

UCTC Final Report

Why Do Inner City Residents Pay Higher Premiums? The Determinants of Automobile Insurance Premiums

Paul M. Ong and Michael A. Stoll

School of Public Affairs
UCLA
3250 Public Policy Bldg., Box 951656
Los Angeles, CA 90095-1656
pmong@ucla.edu and mstoll@ucla.edu

Why Do Inner City Residents Pay Higher Premiums? The Determinants of Automobile Insurance Premiums

Abstract

Auto insurance rates can vary dramatically, with much higher premiums in poor and minority areas than elsewhere, even after accounting for individual characteristics, driving history and coverage. This project used a unique data set to examine the relative influence of place-based socioeconomic characteristics (or redlining) and place-based risk factors on the place-based component of automobile insurance premiums. We used a novel approach of combining tract-level census data and car insurance rate quotes from multiple companies for sub-areas within the city of Los Angeles. The quotes are for a hypothetical individual with identical demographic and auto characteristics, driving records and insurance coverage. This method allowed the individual demographic and driving record to be fixed. Multivariate models are then used to estimate the independent contributions of these risk and redlining factors to the place-based component of the car insurance premium. We find that both risk and redlining factors are associated with variations in insurance costs in the place-based component, with black and poor neighborhoods being adversely affected, although risk factors are stronger predictors. However, even after risk factors are taken into account in the model specification, SES factors remain statistically significant. Moreover, simulations show that redlining factors explain more of the gap in auto insurance premiums between black (and Latino) and white neighborhoods and between poor and nonpoor neighborhoods. The findings do not appear sensitive to the individual characteristics of the hypothetical driver.

Background

This project examined the factors generating higher auto insurance premiums in poor and minority neighborhoods in Los Angeles. Automobile ownership is essential to accessing economic opportunities in modern metropolitan labor markets, and any financial barriers to such access can have adverse impacts on employment and earnings. A key cost is insuring a vehicle, which can be a significant burden for low-income and disadvantaged households. But car insurance rates can vary dramatically, with much higher premiums in poor and minority areas than elsewhere, even after accounting for individual characteristics, driving history and coverage (Ong, 2002). An important policy issue is whether the higher car insurance premiums faced by economically disadvantaged households are based on *de facto* discrimination or on fair and legitimate risk factors. The former explanation is widely known as “redlining,” a discriminatory practice of setting rates for individuals because they live in places with a disproportionately large number of minority and/or poor people. On the other hand, the insurance industry has long asserted that the unequal spatial pattern of car insurance premiums is based on place-based risk factors and that the demographic characteristics of neighborhood are not part of the formula used to set prices.

The debate is not just academic. California, where the Department of Insurance is reconsidering how much weight should be given to place characteristics in setting car insurance premiums, is an example of how this debate can affect public policy. Some, such as civil rights and drivers’ interest groups, advocate entirely eliminating the use of place in setting of car insurance premiums because such a practice lends itself to redlining, especially when effective oversight is absent. On the other hand, others, such as the insurance companies themselves, are fighting for the right to continue its use in order to protect themselves from place-based risks that are not fully revealed in individuals’ driving records but that may affect insurance costs (Friedman, 2006).

Developing a sound policy depends in part on determining the relative role of risk versus race and class in the existing place-based premium structure. Clearly, in this domain, race should have no role in setting such premiums because it would violate civil rights statutes prohibiting racial discrimination in civic and business life. Unfortunately, it is difficult to determine the relative contribution, if any, of these factors in the setting of insurance premiums because the research that the insurance companies use to support their position is treated as proprietary information, and thus not easily evaluated by third parties.

To fill the gap, this project uses a unique data set assembled from numerous information sources to examine the relative influence of place-based socioeconomic characteristics and place-based risk factors on the spatial pattern of automobile insurance premiums in the city of Los Angeles. To examine this, we use a novel approach of combining tract-level census data and car insurance rate quotes from multiple companies for sub-areas within the city of Los Angeles. The quotes are for an individual with identical demographic and auto characteristics, driving records and insurance coverage. This method allows the individual demographic and driving record to be fixed. Multivariate models are then used to estimate the independent contributions of these risk and race and class factors to the place-based component of the car insurance premium.

