American Journal of Business and Management
Vol. 3, No. 3, 2014, 192-198
DOI: 10.11634/216796061403555

Analysis of Effects of Race on Insurance Companies' Operations in Chicago, December 1977 - November 1978

Haotian Zhang

College of Business and Economics, Australian National University, Australia

Dataset *Chicago* is about the residential fire insurance policies issued in Chicago over December 1977 through February 1978; categorized as either voluntary, or involuntary, which includes state-offered Fair Access to Insurance Requirement (FAIR) plans (Wriggins 2010) for people who would be denied insurance because of high risk classification, after Fair Housing Act (FHA) has been in act since 1968 (Calmore 1997). Data size is relatively small as n = 47 Zip codes. Theft, fire and the age of the house are also provided, along with the median income on the size of the expected loss and insolvency. Main purpose of this article is to explore the relationship between insurance activity and the variable race. S-Plus, rather than numerical methods such as risk theory models, was employed by using Data Plots to exam data itself, Regression Fitting to get possible candidates of linear expressions, and Model Selection to pick the best model under different scenarios. The final result proves insurance companies were using race as a determinative factor of underwriting insurance contacts, but not the only determinative factor.

Keywords: African American communities, race, insurance companies

"It's Not a Broken Promise if You Never Meant to Keep It."

Introduction

"Redling", this phenomenon is best referred as "outright refusal of an insurance company or lending institution to provide service solely on the basis of a property's geographical location" (Badain 1980), indicates insurers could use ethnical data to classify and select potential policyholders by canceling insurance policies or refusing to renew. Here, varies of methods are applied to exam whether residential insurance denied information from the data set were statistically correlated with residents' races. The involuntary market (Rice 1996) activity variable (numbers getting FAIR plan) is chosen as the response since this seems to be the best measurement of the denied. It is not a perfect measure given that who got denied insurance may give up and others still may not try at all for that reason. The voluntary market activity variable is not as relevant, though. Furthermore, only the racial composition in the corresponding zip code instead of race denied itself is presented. Finally, money has diminishing marginal return on people's utility, so a proper scaling is needed once processed.

One-way analysis here is not accessible, considering rating variables (race, fire rate, theft rate and age of properties) are not to the exclusion of each other, i.e. one rating variable likely to be influenced by differences in the mix of other rating variables, resulting in collinearities. For instance, a tendency exists, where African American communities, have relatively lower average income than other ethical hoods; the lower income rate, the higher theft rate, and higher consequent chance to get rejected by insurance companies (Joseph 1993). The collinearities of all three will show up in one-way tables of each of them, resulting in three time stronger relationships than they really are.

Consequently, study need to be done about the correlations between all these variables; about initial fitted line using linear regression method; about identifying outliers with different regressions and leverage points; about how to improve the regression and get a relatively accurate relationship between involuntary rate and racial composition along with housing age or fire rate, under control of / eliminating the effects from theft and income facts, by T-statistics, R-square and other relevant model selecting approaches.

Data is given by following zip code:
race: racial decomposition in percent minority;
fire: fires per 100 housing units in 1975;
theft: thefts per 1000 people in 1975;
age: percent of housing units built before 1939;

invol: new FAIR plans & renewals per 100 housing units in first half of 1978; **income**: median family income, divided by 1000 and waiting for future scaling;
vol: omitted in data processing;
Zip: zip code as aggregated information shall now be removed from the mode of regression and only used as reference.

