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.

Model Validation

How Well Do GLMs Fit?

Panelists: Dr. Paul Beinat, Research and Development, EagleEye Analytics

Agenda

- Validation
- Using 2 US portfolios
- Fit Frequency GLMs
  - Fit better than severity – Poisson with Log link
  - GLMs position of strength
- Indicators of Fit
  - Statistical inference
- Validation as Indicator of fit
- Validation using another method
  - General Iteration Algorithms
- Conclusion

Validation

- Data kept aside from modeling
- A test for
  - Model generalization
  - Model prediction – predictive analytics
- Measure how well the model does
- General rule for data
  - 2/3 training, 1/3 validation
- Bayesian view
  - Maximum likelihood leads to overfitting

Portfolio 1

- PPA collision coverage
  - Small portfolio – typical of many companies
- Fit best GLM via statistical inference
  - Explore main effects variables using all combinations
  - Chi squared test for including variable
  - Common design matrix – nested GLMs
  - Optimize deviance
  - 2 years for training, 1 year for validation

Model Statistics

Vars	Coef.	StdErr	z	P>|z|	[95% Conf. Interval]	
Constant	-1.834	0.0261	-70.3916	0	-1.8851	-1.7829
A_0	-0.1295	0.0403	-3.2103	0.0013	-0.2086	-0.0504
MS_m	-0.1346	0.0242	-5.557	0	-0.1821	-0.0871
M_y	-0.247	0.0241	-10.2632	0	-0.2942	-0.1999
Sf	0.0604	0.0229	2.6341	0.0084	0.0155	0.1054
T_-1	0	0	0	0	0	0
T_14	0.0856	0.037	2.3129	0.0207	0.0131	0.1582
T_16	0.125	0.0374	3.3403	0.0008	0.0516	0.1983
T_18	0.1622	0.043	3.7753	0.0002	0.078	0.2465
T_19	0.1658	0.0402	4.1194	0	0.0869	0.2446
T_20	0.2092	0.038	5.5005	0	0.1346	0.2837
V_15	0.0649	0.0255	2.5476	0.0108	0.015	0.1148
V_20	0.1305	0.0473	2.7606	0.0058	0.0379	0.2232
W_-1	0	0	0	0	0	0
W_0	0.1957	0.0467	4.1886	0	0.1041	0.2873
W_15	-0.0591	0.0241	-2.4548	0.0141	-0.1064	-0.0119

Deviance	=	956.9280
Pearson Stat	=	221.1284
AIC	=	0.0596
BIC	=	-568007.1651
Efron pseudo-R2	=	0.2333
McFadden index	=	0.0942

Good P values

Standard errors a little wide, but small data set

Note Efron measure – it’s a weighted one

Validation

Model		
Deviance	=	956.9280
Pearson Stat	=	221.1284
AIC	=	0.0596
BIC	=	-568007.1651
Efron pseudo-R2	=	0.2333
McFadden index	=	0.0942

Validation		
Deviance	=	1997.0177
Pearson Stat	=	633006.3978
AIC	=	0.0808
BIC	=	-557038.8497
Efron pseudo-R2	=	-0.2391
McFadden index	=	-0.2304

- Deviance is about double
  - But not comparable
- Efron is negative
  - Assumes Gaussian errors
  - Says that it does worse than just the mean
- Why so bad on validation?
  - Frequency drift between years – Efron was useful
  - Fix it by splitting data at random 2/3 to 1/3 for training and validation
  - Refit the models

