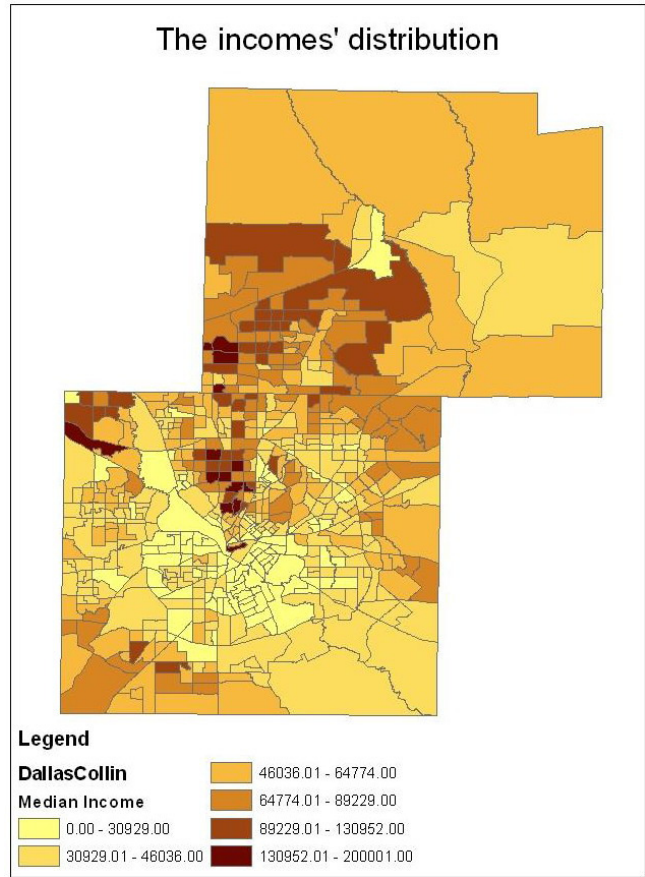


Figure 3. Income Distribution in Dallas-Collin County Area

	Dallas		Collin	
	Coefficients	p-value	Coefficients	p-value
Intercept	2.496	0.00000	5.33900	0.00000
Log_Homevalue	0.06527	0.00000	0.26420	0.00000
T_AP_ASIAN	-0.01237	0.00557	-0.00058	0.81201
T_BPL_FAM	0.01457	0.00001	-0.01300	0.06996
T_ED_PSDEG	0.01378	0.00000	0.00665	0.00005
T_ED_SCHPE	0.02531	0.00000	0.00379	0.55801
T_FAM_AVGS	0.5898	0.00000	0.36230	0.01407
T_LF_PART	0.0447	0.00000	0.00031	0.92388
T_MED_AG	0.0291	0.00000	0.01813	0.00062
T_MS_MARRI	0.0221	0.00000	0.00858	0.00037
totalpop	0.00001937	0.09176	-0.00001	0.31658
R-squared	0.7132		0.93410	
Adjusted R-squared	0.7072		0.92510	

Table 2. Comparison between Dallas and Collin Counties

Log_HomeValue	T_AP_ASIAN	T_BPL_FAM	T_ED_PSDEG	T_ED_SCHPE
4.892465	1.381475	3.774716	<b>6.971421</b>	2.571485
T_FAM_AVGS	T_LF_PART	T_MED_AGE	T_MS_MARRI	totalpop
4.150685	3.367637	2.64031	3.495202	1.184855