Data and Method

The data for this study come from several sources. Car insurance premiums, the outcome variable of primary interest, were collected over the internet for the year 2000 from multiple quotes from multiple insurance companies for each zip code in the city of Los Angeles. Insurance premium

estimates were for the liability component only and were provided by the following website: <http://www.realquote.com>. To capture the “pure” geographic variation of insurance rates, we held the characteristic of the “applicant” constant by using the same demographic profile for every zip code: a 25-year old, employed single mother, who has been driving for 7 years, has taken a driver training course, has one moving violation, but no accidents, and does not smoke. She owns a 1990 Ford Escort LX, two-door hatchback with no anti-theft devices, no anti-lock brakes, and no airbags, and parks on the street. She carries only the minimum insurance required (\$15/30,000 bodily liability, \$5,000 property liability) with no deductibles. The insurance premium for each zip code is the average of quotes from at least a half dozen companies.

Socioeconomic status (SES) factors related to the redlining hypothesis are based on tract-level data from the 2000 Census SF-3 (Summary file 3) data set. We include the percentage of the tract population that is black, Hispanic, and the percentage of the population that is below the poverty line as the main socioeconomic status variables.

The risk variables included in the analysis are those that the insurance companies argue are related to higher risks. These include claim and loss rates as the primary risk factors and accident and crime rates as secondary risk factors. The data on insurance claims come from the California Department of Insurance (CDI). It contains data on private passenger auto liability from 2000 as well as the physical damage experienced broken down by coverage and program type and by zip code. The claim rate is defined as the number of claims per 10,000 insured persons in a zip code per insurance year.

The data on the amount of dollar loss for all insurance companies also come from the California Department of Insurance. The loss rate measures the average loss payments per insured vehicle year normalized by total number of exposures years across zip codes. Exposure years are defined as the number of car-years for which the insurance companies are liable.

Insurance companies also claim that some secondary risk factors, such as the accident and crime rate among others, are taken into account in setting rates. We include these in separate models to evaluate their importance, though we acknowledge their limitations. Moreover, their inclusion does not change the basic findings of the analysis. The accident data come from Los Angeles City Department of Transportation (LADOT). These files document all reported traffic collisions in the city of Los Angeles from 1994 through 2002, with a small number of records for 1993 and 2003. The data are thus restricted to 1994-2002, as the dataset is most complete for these years. The accident locations were geocoded and summed within census tracts, and we use a buffer to capture the accidents that occur on the streets that are on the tract boundaries. From this we defined the accident rate, which is measured as the number of accidents normalized by the area size (in square kilometers) of each tract.

The car crime data was provided by the city of Los Angeles Department of Police (LAPD) for reporting districts (RD) for the year 2000. RDs are the geographical units used by the Police Department to aggregate crime data, and many of these units are equivalent to census tracts. The crime variable used in this study is the aggregated number vehicle thefts. The variable is also normalized by the area size (in square kilometers) of each RD.

Since the insurance data are based on zip codes and the other variables based on census tracts, the former variables are transformed into census tracts by using GIS proportional split method. We used the proportional split method to redistribute spatial values throughout a specific area. The

method assumes that population is evenly distributed over a geographic area. However, this is a rather strong assumption since it is not always the case that population is distributed in this manner. Nevertheless, this method is inexpensive and does generate estimates quickly and impartially without resorting to more accurate, costly techniques. Because the method assumes uniform dispersion of attributes within each zip code, it might cause a problem when there is an extreme unbalance of distribution of attributes (Schlossberg, 2003). However, an analysis of comparing total numbers and patterns of maps indicates that these concerns are limited.

