ISO

And The Winner Is…? Picking a Better Model – Part 1

2018 CAS RPM Seminar

SERVE | ADD VALUE | INNOVATE

Antitrust Notice

- The Casualty Actuarial Society is committed to adhering strictly to the letter and spirit of the antitrust laws. Seminars conducted under the auspices of the CAS are designed solely to provide a forum for the expression of various points of view on topics described in the programs or agendas for such meetings.
- Under no circumstances shall CAS seminars be used as a means for competing companies or firms to reach any understanding – expressed or implied – that restricts competition or in any way impairs the ability of members to exercise independent business judgment regarding matters affecting competition.
- It is the responsibility of all seminar participants to be aware of antitrust regulations, to prevent any written or verbal discussions that appear to violate these laws, and to adhere in every respect to the CAS antitrust compliance policy.

CAS

ISO

© 2018 Insurance Services Office, Inc. All rights reserved. 2

Motivation

- Fit may not be good enough
- Fit may be too good to be true (training, test, holdout data)
- Results may not be stable across data subsets or over time
- Wrong distribution may have been chosen
- Results may be highly influenced by some records
- Model may underperform the current rating plan

© 2018 Insurance Services Office, Inc. All rights reserved. 3

Understanding and Validating a Model

- Model Lift
  - Differentiating between best and worst risks
  - Preventing adverse selection
  - Improving the rating plan
- Goodness of fit
  - Statistics, residual plots, actual versus predicted
- Internal Stability
  - Different data or time period
  - Reliability of parameter estimates

© 2018 Insurance Services Office, Inc. All rights reserved.

Graphical Representations of Model Lift

- Gini index plots
- Simple quantile plots (lift charts)
- Double quantile plots (double lift charts)
- Loss ratio charts

© 2018 Insurance Services Office, Inc. All rights reserved.

Gini Index Applied to Insurance Data

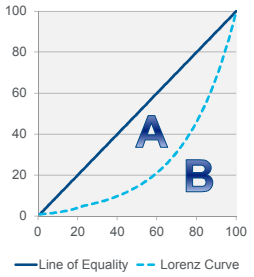
Binary Response
- SAS Proc Logistic
- t = total pairs with different responses
- n_c = concordant pairs
- n_d = discordant pairs
- $t - n_c - n_d$ = tied pairs
- Sommer's D
  = Gini's coefficient
  = $(n_c - n_d) / t$

Lorenz Curve for non-binary insurance response
- Sort holdout data by predicted value and random number
- Calculate cumulative percentages for insurance measure and exposure
- Plot
  - Cumulative exposure in horizontal axis
  - Cumulative insurance measure in vertical axis

© 2018 Insurance Services Office, Inc. All rights reserved.

Gini Index Applied to Insurance Data

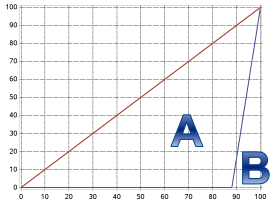
- Measure of how well model classifies risks based on expected pure premium
  - Sort holdout data by predicted pure premium and random number
  - Horizontal axis = cumulative percentage of earned car years
  - Vertical axis = cumulative percentage of reported loss
  - A = Area between line of equality and Lorenz Curve
  - B = Area beneath Lorenz Curve
  - Gini index = A / (A + B)



© 2018 Insurance Services Office, Inc. All rights reserved.

Lorenz Curve — Saturated (Completely Overfit) Model

- Simulated frequency = 12%
- Prediction = Actual
- Sort training data set by model prediction.
- Horizontal axis = percentage of total exposure.
- Vertical axis = Percentage of total claims.
- Gini Index = A / (A+ B) is very high.



© 2018 Insurance Services Office, Inc. All rights reserved.

Lorenz Curve, Gini Index

- Exercise:
  - Model X prediction = expected loss cost
  - Model Y prediction = 0.5 (expected loss cost)
  - Model Z prediction = 2.0 (expected loss cost)
  - Which model has the highest Gini index?

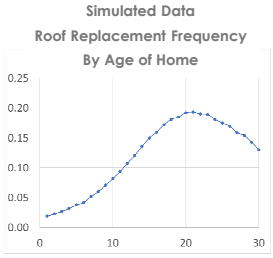
- Model A has a Gini index of 15.9 and B has 15.4
  - Is that difference significant?

© 2018 Insurance Services Office, Inc. All rights reserved.

Simulated Data

Class	Exposure	Claims	Freq.
0	423,294	136,290	0.3220
1	1,076,706	42,745	0.0397

Age	Exposure	Claims	Freq.
1	52,621	986	0.0187
2	52,643	1,192	0.0226
3	52,692	1,443	0.0274
etc.			