Methodology

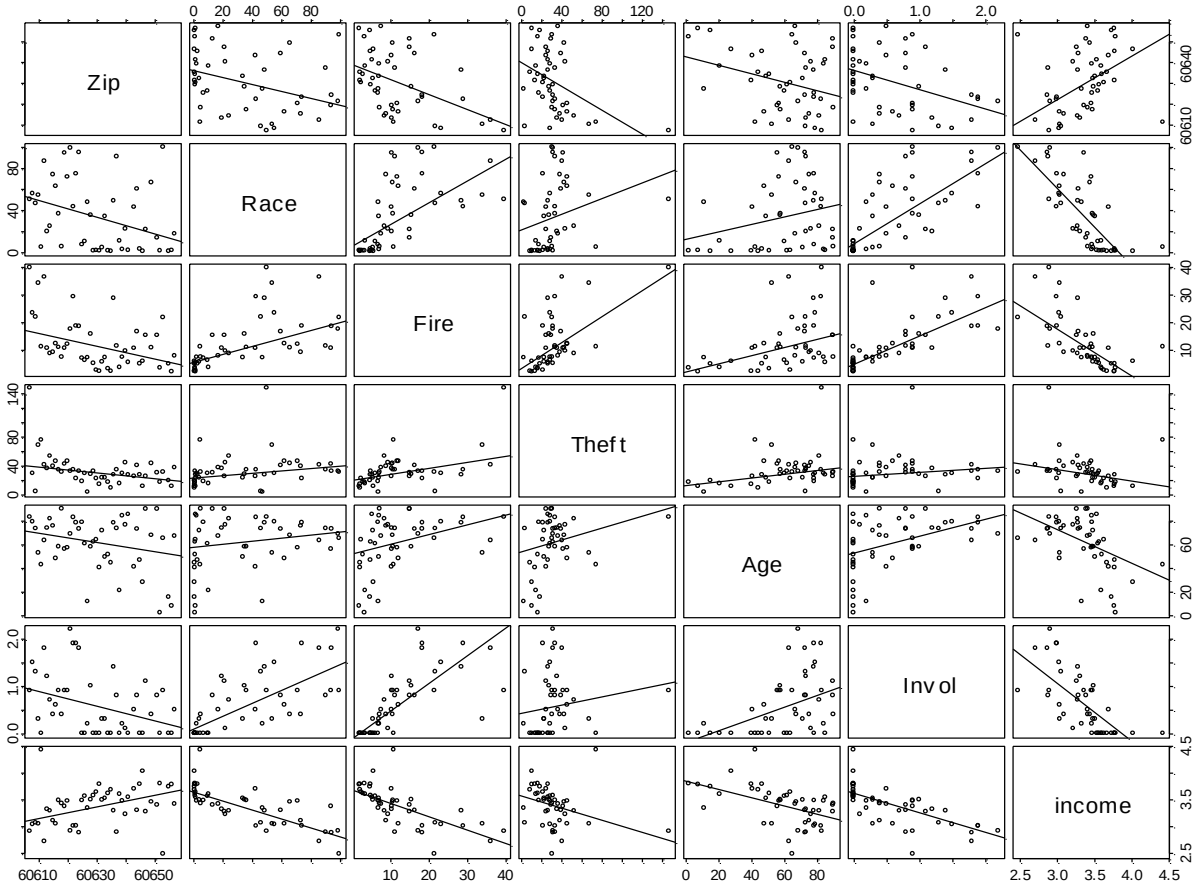
1) Use #Scatter Plot to examine covariates broadly:

a) Strong negative linear relationship between Race and income, suggesting either one of them can be reduced;

b) Theft rate pattern is quite "steep" in above scatter plot, implying sharp advance (decline) over small changes of other variables;

c) Certain amount of zeros are contained in Invol column, might need jitter the data to add noise and separate points in graphs and only in graphs;

d) If there exists racial discrimination, the more minorities in the Zip Code area, the higher chance they got denied by insurance companies (higher involuntary rate) (Cochrane 1991).