Table 1 reports the averages and standard deviations for the dependent (premium) and independent variables (risk and SES) in the analysis. The average car insurance premium in the city of Los Angeles over the study period for the typical driver described above is \$1,041 dollars, with the median estimate slightly lower because of high top end premiums that are pulling up the average estimates. In the sample, the main risk factors indicate an average claim rate of 220 per 10,000 insured persons and an average loss rate of \$212 per insured vehicle. The secondary risk factors show average accident and crime rates of 1,822 and 279 per square kilometers, respectively. Finally, consistent with the demographic composition of the city of Los Angeles, the mean poverty rates is 22.4 percent while the percentage of the population that is black or Latino is 11 and 43 percent, respectively.

More importantly, Table 2 examines the extent to which insurance premiums vary with these place-based characteristics. To do this, for each independent variable on the left hand column, we divide the data into terciles ranked by the values of these respective independent variables. Then, for each tercile category (for each independent variable) we calculate the average car insurance premium. The table clearly shows that insurance premiums are much higher in areas characterized by greater risk. In particular, in areas with greater claim and loss rates, insurance premiums are higher. The same is also true for the accident and crime rate measures of risk. The difference in the insurance premiums for most of the risk variables is about a \$100 or so from the bottom to the middle or from the middle to the top tercile category.

These observed relationships between the risk factors and insurance premiums are verified in the correlation estimates provided in the right side column of the table. The correlations between the primary and secondary risk factors and the insurance premiums are all fairly large in magnitude and statistically significant. More importantly, the strength of the association between the risk factors and premiums are slightly stronger for the primary than secondary risk factors as we should expect.

At the same time, the redlining factors are also correlated with premiums. The correlation values in the last column in Table 2 clearly show that insurance premiums are higher in areas characterized by higher poverty rates and where the percent of the population that is black is higher, as indicated from the bottom to the top of the respective tercile categories. The association between %Latino and insurance premiums is weakly related however, perhaps because of the wider spatial dispersion of this population (than for blacks) and the existence of small Latino enclaves close to non-Hispanic white neighborhoods. Still, the rate of the increase in premiums moving from the bottom to the top terciles for the SES variables (especially for %black) is not as great as that for the risk factors. This is confirmed in the bivariate correlation estimates between the SES factors and insurance premiums. While the correlation estimates are positive and statistically significant between all SES factors and the insurance premiums, they are much more weakly associated than between the primary (and even secondary risk factors) and these premiums.

Table 3 shows data on the simple bivariate correlations among and between the risk factors and SES variables. The table shows that, for example, the poverty rate is correlated with all the risk factors, implying that the correlation between poverty status and insurance premiums may be statistically spurious due to the collinearity between poverty and risk.

On the other hand, the percent of the population that is black variable is either not correlated or only weakly correlated with the risk variables. This may indicate that the observed bivariate relationship between %black and insurance premiums is much more robust than that shown above. Finally, the percent of the population that is Latino variable is correlated with most of the risk variables, but not in a consistent direction.

Multivariate Findings

We use ordinary least squares (OLS) regressions to estimate the influence of the redlining factors and risk factors on insurance premiums. The first model examines the “naïve” redlining model in which only the SES factors are included. Then we examine the “naïve” primary risk factor models where we include only the claim rate and loss rate variables. Next, we examine a model in which both the redlining and primary risk factors are included to examine the independent effect of these factors on auto insurance rate premiums. Model 4 includes only the secondary risk factors such as the accident rate and crime rate. Model 5 includes the primary and secondary risk factor variables to examine their independent contribution on auto insurance premiums since these factors are highly correlated. Finally, Model 6 includes all redlining and risk factor variables in the analysis. Table 4 reports the estimated coefficients for these models.

Model 1 is the “naïve” redlining model, where only the SES factors are included. All coefficients are highly statistically significant and the model explains about a quarter of the variation in the place-based component of the auto insurance premium. The percent of the population that is comprised of people below the poverty line and the percent comprised of blacks are positively related to the insurance premium. In particular, living in a poorer neighborhood (25% poverty rate) would increase insurance premiums by about \$116 relative to living in a neighborhood that has few in poverty (5% poverty rate). The estimated difference between an all black neighborhood and one without any blacks is \$112.