Table 3. VIF of Regression Model 4

	Log_HomeValue	T_AP_ASIAN	T_BPL_FAM	T_ED_PSDEG	T_ED_SCHPE
Log_HomeValue	1				
T_AP_ASIAN	0.31181533	1			
T_BPL_FAM	-0.61732251	-0.30056272	1		
T_ED_PSDEG	<b>0.88488685</b>	0.40286977	-0.6469067	1	
T_ED_SCHPE	-0.39293203	-0.16947922	0.2618053	-0.43512181	1
T_FAM_AVGS	-0.58358624	-0.23315444	0.5024772	-0.68672995	0.6638074
T_LF_PART	0.52551726	0.39930173	-0.5707648	0.6024009	-0.6163465
T_MED_AGE	0.50529242	0.011878	-0.5216322	0.55030178	-0.292623
T_MS_MARRI	0.43898852	0.21304903	-0.6277209	0.42758459	0.2750185
totalpop	-0.02177808	0.06769891	-0.1300915	-0.02447148	0.0779205
	T_FAM_AVGS	T_LF_PART	T_MED_AGE	T_MS_MARRI	totalpop
Log_HomeValue					
T_AP_ASIAN					
T_BPL_FAM					
T_ED_PSDEG					
T_ED_SCHPE					
T_FAM_AVGS	1				
T_LF_PART	-0.671434759	1			
T_MED_AGE	-0.583149907	0.21181777	1		
T_MS_MARRI	-0.007175136	0.06603202	0.34601887	1	
totalpop	0.091705154	0.08707065	-0.22979151	0.153740614	1

Table 4. Correlation Matrix of Regression Model 4

T_AP_ASIAN	T_BPL_FAM	T_ED_PSDEG	T_ED_SCHPE	T_FAM_AVGS
1.362144	3.763907	3.3836	2.527728	<b>4.12974</b>
T_LF_PART	T_MED_AGE	T_MS_MARRI	totalpop	
3.367312	2.63948	3.45474	1.183343	

Table 5. VIF of Regression Model 8

	T_AP_ASIAN	T_BPL_FAM	T_ED_PSDEG	T_ED_SCHPE	T_FAM_AVGS
T_AP_ASIAN	1				
T_BPL_FAM	-0.30056272	1			
T_ED_PSDEG	0.40286977	-0.6469067	1		
T_ED_SCHPE	-0.16947922	0.2618053	-0.43512181	1	
T_FAM_AVGS	-0.23315444	0.5024772	<b>-0.68672995</b>	0.6638074	1
T_LF_PART	0.39930173	-0.5707648	0.6024009	-0.6163465	-0.671434759
T_MED_AGE	0.011878	-0.5216322	0.55030178	-0.292623	-0.583149907
T_MS_MARRI	0.21304903	-0.6277209	0.42758459	0.2750185	-0.007175136
totalpop	0.06769891	-0.1300915	-0.02447148	0.0779205	0.091705154
	T_LF_PART	T_MED_AGE	T_MS_MARRI	totalpop	
T_AP_ASIAN					
T_BPL_FAM					
T_ED_PSDEG					
T_ED_SCHPE					
T_FAM_AVGS					
T_LF_PART	1				
T_MED_AGE	0.21181777	1			
T_MS_MARRI	0.06603202	0.34601887	1		
totalpop	0.08707065	-0.22979151	0.153740614	1	