2) Use #Histograms / densities + #boxplots (sans Zip, and it DOESN'T need same x & y scales) to explore each data group besides Zip code:

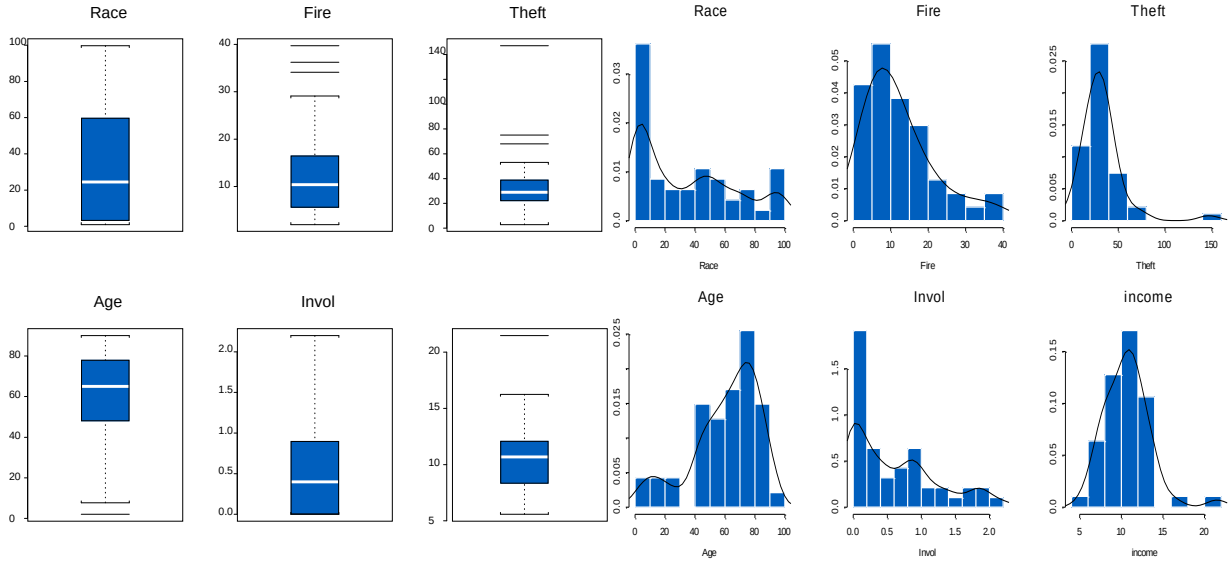
a) Wide range in Race variable, with some Zip codes being entirely minority or non-minority (Smink 2005). Hence either Zip or Race can be removed for the sake of reducing variation;

b) Data including Invol is somehow right skewed, besides Age;

c) Invol with many zeroes might be problematic with limited dependent values, need jittering;

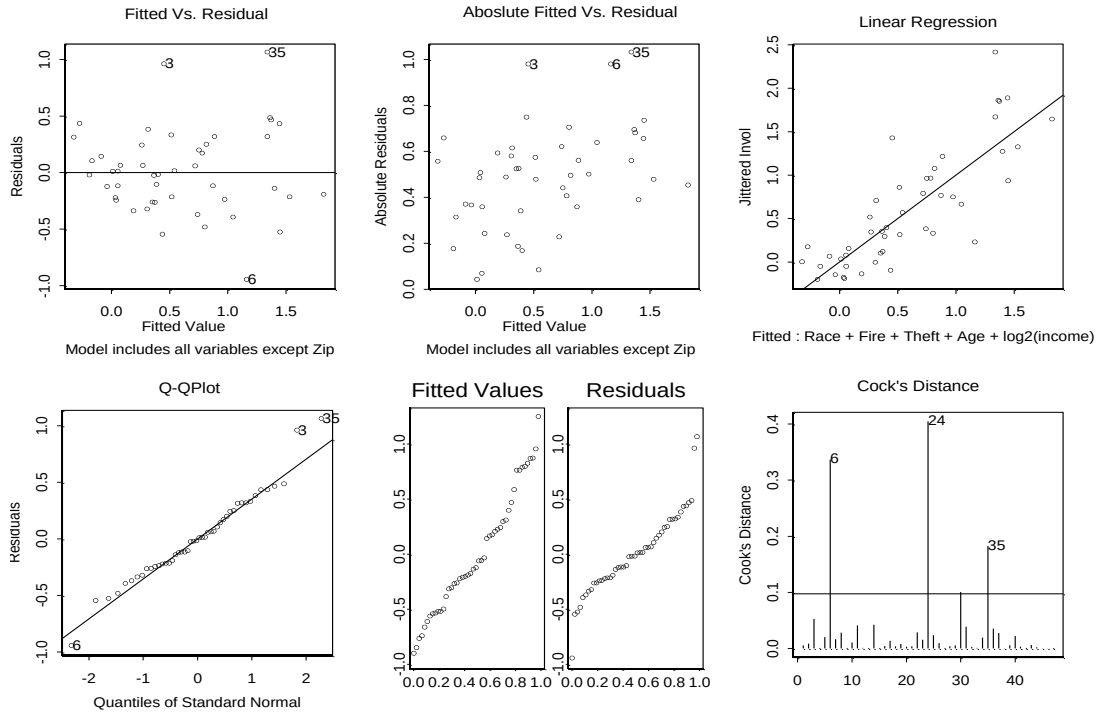
d) There's a "gap" in Age, six observations are smaller or equal to 28 and others are greater or equal to 40;