Interestingly, the estimated coefficient for the percent of Latino is negative. Model 2 includes only the primary risk factors deemed important by insurance companies. The results indicate that all coefficients are positive, highly statistically significant and consistent with expectations based on the risk hypothesis; that is, premiums are higher in neighborhoods with higher driving risks. The model explains about 51% of the variation in premiums, fairly large for a cross-sectional model. The coefficient for the linear loss-rate variable indicates that each additional dollar loss leads to a \$0.15 increase in premium, but the coefficient is not statistically significant. Model 2b includes the square of the loss rate since we assume that there is a curvilinear trend between premiums and the loss rate. The results indicate that the premiums increase with the loss rate but at a slower rate at the high end of the distribution. The inflection point is at \$226 dollars, which is just above the mean value of this independent variable. In other words, premiums start to decline in the upper range of the losses. (Interestingly, the average loss value is lower in predominantly black neighborhoods and very poor neighborhoods than in non-Hispanic white neighborhoods that have very low poverty rates. These types of neighborhoods are discussed later.)

Model 3 includes both the redlining and primary risk factors and indicates that 62 percent of the variation in insurance premiums is explained by these combined factors. More importantly, the results show that both risk and SES are statistically significant, providing evidence of both the redlining and risk hypothesis of setting auto insurance premium levels.

But with the inclusion of the primary risk variables into the model specification, the coefficient estimate on the percent poverty decreases by 47 percent, indicating that much of the positive relationship between the poverty rate of an area and the insurance rate premium is driven by the higher primary risk factors in such areas. Interestingly, the coefficient estimate on percent black becomes slightly, but not statistically significantly larger, indicating that primary risk factors are not driving the higher insurance premiums in areas where there is a higher percentage of the population that is black. Finally, the negative coefficient on the percent Latino variable becomes less negative with the inclusion of the primary risk factor variables suggesting that much of the negative relationship between the poverty rate of an area and the insurance rate premium is influenced by lower primary risk factors in such areas.

Model 4 includes only the secondary risk factors. The coefficient estimates are highly statistically significant and demonstrate that both the accident rate and crime rate of an area are positively associated with the place-based component of the auto insurance premium. The point estimate for the accident rate indicates that an increase in accidents over a square kilometer by 1,000 is associated with an increase the car insurance premium of about \$40.00.

Model 5 includes only the primary and secondary risk factors into the model specification. The inclusion of these variables in most cases leads to slightly lower point estimates of the association of the primary and secondary risk factors and the insurance premiums. Still, when both are included, all risk factors remain positive and statistically significantly related to insurance premiums.

Model 6 includes the SES variables and both the primary and secondary risk factors into the model specification. About 64 percent of the variation in insurance premiums is explained with all variables in the analysis included. In addition, the results show that both the SES and most risk factor variables are statistically significant, providing strong evidence of both the redlining and risk hypothesis of setting auto insurance premium levels. With the inclusion of the primary and secondary risk variables into the model specification, the coefficient estimate on the percent poverty decreases by 32 percent from the coefficients estimate in Model 3, indicating that much of the positive relationship between the poverty rate of an area and the insurance rate premium is driven by the higher primary and secondary risk factors in such areas. However, like the results in Model 3, the coefficient estimate on percent black becomes slightly larger with the inclusion of the secondary risk factors into the specification suggesting that like the primary risk factors, the secondary risk factors are not driving the higher insurance premiums in black areas.

To estimate the relative role of redlining and risk, we use results from Model 3 to simulate insurance premiums because of the concerns about the secondary risk factors. However, we compare these results with decompositions from Model 6 since the latter model's results provide more conservative estimates of the role of race in setting premiums. We include four types of neighborhoods: (1) non-Hispanic white (NHW) and non-poor tracts where blacks, Latinos and those below poverty individually comprise no more than 5% of the population, (2) black tracts where blacks comprise at least 75% of the population, (3) Latino tracts where Latinos comprise at least 75%

of the populations, and (4) poor tracts where those below the poverty line comprise at least 40% of the population.