Table 6. Correlation Matrix of Regression Model 8

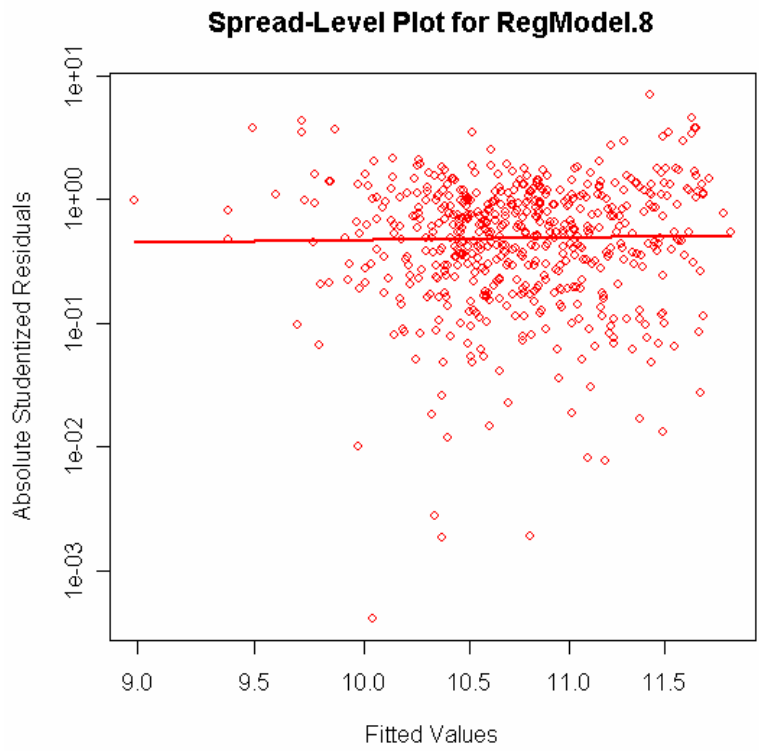


Figure 4. Studentized Residuals against Fitted Values in Regression Model 8

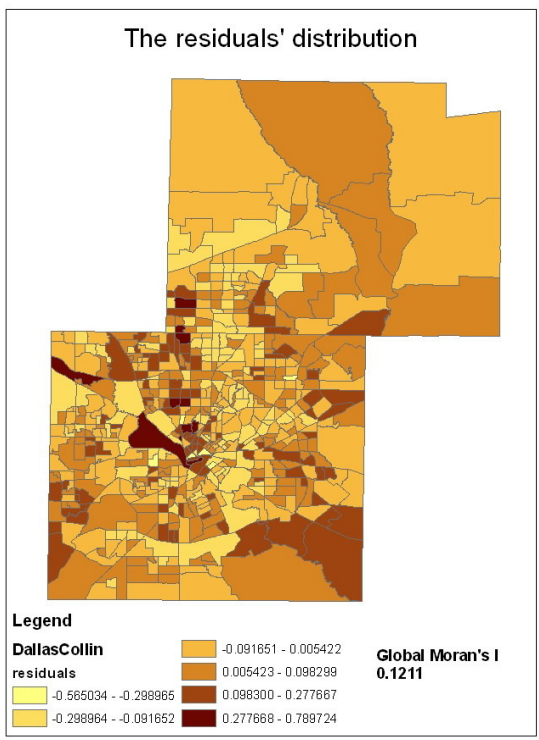


Figure 5. Residual Spatial Distribution

Variable		Model I
Intercept	(Intercept)	-0.000000000000001479
		(-0.0000000000000136)
Asian (%)	T_AP_ASIAN	-0.04938*
		(-3.878)
Poverty Line Families (%)	T_BPL_FAM	-0.3335*
		(-15.758)
Post Secondary Degree (%)	T_ED_PSDEG	0.5739*
		(28.596)
School Children (%)	T_ED_SCHPE	0.1023*
		(5.896)
Average Family Size (Average)	T_FAM_AVGS	0.1195*
		(5.392)
Labor Force Participation (%)	T_LF_PART	0.0401.
		(2.003)
Median Age (Number)	T_MED_AGE	0.06345*
		(3.58)
Married Individuals (%)	T_MS_MARRI	0.277*
		(13.661)
Tract Total Population (Number)	totalpop	-0.02202.
		(-1.856)
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1		

Table 7. Beta weights of Regression Model 8 (T-stat's in parentheses)