e) Meanwhile income seems to follow a bell-shaped pattern, so transformation to log2(income) = **Income** is applied for better interpretation purpose – such diminishing marginal return is more realistic as money's nature.



3) Find outliers and remove them from the model: First, check overall (almost) full model fitted without Zip for regression as Involuntary rate against Race, Fire, Theft, Age, and log2(income), with #jittering amount = 0.5, getting:

Invol = -1.1855 + 0.0095 Race + 0.0399 Fire - 0.0103 Theft +0.0083 Age + 0.2397 Income with following diagnostic plots:



Given S-Plus flagged out ZIP codes 60610, 60607 and 60621 (No. 3, 24 and 35 correspondingly) as influential observations with high leverage from both #Fitted vs. Residual and #Cooks distance with 4/(47-5-1) cut-off;

a) In summary, Theft has a negative coefficient while it has a positive relationship with involuntary rate from scatterplot, as a consequence of Race – Theft collinear relationship;

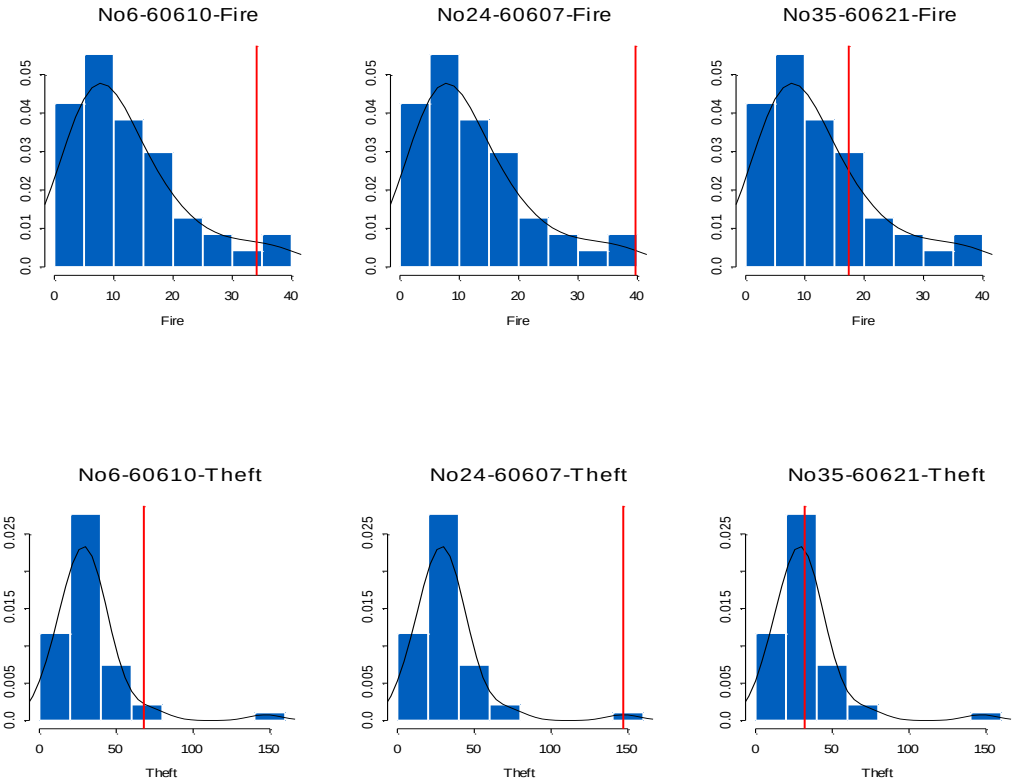
b) This All-in-one model has a roughly constant variance with zero mean from fitted vs. residual, a normality validation from normal Q-Q plot (evenly straight line, just little bit light tailed) and an independence from regression line.