The first set of simulations is based on multiplying the parameters from Model 3 by the average demographic and risk values for each of the four types of neighborhoods. To estimate the impact of differences in risk, the neighborhood-specific average risk values are replaced by the average risk values for NHW and non-poor tracts. Finally, to estimate the impact of differences in demographic characteristics, the neighborhood-specific average demographic values are replaced by the average demographic values for NHW and non-poor tracts. Figure 1 reports the results.

Figure 1 indicates that the relative importance of redlining and risk varies by neighborhood type. The estimates for black neighborhoods reveal that when such neighborhoods are given the same level of neighborhood driving risk factors as that in NHW, non-poor neighborhoods, the simulated average insurance premium declines by a trivial amount. However, when black neighborhoods are given the same racial composition as NHW, non-poor neighborhoods (and given the same level of neighborhood driving risk factors as that in these neighborhoods) the simulated average insurance premium declines by a significant amount. This implies that in black neighborhoods, differences in risk account for a very small part of the higher insurance premiums relative to that for NHW, non-poor neighborhoods, while differences in demographic composition explain an overwhelming majority of the gap. More precisely, about 3 percent of the \$154 gap in insurance premiums between these areas is explained by risk factors while 97 percent of the gap is explained by differences in demographic characteristics. When results from Model 6 are used in these decompositions, about 11 percent of the gap in insurance premiums is explained by risk factors while 89 percent of the gap is explained by differences in demographic characteristics.

The premium gap between Latino neighborhoods and NHW and non-poor neighborhoods is relatively small (\$48). However, differences in demographic composition explain the entire gap (100 percent). When results from Model 6 are used in these decompositions, about 73 percent of the gap in insurance premiums between Latino neighborhoods and NHW and non-poor neighborhoods is explained by differences in demographic characteristics.

The premium gap with NHW and non-poor for poor neighborhoods is similar to that encountered by black neighborhoods (\$156), but risk factors explain slightly more of the gap (24 percent) than in black neighborhoods. Still, as is the case in black neighborhoods, demographic differences explain more of the gap between poor and the non-poor neighborhoods (76 percent) than risk factors. Again, when results from Model 6 are used in these decompositions, demographic factors account for less of the gap than in Model 3 (about 55 percent), but is still the larger influence in the gap as compared to risk factors. Taken as a whole, in both the simulations shown here based on Model 3 and the more conservative estimates (around the role of redlining) from Model 6, the results indicate that redlining contributes more to higher premiums in black, poor and Latino neighborhoods than risk.

To address the concern of whether the average automobile insurance premiums for our hypothetical driver serve as a reasonable proxy for other types of drivers, we analyze more recent data regarding the spatial variations in premiums for several types of drivers. For this exercise, we gathered 125 current quotes for semi-annual premiums from the website of a major insurance firm for areas that are representative of the various communities in Los Angeles. The quotes cover 25 zip

code areas, with quotes for six types of individuals for each zip code. The six types vary by gender, driving record, age and marital status.¹ The coverage included only the legally required minimum coverage of liability and bodily injury. Each of the six types of drivers is the owner of a hypothetically identical six-year old Honda Civic sedan, which is driven 12,000 per year and 5 days a week to work with a distance of 10 miles each way. The vehicle is a six-year old Honda Civic sedan. The quotes were downloaded September 22, 23 and 24, 2006.

One of the hypothetical type of person is equivalent to the hypothetical driver used to construct the original 2000 data set, and the geographic variations in her premiums are highly correlated with the geographic variations for the other five types of individuals (for each pair-wise comparison, $r > .9999$ and $p < .0001$ with $n = 25$). The regression results are available from the authors and show that within a given zip code area the premium for the first type of person is a highly predictive of the premiums for the other types of drivers. The estimated model has an adjusted r-square of .987. The high explanatory power of the model means that it is capturing most of the variations associated with the formula used by the insurance company to determine rates, which is unknown to us. The coefficients for the dummy variables are all statistically significant ($p < .0001$), and follows a predictable well known pattern – males pay more, those with an accident pay more, and married older drivers pay less. The estimated coefficient for the hypothetical driver is .992, indicating that each dollar difference in premium across zip code areas for the hypothetical driver translates into a dollar difference for other drivers.