New Portfolio 1 Model

Vars	Coef.	StdErr	z	P>|z|	[95% Conf. Interval]	
constant	-2.1733	0.0354	-61.3438	0.0000	-2.2427	-2.1038
A0	-0.0879	0.0318	-2.7668	0.0057	-0.1501	-0.0256
A10000	0.1416	0.0271	5.2159	0.0000	0.0884	0.1948
A12500	0.2736	0.0392	6.9776	0.0000	0.1967	0.3504
A15000	0.2613	0.0446	5.8618	0.0000	0.1739	0.3487
MM	-0.1306	0.0230	-5.6800	0.0000	-0.1756	-0.0855
MY	-0.2600	0.0232	-11.2044	0.0000	-0.3054	-0.2145
N-1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
N6	-0.0707	0.0250	-2.8303	0.0047	-0.1196	-0.0217
N8	-0.0680	0.0294	-2.3166	0.0205	-0.1256	-0.0105
SF	0.0637	0.0221	2.8765	0.0040	0.0203	0.1071
T-1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
T10	0.0866	0.0386	2.2437	0.0249	0.0109	0.1622
T12	0.1380	0.0373	3.7015	0.0002	0.0649	0.2110
T14	0.1408	0.0395	3.5606	0.0004	0.0633	0.2183
T16	0.1621	0.0387	4.1906	0.0000	0.0863	0.2380
T18	0.1943	0.0452	4.2997	0.0000	0.1058	0.2829
T19	0.2117	0.0424	4.9963	0.0000	0.1286	0.2947
T20	0.2713	0.0391	6.9396	0.0000	0.1947	0.3479
V15	0.0713	0.0243	2.9380	0.0033	0.0237	0.1188
V20	0.1752	0.0439	3.9909	0.0001	0.0892	0.2613
W-1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
W0	0.2378	0.0417	5.7017	0.0000	0.1561	0.3195
W5	0.1866	0.0306	6.1077	0.0000	0.1267	0.2465

Deviance	=	6356.6198
Pearson Stat	=	5845503.9905
AIC	=	0.2081
BIC	=	-758078.9118
Efron pseudo-R2	=	0.0145
McFadden index	=	0.0350

Good P values

Good Zs

Look at the Efron

Validation

Model			Validation		
Deviance	=	6356.6198	Deviance	=	5007.7991
Pearson Stat	=	5845503.9905	Pearson Stat	=	8822293.6783
AIC	=	0.2081	AIC	=	0.2869
BIC	=	-758078.9118	BIC	=	-363860.2653
Efron pseudo-R2	=	0.0145	Efron pseudo-R2	=	0.0043
McFadden index	=	0.0350	McFadden index	=	0.0163

- The Deviance of validation is better than training
  - But validation has half the data
- What is going on with the Efron?
- Is this model good or not?
  - It had good model statistics!
  - But it does marginally better than just the mean!
- If deviance on validation is the measure of accuracy then
  - The best model is simpler than this one
  - It has a worse model deviance!
- Indicates this model may be overfitting

Validation Measure

- Deviance based
  - Uses Poisson error structure
  - But deviance varies from data set to data set and model to model
- Want a relative deviance measure

$$I = 1 - \frac{d_m}{d_0}$$

- 1 minus the deviance of the model divided by the deviance of the null model
  - The relative improvement over the null model

Validation

Model

Deviance	=	6356.6198
Pearson Stat	=	5845503.9905
AIC	=	0.2081
BIC	=	-758078.9118
Efron pseudo-R2	=	0.0145
McFadden index	=	0.0350

Validation

Deviance	=	5007.7991
Pearson Stat	=	8822293.6783
AIC	=	0.2869
BIC	=	-363860.2653
Efron pseudo-R2	=	0.0043
McFadden index	=	0.0163

- Model deviance improvement I = 7.5%
  - In deviance terms it does not much better than using just the average
  - Consistent with Efron message
- Validation – I = 3.2%
  - Less than half the signal present in validation
  - Model overfits
  - Consistent with Bayesian view
- Is it just this portfolio?

Homeowners

Vars	Coef.	StdErr	z	P>|z|	[95% Conf. Interval]	
Constant	-7.6551	0.0311	-245.8342	0.0000	-7.7162	-7.5941
e_0	1.2792	0.4478	2.8563	0.0043	0.4014	2.1569
e_25	0.8971	0.1870	4.7966	0.0000	0.5306	1.2637
e_50	0.3154	0.1009	3.1247	0.0018	0.1176	0.5132
e_75	0.3316	0.0808	4.1024	0.0000	0.1732	0.4901
t_42	0.4398	0.0327	13.4502	0.0000	0.3757	0.5039
t_41	0.6084	0.0358	17.0092	0.0000	0.5383	0.6785
t_43	-0.2101	0.0838	-2.5082	0.0121	-0.3743	-0.0459
t_4	0.1924	0.0751	2.5624	0.0104	0.0452	0.3396
t_31	0.2268	0.0684	3.3146	0.0009	0.0927	0.3610