Now potential outliers are being examined in Theft and Fire cases in #Histograms:

a) t is ostensible that No. 3, 24 and 35 of our data set with zip codes 60610, 60607 and 60621 correspondingly could have high leverage, i.e. they could be potential outliers;

b) No. 3 and 24 show in high Fire and Theft rates, No. 35 shows otherwise;

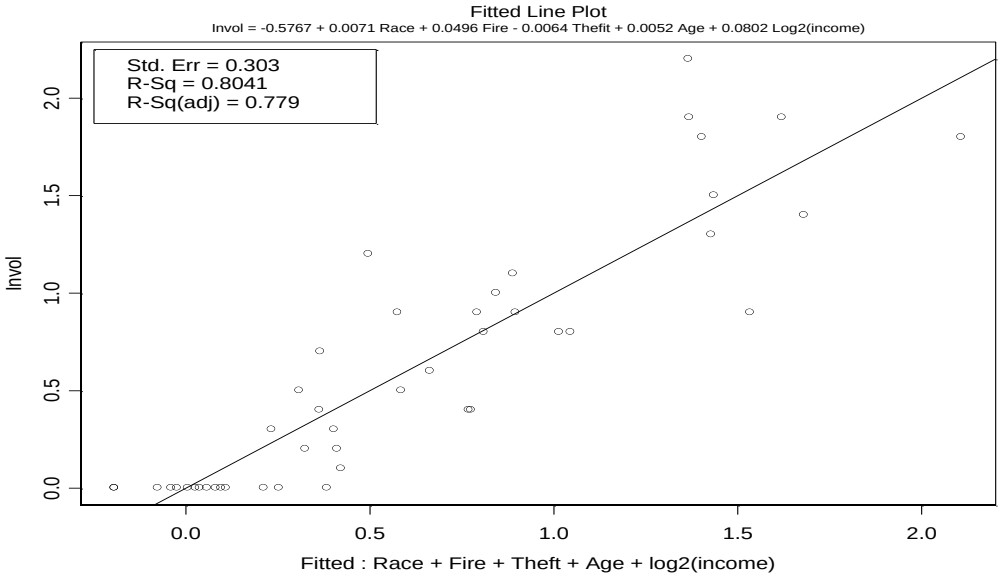
c) It is accurate to remove No. 3 and 24 of high Fire and Theft regions from the possible model since they are outliers and have strong influence on variables besides Race and Age; yet retaining No. 35.



4) Run new regression again, this time without c(6,24):

A Linear regression with not-so-bad fit (without jittering, too gimmick) and following Test Statistics – t value Pr(>|t|) (Intercept) -0.5340 0.5964

Race 2.6160 0.0126
Fire 5.7931 0.0000
Theft -1.4794 0.1471
Age 1.7863 0.0818
Income 0.2885 0.7745



a) Theft and Age now no longer significant at 5% level;
b) Race's P-value = 0.0126 < 0.05 is still statistically significant;
c) Fire is now the most significant variable among others;
d) Given that Income has a relatively large P-value in the rejecting zone comparing to the rest of variables, it can be eliminated judging by T-Statistics approach.

