Dr Frankstein Build The GLM What Could Go Wrong?

Dr Paul Beinat[1,2,3,4,5]

1. NeuronWorks
2. Centre for Quantum Computation and Intelligent Systems, UTS
3. Advanced Analytics Institute, UTS
4. Data Science Australia, UTS, UMelb, MonashU, UNSW, UQ
5. EagleEye Analytics

Agenda

- The "systematic/random" conjecture
- Experimental design and Results
- Linear as in GLM
- When It's Not Poisson
- Standard Errors
- Ensembles
- Discussion

The systematic/random conjecture

- Multivariate regression has been taught as

$$Y = X.A + \varepsilon$$

(Note: this is the "fancy" version of Y = signal + noise)[1]

- *Signal + noise = systematic and random components*

1 Introduction to Ratemaking *Multivariate Methods*
http://www.casact.org/education/rpm/2009/handouts/cooksey.pdf

The systematic/random conjecture

If the MLR assumptions don't work well for insurance, then change them! With the same general approach, but the following assumptions, you've transitioned from MLRs to GLMs.

1. (*Random Component*) Observations are independent, but come from one of the family of exponential distributions.
2. (*Systematic Component*) X.A is called the linear predictor, or η.
3. (*Link function*) The expected value of Y, E(Y), is equal to g-1(η).

1 Introduction to Ratemaking *Multivariate Methods*
http://www.casact.org/education/rpm/2009/handouts/cooksey.pdf

The systematic/random conjecture

- Is this appropriate?
  - Are the components separable
    - Signal and noise
    - Systematic and random
  - Do modern regression techniques achieve this?
  - Is maximum likelihood regression stable?
  - How can we tell?
- We need a controlled experiment

Experimental Design

- Aim
  - Design of a controlled experiment to test how much data noise impacts regression
- Data
- Data strategy
- Regression approach
- Comparison metrics

Experimental Design

- Data
  - One experience data set
  - PPA Collision coverage of 3 years
  - 1.8M years exposure, 75k claims
  - Very well-behaved data

Experimental Design

- Data Strategy
  - Divide experience data into 2 samples
    - Exclusively at random
    - Each sample has the same joint distribution of variables
    - Each sample has
      - 900k years exposure
      - 37.5k claims

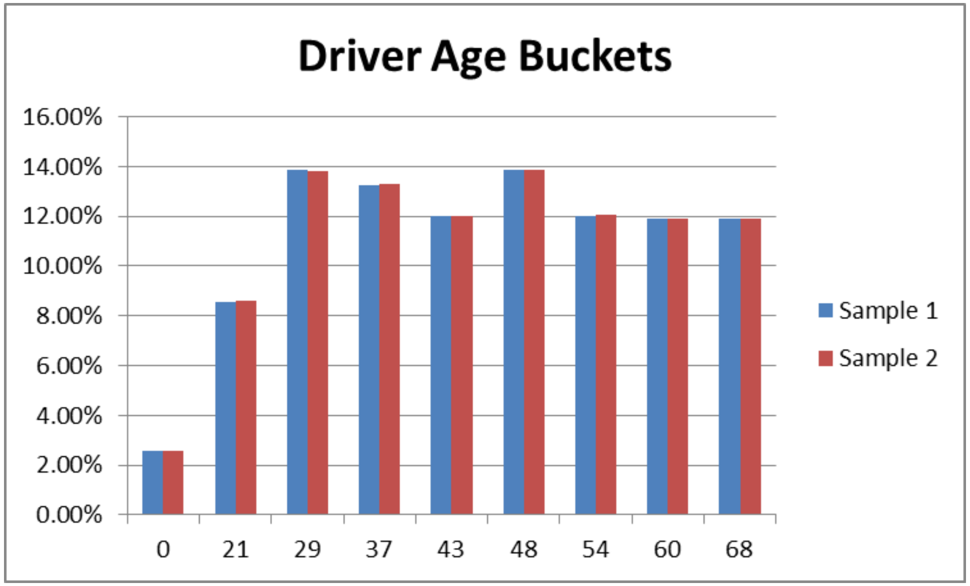
	Claim Frequency	Claim Severity	Loss Cost
Sample 1	4.19%	3680	155
Sample 2	4.16%	3718	154

Experimental Design

- Data Strategy
  - Variables selection
    - Top 50 variables by signal strength chosen by
      - Claim frequency
      - Claim severity
      - Loss cost
    - Based on all the data
  - Variables bucketing (binning) automatically
    - Categorical variables
      - Categories that have at least 10% exposure
    - Ordinal variables
      - Optimal buckets by signal type (up to 10 but usually much less - 2)
    - Based on all the data

Experimental Design

- Variable bucketing
  - Driver age buckets for claims frequency – relative exposure
    - Less buckets for severity and loss cost
  - Sample consistency(unsurprising)