Deviance = 2286.8363
Pearson Stat = 21.5525
AIC = 0.0003
BIC =-321306723.6643
Efron pseudo-R2 = 0.4449
McFadden index = 0.2774

Good P values

Good Zs

Much better Efron

Again looks like a good model

Validation

Model

Deviance	=	2286.8363
Pearson Stat	=	21.5525
AIC	=	0.0003
BIC	=	-321306723.6643
Efron pseudo-R2	=	0.4449
McFadden index	=	0.2774

Validation

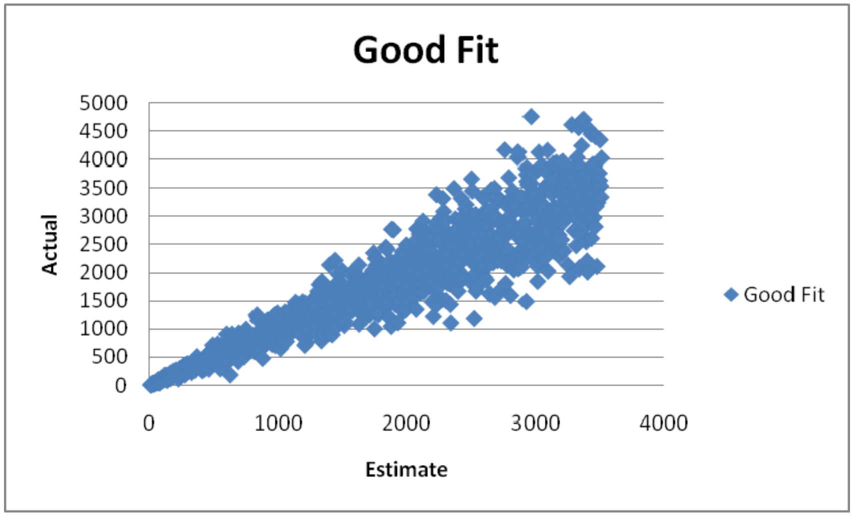
Deviance	=	4971.1931
Pearson Stat	=	15.6539
AIC	=	0.0004
BIC	=	-324074651.0975
Efron pseudo-R2	=	-0.6298
McFadden index	=	-0.2382

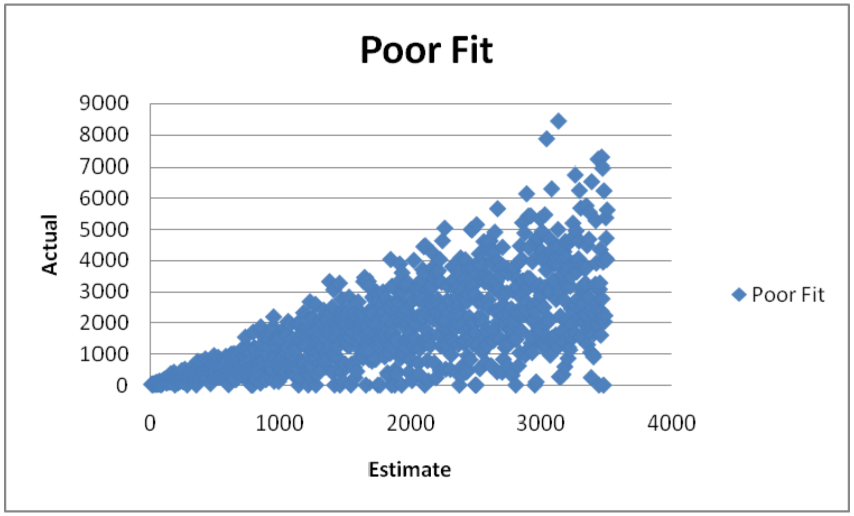
- Modeling on 1 year validation on the next year
- Validation
  - Efron is very negative
  - The mean would do much better than our model!
- Deviance improvement measure
  - Model – I = 49%
  - Validation – I = -40%
- It's a catastrophe
  - Driven by one variable - found on investigation
  - Model fitted on all the data does not show this problem!
  - Shows power of validation

What About Gini?