Model selection methods using direction= #BACKWARD, #FORWARD and #BOTH(STEPWISE) methods (Derksen 1992) (pure graphics doesn't work here) to see if the variable is statistically significant at 5% level:

a) Step "An Information Criterion" AIC method, the lesser AIC is, the less "over fitting" (Akaike 1976):

I. ***Backward approach:***

Regression model Invol ~ Race + Fire + Theft + Age + Income starts with AIC= 4.6885:

	Df	Sum of Sq	RSS	AIC
<none>			3.585335	4.688515
Race	1	0.629120	4.214455	5.133772
Fire	1	3.085206	6.670541	7.589858
Theft	1	0.201190	3.786525	4.705842
Age	1	0.293351	3.878686	4.798003
Income	1	0.007649	3.592984	4.512301

Since Income's AIC = 4.51 < AIO model's AIC 4.69, Variable log2(income) should be removed; new fitted as Invol ~ Race + Fire + Theft + Age with AIC= 4.5123:

	Df	Sum of Sq	RSS	AIC
<none>			3.592984	4.512301
Race	1	1.120840	4.713824	5.449277
Fire	1	3.194330	6.787315	7.522768
Theft	1	0.218066	3.811050	4.546503
Age	1	0.362397	3.955381	4.690835

No variables' AIC is smaller than 5.7909. Hence the model containing Race, Fire, Theft and Age is de facto the best fit after eliminating Income for the sake of less linearity. Coefficients:

(Intercept)	Race	Fire	Theft	Age
-0.2678703	0.006489357	0.04905728	-0.005809145	0.004687798

II. ***Forward approach:***

Regression model Invol ~ 1 (only intercept), with AIC = 19.1309:

	Df	Sum of Sq	RSS	AIC
<none>			18.29911	19.13089
Race	1	9.58226	8.71685	10.38041
Fire	1	13.34416	4.95495	6.61851
Theft	1	0.72497	17.57414	19.23769
Age	1	4.04815	14.25096	15.91452
Income	1	9.66611	8.63300	10.29656

Since Fire's AIC = 6.6185 < AIO model's AIC 19.1309 and stays smallest, Variable Fire should be added; new fitted as Invol ~ Fire with AIC= 6.6185:

	Df	Sum of Sq	RSS	AIC
<none>			4.954954	6.618510
Race	1	0.9121498	4.042805	6.538138
Theft	1	0.0132249	4.941730	7.437063
Age	1	0.1790148	4.775940	7.271273
Income	1	0.7088172	4.246137	6.741471

Since Race's AIC = 6.5381< AIO model's AIC 6.6185 and stays smallest, Variable Race should be added; new fitted as Invol ~ Fire + Race with AIC = 6.5381:

	Df	Sum of Sq	RSS	AIC
<none>			4.042805	6.538138

Theft 1 0.0874233 3.955381 7.282492
Age 1 0.2317545 3.811050 7.138161
Income 1 0.1232439 3.919561 7.246672
No variables' AIC is larger than 6.5381. Hence the model containing Race and Fire is de facto the best fit after adding Income for the sake of less linearity.
Coefficients:
(Intercept) Fire Race
-0.1913249 0.0546643 0.005712017

III. ***Stepwise approach gives the same result as backward.***

- Hence within these three method, two-variable model's AIC = 6.5381 > AIC= 4.6885 of four-variable model; the fitted model with only Race, Fire, Theft and Age shall be accepted.

b) Step "Bayesian Information Criterion" (BIC) method, the lesser absolute value of BIC is, the less "over fitting" (Tamura 1991); procedure similar to AIC method:

Regression model involct ~ race + fire + theft + age + Income, with BIC = -91:

Df Sum of Sq RSS BIC
Income 1 0.00765 3.5930 -94.712 ---------- Eliminate log2(income)
Theft 1 0.20119 3.7865 -92.351
Age 1 0.29335 3.8787 -91.269
<none> 3.5853 -91.002
Race 1 0.62912 4.2145 -87.533
Fire 1 3.08521 6.6705 -66.870
Regression model involct ~ race + fire + theft + age, with BIC = -94.71:

Df Sum of Sq RSS BIC
theft 1 0.2181 3.8111 -95.867----- Eliminate Theft
<none> 3.5930 -94.712
age 1 0.3624 3.9554 -94.195
race 1 1.1208 4.7138 -86.301
fire 1 3.1943 6.7873 -69.896

Regression model involct ~ race + fire + age, with BIC = -95.87:
Df Sum of Sq RSS BIC
age 1 0.2318 4.0428 -97.018--------- Eliminate Age
<none> 3.8111 -95.867
race 1 0.9649 4.7759 -89.518
fire 1 3.2632 7.0743 -71.839
Regression model involct ~ race + fire, with BIC = -97.02:
Df Sum of Sq RSS BIC-----Can't eliminate anymore
<none> 4.0428 -97.018
- race 1 0.9121 4.9550 -91.669
- fire 1 4.6741 8.7169 -66.250
Hence the model containing Race and Fire is de facto the best fit after adding Income for the sake of less linearity.
Coefficients: (Intercept) race fire
-0.191325 0.005712 0.054664

a) Adjusted R^2 Selection:
As a result, it is pretty obvious that model contains 4 variables as Race, Fire, Theft and Age is the best fitted for adjusted R^2 Selection, given that for selecting models within the range of sizes, the higher adjusted R^2 is, the better dependent variable "explained" (Harel 2009) by independent variables.

b) Mellow's C_p Selection:
Judging from the CP Plot, Model 1234 has the smallest Cp value, i.e. Model with Race, Fire, Theft and Age is the best fitting model according to Mellow's C_p Selection (Yu 2000); meanwhile model 124 (containing only Race, Fire and Age) is not bad too, as second runner up.

Result

So far, instead of one-model-to-rule-them-all, there are three competing variable selection options:

R	F	T	A	I	R-Sq Adj	R-Sq	RSE	AIC	BIC	race-pvalue	AIC	BIC	R2	CP
x	x	x	x		0.7840	0.8037	0.3000	25.95890	36.79887	0.00110	x		x	x
x	x		x		0.7765	0.7917	0.3049	26.61037	35.64368	0.00250			x	
x	x				0.7686	0.7791	0.3103	27.26690	34.49355	0.00366		x		

I.e. Invol = -0.2678703 + 0.006489357 Race + 0.04905728 Fire-0.005809145 Theft+ 0.004687798 Age;
Invol = -0.353968 + 0.005885 Race + 0.049547 Fire + 0.003566 Age;
Invol = -0.191325 + 0.005712 Race + 0.054664 Fire.

a) Race DOES seem to be significant at 5% level of all of them by P-Value;
b) While controlling other variables, such as Theft rate and Income, Race still maintains a quite strong (positive) relationship with Involuntary insurance rate;
c) Fire rate is the most substantial factor in regression model, so that insurance companies DID consider hard cold facts prior to racial and ethnical background;
d) All models above did a relatively good job at independent-dependent relationships explanatory coverage and avoiding over fitting simultaneously;

In all, "redlining" existed, it might play some roles in law suit or political campaign; Nonetheless, the reason behind such "initial selection" (actuarial term here) was legitimate, at least from statistic side.

Further exploration and potential flaws of analysis

The whole collection of data was based on Zip code, which was neglected from the early work of this article. Since it is not individual but aggregated, other potential important factors were omitted. Say, past insurance information, tornado, and rental. Other factors correlated with existing factors were unknown, either.

Furthermore, it is possible that the proximity of one Zip code to another sharing similar economic experience affects the dependency on observation; in opposite, the assumption of even proportions of FAIR plans across those Zip codes could be off the mark (Sutter 2009) – either way, more work need to be done such as division of data into suburbs, communities or streets. Those groups fitting ecological fallacy would be more homogenous (Kennedy 1998) thus helpful.