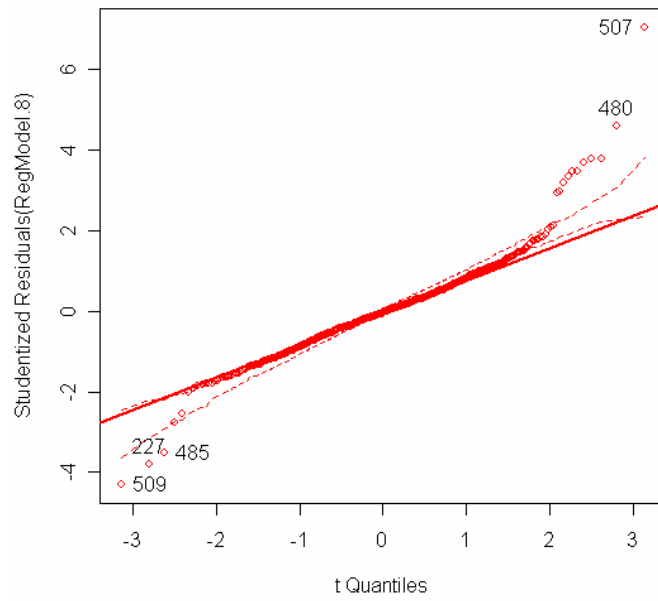


Figure 6. Residual-Quantile Comparison

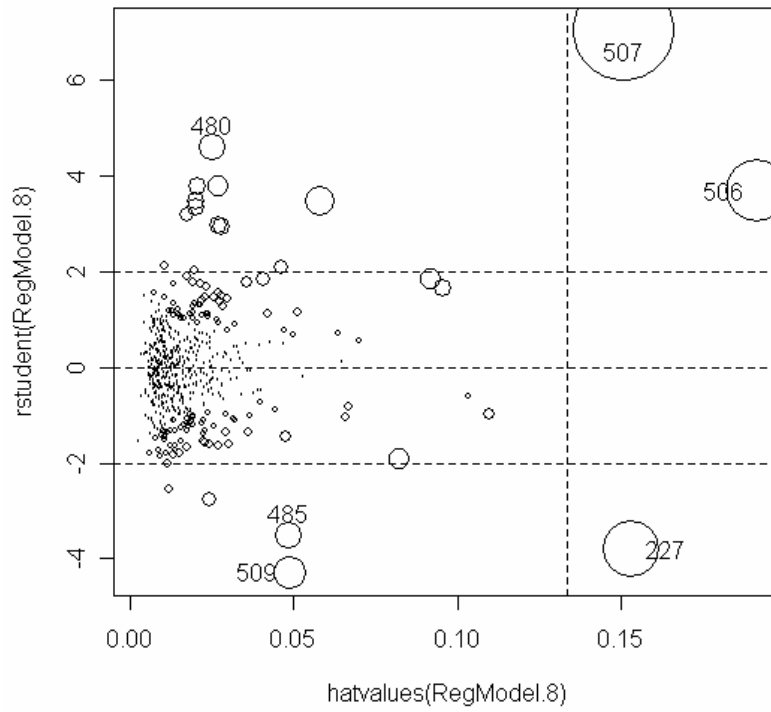


Figure 7. Influence Plot

Observation	Residual	Standardized Residual	Studentized Residual	Leverage (Hat)	Cook's Distance
227	-0.4375901	-3.766746911	-3.813080019	0.153086221	<b>0.256466</b>
480	0.5640551	4.525519274	4.608477143	0.025137493	0.05280991
485	-0.4293264	-3.486184231	-3.522346438	0.048274594	0.06164639
507	0.7861443	6.756840207	7.051338733	0.150517777	<b>0.8089484</b>
509	-0.5218008	-4.237990284	-4.305387581	0.048680027	<b>0.09190605</b>

Table 8. Outlier Analysis

	min.	median	mean	max.
Residual Summary	-0.5218	-0.004557	5.23E-19	0.7861
Leverage Summary	0.00269	0.01364	0.01795	0.1915
Cook's Distance Summary	2.09E-10	0.0004102	0.004614	0.8089
Standardized Residual Summary	-4.238	-0.03623	0.001308	6.757
Studentized Residual Summary	-4.305	-0.0362	0.002195	7.051