- Previously proposed as validation measure
- What are Gini measures of previous model
  - Training 0.2930
  - Validation 0.2453
  - Seems to indicate a good model
  - Divergent view to R^2 and deviance improvement measure deterioration
- Why?

Example





- Both have similar Gini values
- Very different fits
- Gini measures lift, not fit
- This can lead to adverse selection

Model without problem variable

Model

Deviance	= 281.6758
Pearson Stat	= 0.0502
AIC	= 0.0000
BIC	= -321309298.9510
Efron pseudo-R2	= 0.7623
McFadden index	= 0.5182

Validation

Deviance	= 479.7100
Pearson Stat	= 0.1340
AIC	= 0.0001
BIC	= -324079612.3232
Efron pseudo-R2	= 0.5667
McFadden index	= 0.4076

- Model
  - Efron improves from 0.449 to 0.762
  - Deviance improvement from I=49% to 75%
- Validation
  - Efron improves from -0.6298 to 0.5667
  - Deviance improvement from I=-40% to 58%
- This is a real model
  - Does much better than the mean
  - But model statistics no better than previous model!
  - Consistent overfitting - loses 25% of the signal on validation

Why?

- Bayesians predict this overfitting
- It can only arise from methodology
  - Maximum Likelihood
  - Poisson error structure
  - Log link
- Can these be relaxed and can something do better
  - GIAs
    - Minimum bias methods
    - Iteratively estimate relativities
    - Relax distribution and link constraints via P, Q and K parameters

Fu L and Wu P General Iteration Algorithm for Classification Ratemaking [Journal] // Variance. - [s.l.] : CAS. - 02 : Vol. 01.

Measures of Fit

- Weighted average bias(WAB)

$$\text{WAB} = \frac{\sum_i w_i |y_i - \hat{y}_i|}{\sum_i w_i}$$

- Weighted average relative bias (WAQB)

$$\text{WAQB} = \frac{\sum_i w_i \frac{|y_i - \hat{y}_i|}{\hat{y}_i}}{\sum_i w_i}$$

- Weighted Chi squared (WChi)

$$\text{WChi} = \frac{\sum_i w_i \frac{(y_i - \hat{y}_i)^2}{\hat{y}_i}}{\sum_i w_i}$$

- Composite measure
  - Weighted average bias and Chi squared
    - WABWChi

$$\text{WABWChi} = \sqrt{WAB \times WChi}$$

- Measured using design matrix

PPA Results

- Using validation measures
  - Data split at random 2/3, 1/3 training and validation
  - Same variables as GLM used

- 632 GIAs fitted (combinations of P, Q and K)
  - Using WAB GLM is 2nd best
  - Using WAQB GLM is 428th best
  - Using WChi GLM is 394th best
  - Using WABWChi GLM is 212th best
  - Relativities can be very different from GLM model

Homeowners Results

- Using validation measures
  - Data split at one year training and next year validation
  - Same variables as GLM used

- 632 GIAs fitted (combinations of P, Q and K)
  - Using WAB GLM is 22^{nd} best
  - Using WAQB GLM is 20^{th} best
  - Using WChi GLM is 7^{th} best
  - Using WABWChi GLM is 8^{th} best
  - Relativities can also be very different from GLM model

Generality of Results

- Used large data sets
  - Among others a top 5 insurer
- Experiments where GLMs fitted by consulting actuaries
  - Same results
- Results are an attribute of the method
  - Not of the data
  - Not of the analyst fitting the GLM

Conclusions

- Model statistics can be misleading
  - Validation indicates this
- GLM assumptions not optimal
  - GIAs demonstrate this
- Maximum likelihood
  - Log likelihood surface is very shallow
  - Maximum of surface dependent on idiosyncrasies
    - Not immune to squared error problems
  - Models overfit
    - Validation indicates this
    - Bayesians think they have a proof