These findings provide evidence consistent with the idea that the geographic variation for our one hypothetical driver does a very reasonable job of capturing the geographic variation for other types of drivers.

Conclusion

In this project, we used the best data available to academic researchers (not working directly for the insurance industry) and a novel approach to examine the relative importance of risk and redlining factors in the setting of the place-based component of auto insurance premiums. The basic results showed that both risk and redlining factors are associated with auto insurance premiums in the expected direction, although the risk factors were found to be statistically more important determinants of the premiums. However, even after risk factors are taken into account, SES factors remain statistically significant. Moreover, simulations of the regression equations demonstrate that redlining factors explain more of the gap in the premiums across black (and Latino) and white neighborhoods and between poor and nonpoor neighborhoods.

Journal Paper: A journal paper based on this study, entitled “Risk or Redlining: A Spatial Analysis of Auto Insurance Premiums in Los Angeles”, was submitted to the *Journal of Policy Analysis and Management*.

¹ The drivers include: (1) single female, age 26 with a clean driving record; (2) single male, age 26 with a clean driving record; (3) single female, age 26 with an accident 4 years ago; (4) single male, age 26 with an accident 4 years ago; (5) married female, age 36 with a clean driving record; and (6) married male, age 36 with a clean driving record.

References

- Friedman, J. (2006), Steadfast believer in numbers: The state's plan to revamp the setting of car insurance rates doesn't add up for Mercury's George Joseph. Los Angeles Times, Business Section, February 5, 2006.
- Ong, P. & Sung, H. G. (2003). Exploratory study of spatial variation in car insurance premium, traffic volume and vehicle accidents. Berkeley, CA: University of California Transportation Center Working paper no. 654.

