Auto Insurance Premium Pricing Model

John Wensink

MIS 480 Senior Capstone

Colorado State University-Global- Campus

Chris den Heijer

January 10, 2020

Auto Insurance Premium Pricing Model

Hello and thank you for your time. As requested, my organization has completed our analysis of a sample of your company’s claims data for the purpose of creating a model to calculate premium accurately based on a set of defined risk factors. We are pleased to submit this analysis to you as a business intelligence solution that your underwriting department can begin testing today. As a brief summary of our team’s work, data from N=1000 insured member’s first reported auto claims were sampled at random from a population of your company’s auto damage claims. Historical customer profile data was added to each damage claim to examine for possible correlations with eight specific risk factors. Continuous variables including a customer’s age, the year of the insured member’s vehicle which was involved in an accident, a member’s umbrella policy limits, we also looked at the duration of a member’s policy before his or her first accident. We examined an insured member’s sex, education level, relationship status, the model of the insured vehicle, and policy base state for possible categorical/boolean variables. For all tests, the dependent variable was the claim’s total Dollar amount payout. The factors present in this bottom-line payout amount are from three sources; damage to the insured vehicle, damage to other’s personal property, and bodily injury claims/attorney fees as a result of our insured member’s negligence. SAS University Edition (SUE) was used under an educational license to perform descriptive analytics on the data sample, including Multi-Factor Analysis of Variance (ANOVA), Principal Component Analysis (PCA), T-Tests, as well as their graphical presentations. For this analysis, we set significance level α=0.05, and state our null hypothesis as **there is no correlation between a risk factor and the claim’s total payout amount.**

**Dataset**

To get started, our team began the data familiarization process. We ran summary statistics (SAS, 2020) on the data with total claim amount as the analysis variable and the risk factors as classification variables. The first insight gleaned from this was that there were obviously too many auto models and not enough observations to deliver any meaningful insights so the risk factor auto model was replaced with the risk factor auto make. Another insight delivered through first impressions of the data was that payout amounts for claims for bodily injury were orders of magnitude higher than simple claims for property damage and damage to the insured vehicle. Our team decided to normalize the total claim payout using logarithmic scaling so that the distribution of claim payments more closely matches that of the potential risk factors.

data work.transform;

set WORK.IMPORT;

tr1\_total\_claim\_amount=LOG(X);

run;

A relatively simple, but important transformation, we will now run our queries on the resultant table:

PROC SQL;

CREATE TABLE WORK.query AS

SELECT months\_as\_customer , age , policy\_number , policy\_bind\_date , policy\_state , policy\_csl , policy\_deductable , policy\_annual\_premium , umbrella\_limit , insured\_zip , insured\_sex , insured\_education\_level , insured\_occupation , insured\_hobbies , insured\_relationship , capital\_gains , capital\_loss , incident\_date , incident\_type , collision\_type , incident\_severity , authorities\_contacted , incident\_state , incident\_city , incident\_location , incident\_hour\_of\_the\_day , number\_of\_vehicles\_involved , property\_damage , bodily\_injuries , witnesses FROM WORK.'TRANSFORM'n;

run;

Now that we have done some exploratory data analysis and normalized our tables we are ready to get to work. As our null hypothesis that there is no relationship between the dependent variable (total payout) and a possible independent variable, we are looking to use univariate and multivariate ANOVA to calculate that the probability of observing the F-statistic that is at least as high as our study calculated. As such, a higher F Value and lower Pr > F shows a stronger relationship between the variables. For this study, as we have a relatively small sample size, we will set our confidence level to the standard alpha = 0.05, with a confidence interval of 95%. As observed, we must discard five of the eight indicators of risk which had interested us, as they are below the threshold for statistical significance at confidence level alpha = 0.05. That being said we are confident that Age, Sex, and Education Level will be significant predictors of a claim’s payout dollar amount.



From this ANOVA we see that age, gender, and education level are the more influential factors in this sample’s total claims payout. As such these are the factors that we will further explore.