Table 9. Influence Analysis Summary

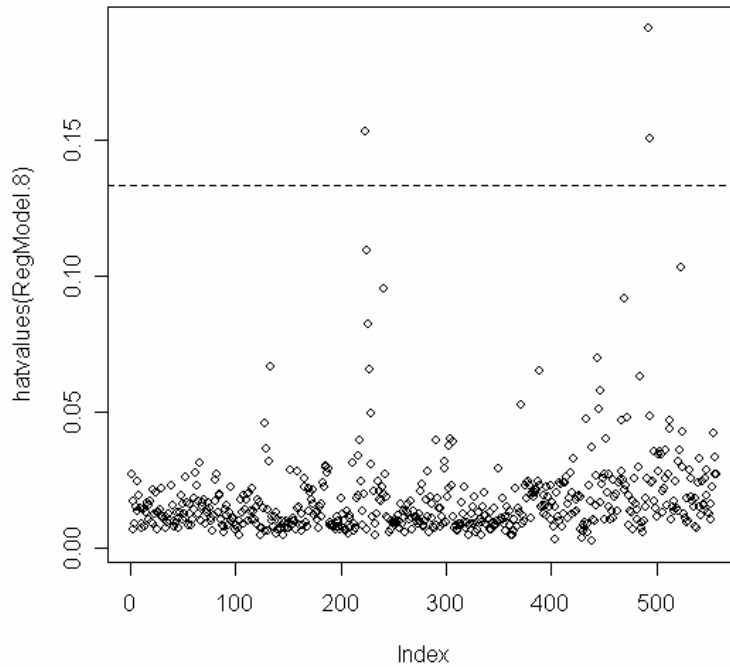


Figure 8. Hat Value

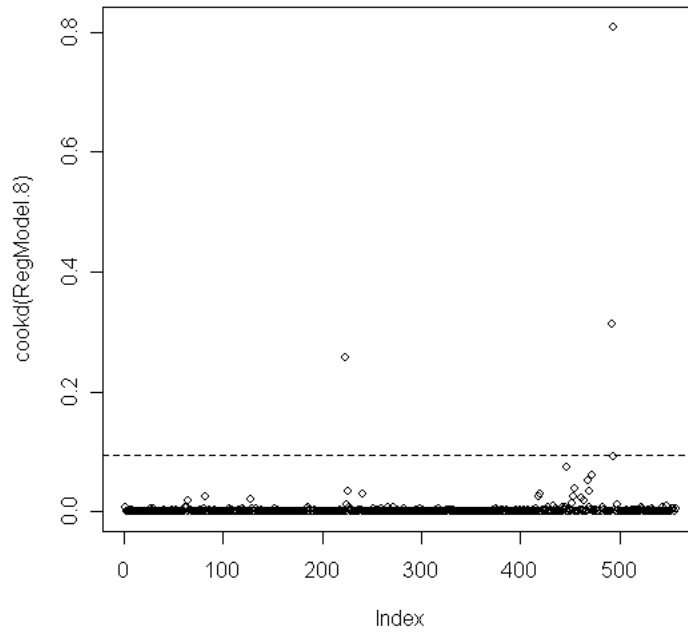


Figure 9. Cook Distance

Observation	(Intercept)	T_AP_ASIAN	T_BPL_FAM	T_ED_PSDEG	T_ED_SCHPE
227	-0.07812833	<b>-0.21256925</b>	-0.05136834	<b>0.13899328</b>	0.05596637
480	-0.01764979	<b>0.095356104</b>	-0.03390647	-0.030387988	-0.0012488
485	-0.00626888	0.04341907	-0.00668965	-0.033564027	-0.013545782
507	0.019166184	-0.078583621	0.000317009	0.049504086	-0.053107632
509	-0.00153835	-0.00255535	0.002312766	-0.000054753	-0.003127922
Observation	T_FAM_AVGS	T_LF_PART	T_MED_AGE	T_MS_MARRI	totalpop
227	<b>0.09687525</b>	0.05493478	0.02409191	<b>-0.09630902</b>	0.08248969
480	0.043647718	0.010815377	0.008337372	-0.030704825	-0.014116199
485	0.004971095	-0.005517761	0.022234937	0.01320462	-0.000557397
507	0.002669647	0.030229103	-0.04781269	-0.005199424	-0.030614714
509	0.000308792	0.004000573	-0.00517732	0.008052507	0.000401828

Table 10. DFBETAS of Influential Census Tracts