Note: Something really interesting discovered when checking the map attached – Once we split data into north (n = 24) and south (n = 21) according to the map and column one Zip: P values of Race are equal to 0.00603 and 0.0873 accordingly. That means, after "diluting" the data into smaller subsets, race can no longer be significant in some way.

Last but not the least, sizes of both data set (n = 47) and effects (maximum response value is 2.2%) from predictor variables are not large at all; discrimination or not, few people would be affected at statistical base. Until larger sample set is applied, Generalized Linear Modelling (GLM) cannot be used to explore exactly which variable is significant.

Furthermore, as banned officially since the 70s (Grogan & Proscio, 2000) although Redlining may still goes in a less overt way, it can be nowhere pervasive under a multi-value, modern and dynamic world. Different race, religion, sex orientation, or disability, people all over the world are protected policy wise under variegated legislations. Company needs more than numbers to justify its policy design. Say, explanations will be examined fastidiously by LGBT communities in spite of a higher AIDS rate as collinearity.

The bottom-line is no right or wrong here. Both demand and supply in insurance industry surely have their views and needs. When they match, it is a business; when they are not on the same page, there must be a reason behind this. Business is anything but charity and moral. Someone's whole life could be completely screwed up under some other's single line of code in S-Plus; life has never been easy. As said by *Enoch Thompson* from *Boardwalk Empire, "We all have to decide how much sin we can live with".*

References

Akaike, H. (1976). An information criterion (AIC). *Math Sci*, *14*(153), 5-9.

Badain, D. I. (1980). Insurance redlining and the future of the urban core.*Colum. JL & Soc. Probs.*, *16*, 1.

Calmore, J. O. (1997). Race/ism Lost and Found: The Fair Housing Act at Thirty. *U. Miami L. Rev.*, *52*, 1067.

Cochrane, J. H. (1991). A simple test of consumption insurance. *Journal of political economy*, 957-976.

Derksen, S., & Keselman, H. J. (1992). Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*,*45*(2), 265-282.

Grogan, P. S., & Proscio, T. (2000). *Comeback cities: A blueprint for urban neighborhood revival*. Basic Books.

Harel, O. (2009). The estimation of R 2 and adjusted R 2 in incomplete data sets using multiple imputation. *Journal of Applied Statistics*, *36*(10), 1109-1118.

Joseph P. Newhouse, & Rand Corporation. Insurance Experiment Group (Eds.). (1993). *Free for all?: lessons from the RAND health insurance experiment*. Harvard University Press.

Kennedy, B. P., Kawachi, I., Glass, R., & Prothrow-Stith, D. (1998). Income distribution, socioeconomic status, and self rated health in the United States: multilevel analysis. *Bmj*, *317*(7163), 917-921.

Rice, W. E. (1996). Race, Gender, Redlining, and the Discriminatory Access to Loans, Credit, and Insurance: An Historical and Empirical Analysis of Consumers Who Sued Lenders and Insurers in Federal and State Courts, 1950-1995. *San Diego L. Rev.*, *33*, 583.

Smink, D. S., Fishman, S. J., Kleinman, K., & Finkelstein, J. A. (2005). Effects of race, insurance status, and hospital volume on perforated appendicitis in children. *Pediatrics*, *115*(4), 920-925.

Sutter, D. (2009). *Policy uncertainty and the market for wind insurance*. Mercatus Center, George Mason University.

Tamura, Y., Sato, T., Ooe, M., & Ishiguro, M. (1991). A procedure for tidal analysis with a Bayesian information criterion. *Geophysical Journal International*, *104*(3), 507-516.

Wriggins, J. B. (2010). Automobile Injuries as Injuries with Remedies: Driving, Insurance, Torts, and Changing the Choice Architecture of Auto Insurance Pricing. *Loy. LAL Rev.*, *44*, 69. 1

Yu, C. H. (2000). An overview of remedial tools for collinearity in SAS. In *Proceedings of 2000 Western Users of SAS Software Conference* (1),196-201.