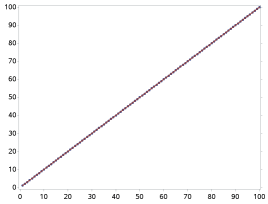


© 2018 Insurance Services Office, Inc. All rights reserved.

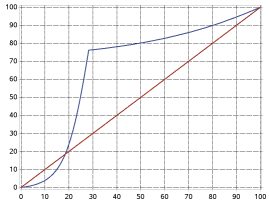
Lorenz Curve — Be Very, Very Careful!

Include a random sort key

Lorenz Curve for Simulated Data
In Random Order



Lorenz Curve for Simulated Data
In Roof Class and Home Age Order



© 2018 Insurance Services Office, Inc. All rights reserved.

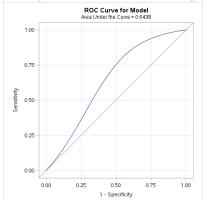
Simulated Data and Logistic Regression Diagnostics

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.0010	0.008090	137674.93	<.0001
age	1	0.0595	0.000399	22186.78	<.0001

Association of Predicted Probabilities and Observed Responses

Percent Concordant	62.8	Somers' D	0.288
Percent Discordant	34.0	Gamma	0.297
Percent Tied	3.2	Tau-a	0.061
Pairs	85289704045	c	0.644

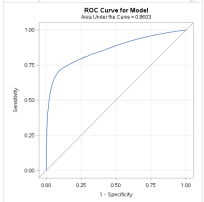


Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.8427	0.00978	84525.0302	<.0001
age	1	0.1652	0.000623	70232.7406	<.0001
group	1	-3.8578	0.0107	129502.583	<.0001

Association of Predicted Probabilities and Observed Responses

Percent Concordant	85.5	Somers' D	0.721
Percent Discordant	13.5	Gamma	0.726
Percent Tied	1.0	Tau-a	0.152
Pairs	85289704045	c	0.860

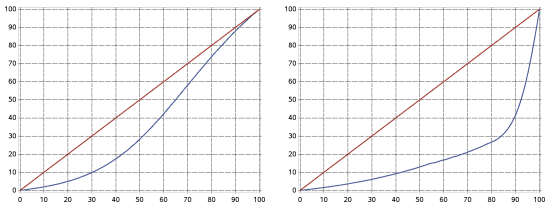


© 2018 Insurance Services Office, Inc. All rights reserved.

Lorenz Curve For Simulated Data and Logistic Regression

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	62.8	Somers' D	0.288
Percent Discordant	34.0	Gamma	0.297
Percent Tied	3.2	Tau-a	0.061
Pairs	85289704045	c	0.644

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	85.5	Somers' D	0.721
Percent Discordant	13.5	Gamma	0.728
Percent Tied	1.0	Tau-a	0.152
Pairs	85289704045	c	0.860



© 2018 Insurance Services Office, Inc. All rights reserved.

Lift Charts or Quantile Plots

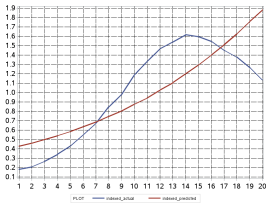
- Creating a quantile plot
  - Use holdout sample.
  - Sort data based on predicted values and random number.
  - Subdivide sorted data into quantiles with equal weight.
    - Use exposure weights for frequency and pure premium.
    - Use claim count weight for severity.
  - Calculate average actual value and average predicted value for each quantile and index to their overall average.

© 2018 Insurance Services Office, Inc. All rights reserved.

Lift Charts or Quantile Plots for Simulated Data

Bin	Exposure	Actual Claims	Pred Claims	Actual	Pred
1	30,004	642	1,530	0.15	0.43
2	30,004	742	1,647	0.21	0.46
3	30,004	957	1,789	0.27	0.50
4	30,004	1,220	1,937	0.34	0.54
5	30,004	1,553	2,096	0.43	0.58
6	30,004	1,955	2,277	0.55	0.63
7	30,004	2,386	2,456	0.67	0.68
8	30,004	3,021	2,676	0.85	0.75
9	30,004	3,497	2,882	0.98	0.80
10	30,005	4,240	3,142	1.19	0.88
11	30,004	4,754	3,382	1.33	0.94
12	30,004	5,237	3,683	1.47	1.03
13	30,004	5,496	3,965	1.54	1.11
14	30,004	5,773	4,306	1.62	1.20
15	30,004	5,701	4,640	1.60	1.29
16	30,004	5,525	5,021	1.55	1.40
17	30,004	5,181	5,417	1.45	1.51
18	30,004	4,920	5,831	1.38	1.63
19	30,004	4,532	6,300	1.27	1.76
20	30,005	4,060	6,743	1.14	1.88
All	600,082	71,382	71,720	1.00	1.00

Model using Home Age only

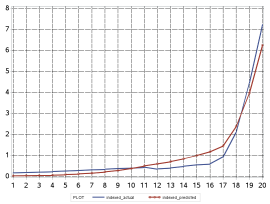


© 2018 Insurance Services Office, Inc. All rights reserved.