Figure 1: Simulation Results by Neighborhood Type

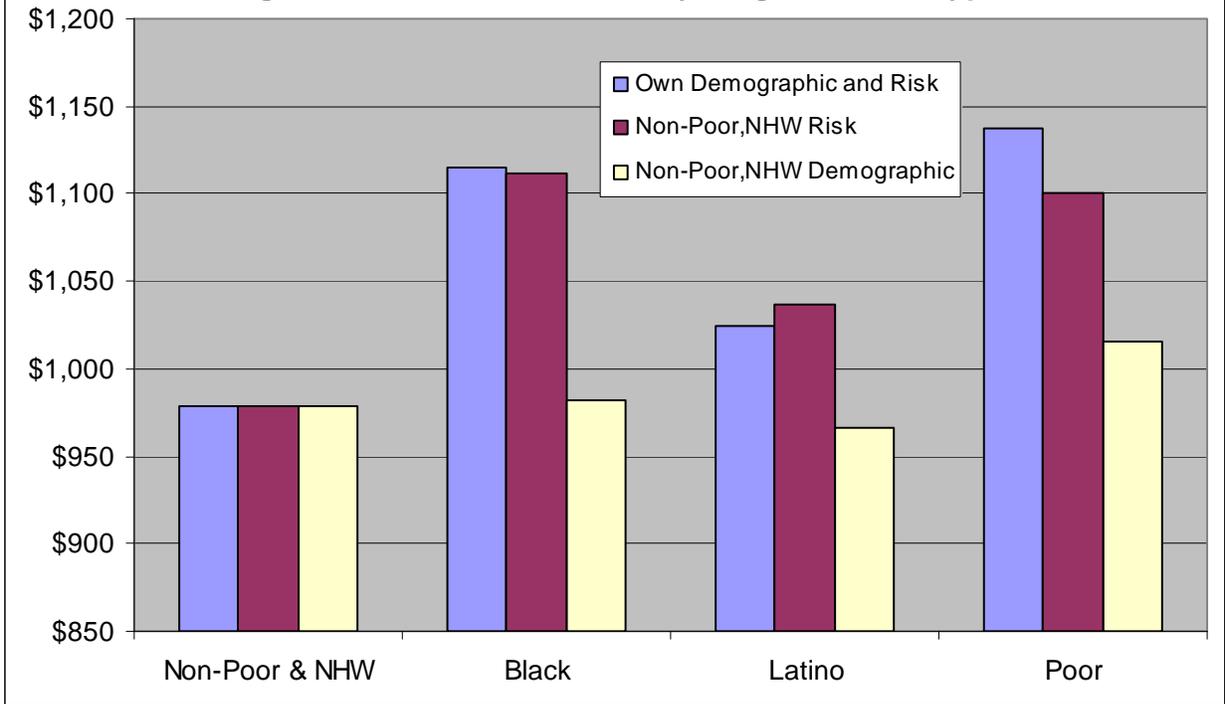


Table 1. Descriptive Statistics

	Variables	Mean	Median	Std Deviation
Dependent Variable	Premium	1,041.4	1,028.1	151.5
Neighborhood Socioeconomic Status	% Poverty	22.4	20.4	14.2
	% Black	10.8	4.1	16.7
	% Latino	43.2	44.3	27.7
Insurance-based Risk	Claim rate	220.1	212.9	48.4
	Loss rate	212.7	204.8	53.3
Other Risk Factors	Accident rate	1,821.7	1,359.0	1,659.0
	Crime rate	279.2	221.3	235.1

Table 2. Average Insurance Premiums by Risk and SES Characteristics of Neighborhoods
(Trisected by each variable)

Mean	Insurance Premium			Correlation to Premium
	Bottom 3rd	Middle 3rd	Top 3rd	
% Poverty	977.4	1017.9	1128.2	0.4089***
% Black	998.6	1034.3	1090.9	0.2677***
% Latino	1015.8	1059.5	1048.8	0.0761*
Claim Rate	903.8	1060.4	1159.2	0.7139***
Loss Rate	939.4	1019.9	1164.1	0.6230***
Accident Rate	933.4	1033.0	1156.9	0.6143***
Crime Rate	945.4	1033.9	1144.2	0.5312***

P-value: * p<0.05; ** p<0.01; *** p<0.001

Table 3. Correlations among Variables

	% Poverty	% Black	% Latino	Losses	Accidents	Crime
% Poverty	1	0.2273***	0.6328***	0.1544***	0.5170***	0.5840***
% Black	0.2273***	1	-0.0773*	0.0054	0.0261	0.0848*
% Latino	0.6328***	-0.0773*	1	-0.1059**	0.2576***	0.2523***
Claims Rate	0.2724***	0.0772*	0.0538	0.8530***	0.6361***	0.5625***
Losses Rate	0.1544***	0.0054	-0.1059**	1	0.5463***	0.4553***
Accidents Rate	0.5170*	0.0261**	-0.2576***	0.5463***	1	0.7888***
Crime Rate	0.5840***	0.0848*	0.2523***	0.4553***	0.7888***	1

P-value: * p<0.05; ** p<0.01; *** p<0.001

Table 4. Regression Result – Dependent Variable: Insurance Premium

	Model 1	Model 2a	Model 2b	Model 3	Model 4	Model 5	Model 6
Constant	960.2	548.4	295.7	267.1	933.2	278.1	298.7
% Poverty	5.822***	N/A	N/A	3.106***	N/A	N/A	2.127***
% Black	1.122***	N/A	N/A	1.207***	N/A	N/A	1.410***
% Latino	1.416***	N/A	N/A	-0.409*	N/A	N/A	-0.358*
Claim rate	N/A	2.097***	2.225***	1.778***	N/A	1.588***	1.467***
Loss rate	N/A	0.147	2.264***	2.644***	N/A	3.173***	2.866***
Loss rate squared	N/A	N/A	0.005***	0.005***	N/A	0.007***	0.006***
Accident rate	N/A	N/A	N/A	N/A	0.047***	0.023***	0.023***
Crime rate	N/A	N/A	N/A	N/A	0.079**	0.061*	-0.014
Number of Observation	836	836	836	836	836	836	836
Adjusted R-Square	0.233	0.509	0.530	0.619	0.382	0.592	0.644

P-value: * p<0.05; ** p<0.01; *** p<0.001