Age is likely to be the most important contributing factor in risk as it applies to insurance actuarial analysis. To model this, we will use a simple linear regression model with claim total payout as the dependent variable and age as the continuous variable:

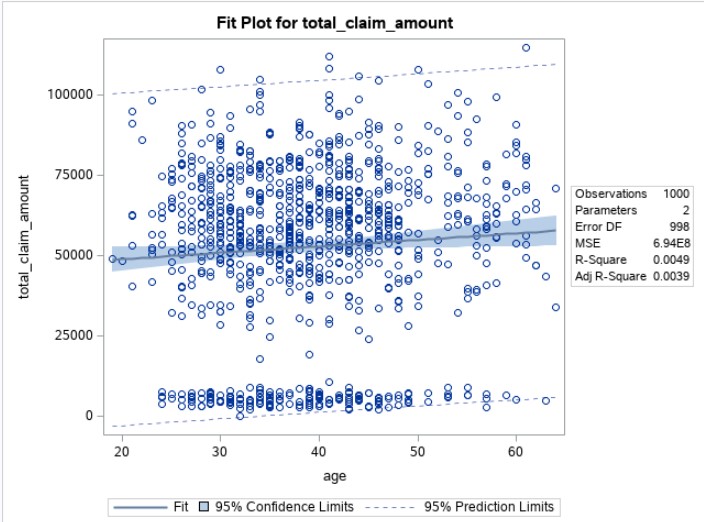
proc reg data=WORK.IMPORT alpha=0.05 plots(only)=(diagnostics residuals fitplot

observedbypredicted);

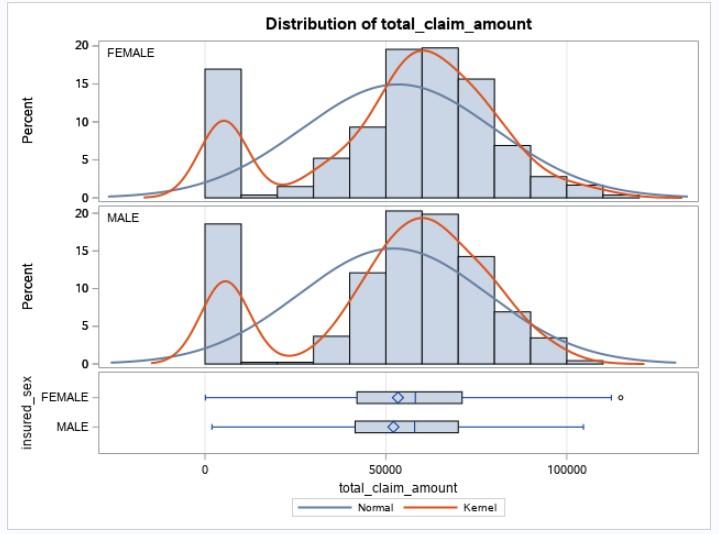
model total\_claim\_amount=age /;

run;

which yields a regression slope of approximately y = 133x + 47,300. This can be applied to our model as a 1.33% surcharge for drivers older than 43 years old for each year they age, compounded yearly. Ideally, we would like for our sample to be statistically significant before applying premium surcharges in real life.