Lift Charts or Quantile Plots for Simulated Data

Bin	Exposure	Actual Claims	Pred Claims	Actual	Pred
1	30,004	532	51	0.15	0.01
2	30,004	599	78	0.17	0.02
3	30,004	688	117	0.19	0.03
4	30,004	752	171	0.21	0.05
5	30,004	842	250	0.24	0.07
6	30,004	925	357	0.26	0.10
7	30,004	1,076	503	0.30	0.14
8	30,004	1,153	700	0.32	0.20
9	30,004	1,278	961	0.36	0.27
10	30,005	1,366	1,304	0.38	0.36
11	30,004	1,527	1,730	0.43	0.48
12	30,004	1,242	2,097	0.35	0.58
13	30,004	1,400	2,484	0.39	0.69
14	30,004	1,666	2,962	0.47	0.83
15	30,004	1,904	3,513	0.53	0.98
16	30,004	2,064	4,139	0.58	1.15
17	30,004	3,257	5,143	0.91	1.43
18	30,004	7,492	8,484	2.10	2.37
19	30,004	15,856	14,139	4.44	3.94
20	30,005	25,763	22,288	7.22	6.22
All	600,082	71382	71,471	1.00	1.00

Model: Roof Class & Home Age



© 2018 Insurance Services Office, Inc. All rights reserved.

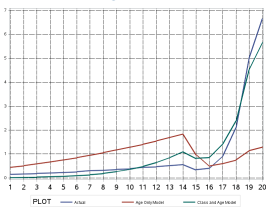
Double Lift Charts

- Creating a double lift chart
  - Sort holdout data by
    - Ratio of model prediction to competing model
    - Random number.
  - Subdivide sorted data into quantiles with equal exposure.
  - For each quantile calculate average actual value, average predicted loss cost and average loss cost underlying the current manual premium or competing model prediction.
  - Index the quantile averages to the overall averages.

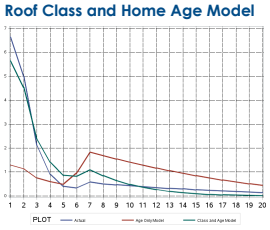
© 2018 Insurance Services Office, Inc. All rights reserved.

Double Lift Chart For Simulated Data

Sorted by Ratio of Roof Class and Home Age Model To Home Age Only Model



Sorted by Ratio of Home Age Only Model to Roof Class and Home Age Model



© 2018 Insurance Services Office, Inc. All rights reserved.

Code Used for Data Simulation

```
set.seed(3)
m=1500000
d <- data.frame(year=c(rep(2017, m)), policy_number=c(1:m), exposure=rep(1, m))
p <- ( c(rep(30,30)) - seq(from=0, to=3, length.out = 30) )
p <- p / sum(p)
d$age <- sample(c(1:30), m, replace = TRUE, prob=(p))
p <- 0.50 - ( d$age / 70 )
random_draw <- runif(m, min=0, max=1)
d$class <- ifelse( random_draw < p, 0, 1 )
d_lp <- -4.00 - 0.10 * d$class + 0.25 * d$age - 0.20 * d$class * d$age
expected_claim_prob <- exp(d_lp) / ( exp(d_lp) + 1 )
random_draw <- runif(m, min=0, max=1)
d$claim_ind <- ifelse( random_draw < expected_claim_prob, 1, 0 )
random_draw <- runif(m, min=0, max=1)
d$sample <- ifelse( random_draw < 0.6, "Train", "Test" )
```


© 2018 Insurance Services Office, Inc. All rights reserved.

References

- Goldburd, M., Khare, A., and Tevet, D., Generalized Linear Models for Insurance Rating, Casualty Actuarial Society, 2016
- SAS Institute Inc. 2017. SAS/STAT® 14.3 User's Guide. Cary, NC: SAS Institute Inc.
- De Jong, P. and Heller, G. Z., Generalized Linear Models for Insurance Data, Cambridge University Press, 2008
- Dickey, D. A., "Finding the Gold in Your Data: An Overview of Data Mining", SAS Global Forum 2013
- Frees, E.W., Derrig, R. A., and Meyers, G., Predictive Modeling Applications in Actuarial Science, Cambridge University Press, 2014
- Jaffery, T. and Liu, S. X., "Measuring Campaign Performance by Using Cumulative Gain and Lift Chart", SAS Global Forum, 2009

© 2018 Insurance Services Office, Inc. All rights reserved.

This material was used exclusively as an exhibit to an oral presentation.
It may not be, nor should it be relied upon as reflecting, a complete record of the discussion.

Contact: Hernan.Medina@Verisk.com

ISO

© 2018 Insurance Services Office, Inc. All rights reserved.


ISO