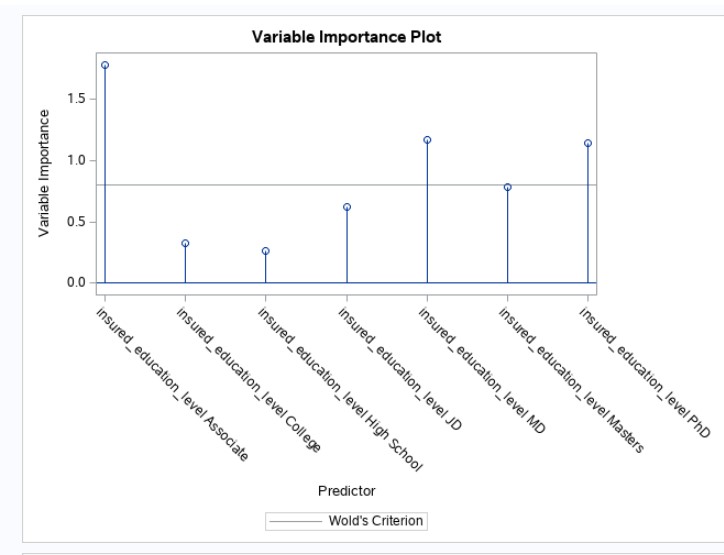


Moving on to the area of an insured member’s sex correlating as a risk factor, we went with a T-Test (University of California, 2020) to examine for correlation between the insured member’s sex versus the total claim payout yielded. which produced some interesting results.



The results of the two-sample T-Test did not show a significant relationship between the sex of the insured member and the claim’s total payout. As such no differentiation will be made between male and female applicants

Moving on to the third piece of our model, education level, our team analyzed total claim payout as the dependent variable and examined how an insured member’s level of education contributes to his or her overall risk profile, all else being equal. To this end, we utilized a partial least squares regression model (SAS, 2012) to determine the importance of different levels of education as they are a factor of the sample’s total claims payout. The results were not what we expected but we do believe there is actionable intelligence here that may say more about our population than is obvious through intuition.



Here we see a variable importance plot with the different levels of education on the independent axis. Surprisingly, the education level of an associate’s degree was discovered to have the highest variable importance score for projection. This means that a measure of the contribution of the level of education as it is described by the variance of each component. Rather than penalizing people for education levels that may not intuitively make sense, we should reward customers with a sliding scale of premium deductions. Skipping high school and starting with a college degree the customer will receive a 0.5% premium credit for an associate’s or a bachelor’s level degree. Masters degrees and Juris Doctors will be entitled to a 1.0% reduction in premium, with MD’s and Ph.D. 's receiving the maximum 1.5% reduction in premium.

**Implementation**

Now that we’ve weighed variables from inferences gained in SAS, it is time to build a premium pricing model where a user can input his or her personal data in an in-script questionnaire. A rate will be quoted, and the user’s data will be uploaded to the data warehouse setup on a PostgreSQL server. The existing training data will live there and new rows will be added using an ETL process, possibly with the aid of Pentaho, although that is a colossal file and it may not be necessary as the questionnaire will be a simple page with simple data types like tinyint, date, and strings. Creating a model in Python will be a constantly evolving model as our dataset becomes more robust. Vineet & Jannos (2017) describe a program on our organization shall imitate. Although the scope of this project will be limited, if our model evolves through testing and meets the needs of the organization, our shortage of what could be modeled in an object-based language like Python, the only limitation would be the quality of data and the programmer’s creativity.

References

SAS. (2012). Partial Least Squares Regression. Retrieved January 10, 2021, from <https://www.sas.com/content/dam/SAS/en_ca/User%20Group%20Presentations/TASS/Cai-PLSRegression.pdf>

SAS. (2018). N-Way ANOVA Task: Building a Model. Retrieved December 14, 2020, from <https://documentation.sas.com/?cdcId=webeditorcdc>

SAS. (2020). Creating Summary Statistics for a Table. Retrieved December 14, 2020, from <https://documentation.sas.com/?docsetId=etlug>

Shah, B. (2018, August 20). Auto Insurance Claims Data. Retrieved December 14, 2020, from <https://www.kaggle.com/buntyshah/auto-insurance-claims-data>

The University of California. (2020). HOME. Retrieved December 14, 2020, from <https://stats.idre.ucla.edu/sas/output/proc-ttest/>

Vineet, & Jannos. (2017, September 16). Insurance premium calculator exercise. Retrieved December 14, 2020, from <https://codereview.stackexchange.com/questions/175816/insurance-premium-calculator-exercise>