Experimental Design

- Regression Approach
  - Divide each sample into training and validation data
    - 70%-30% at random
  - Compare traditional and modern methods
    - GLM
    - Matrix ensemble
    - Claim Frequency

Experimental Design

- GLM
  - Log link
  - Poisson, gamma, Tweedie distributions
  - Forward stepwise model selection
- Matrix ensemble
  - Members (base learners) placed in a matrix so that
    - Columns perform variance reduction
    - Rows perform bias reduction
    - Talon base learner
  - Un-optimized choices
    - 10% exposure for claim frequency
    - 4000 claims for severity and loss cost
    - Matrix size is 10 columns and 50 rows

Experimental Design

- Selecting models
  - GLM based on:
    - AIC as calculated on the validation data
    - BIC as calculated on the validation data
    - The minimum of the appropriate deviance as calculated on the validation data.
    - AIC as calculated on the training data
    - BIC as calculated on the training data
  - Matrix ensemble
    - No choice possible – model is self defining

Experimental Design

- Comparison metrics
  - For each model chosen apply the model to both data samples and then compare observed estimates
    - $Dispersion = 2(e_1 - e_2) / (e_1 + e_2)$
    - $Difference = (e_1 - e_2)$

Where e_1, e_2 represent estimates from sample 1 and 2 based models

Experimental Design

- Accumulate

Dispersion	Difference		
-0.5	< -0.5	-200	< -200
-0.2	[-0.5,-0.2)	-150	[-200,-150)
-0.1	[-0.2,-0.1)	-100	[-150,-100)
-0.05	[-0.1,-0.2)	-75	[-100,-75)
-0.02	[-0.05,-0.02)	-50	(-75,-50)
-0.01	[-0.02,-0.02)	-20	[-50,-20)
0	(-0.01,0.01)	0	(-20,20)
0.01	[0.01.0.02)	20	[20,50)
0.02	[0.02,0.05)	50	[50,75)
0.05	[0.05,0.1)	75	[75,100)
0.1	[0.1,0.2)	100	[100,150]
0.2	[0.2.0.5]	150	[150,200)
0.5	> 0.5	200	>200

Results – Claim Frequency

- Select optimum GLM models
  - How many variables to include
  - Note training and validation based AIC and BIC

	GLM Training AIC	GLM Training BIC	GLM Validation AIC	GLM Validation BIC	GLM Validation Deviance
Sample 1 Iteration	35	15	18	15	42
Sample 2 Iteration	32	14	18	13	36

Results – Claim Frequency

- Dispersion of estimates
  - Based on applying pairs of models to all the data

Dispersion	GLM Training AIC	GLM Training BIC	GLM Validation AIC	GLM Validation BIC	GLM Validation Deviance	Ensemble
-0.5	0%	0%	0%	0%	0%	0%
-0.2	3%	3%	4%	6%	3%	0%
-0.1	11%	16%	11%	13%	11%	5%
-0.05	12%	11%	12%	13%	12%	14%
-0.02	10%	7%	9%	8%	10%	15%
-0.01	4%	3%	3%	3%	4%	6%
0	7%	6%	7%	5%	7%	13%
0.01	4%	3%	4%	3%	4%	6%
0.02	11%	10%	11%	7%	11%	16%
0.05	16%	17%	16%	12%	16%	18%
0.1	17%	19%	17%	22%	17%	7%
0.2	5%	5%	5%	8%	5%	0%
0.5	0%	0%	0%	0%	0%	0%
Mean	8.94%	9.85%	9.31%	11.19%	8.86%	5.13%

Results – Claim Frequency

- Maximum, minimum observed values

	GLM Training AIC	GLM Training BIC	GLM Validation AIC	GLM Validation BIC	GLM Validation Deviance	Ensemble
Model 1						
Min	0.003922	0.009872	0.007223	0.010492	0.003896	0.01135
Max	1.811758	1.362593	1.880287	1.298176	1.8865	0.10384
Model 2						
Min	0.000016	0.011671	0.010372	0.011671	0.000016	0.01338
Max	1.548286	1.303763	1.535855	1.303763	1.510621	0.10639
Ratios						
Model 1	462	138	260	124	484	9
Model 2	96768	112	148	112	94414	8

Discussion

- Forward stepwise procedure
  - "I can fit better models than that!"
  - Better?
    - Data is given – samples 1 and 2
    - Used conservative variable bucketing approach
      - Relatively few beta values
    - Forward stepwise very close to forward stepwise by 3
  - What would better mean?

Discussion

- Lower AIC
  - Proposed by many practitioners
  - Fits a lot of variables
    - Approximately 30 for frequency
  - Different when calculated on validation data
    - Far fewer variables indicated
    - Is not a reliable complexity measure
  - Models are not more consistent
- Lower BIC
  - Heavier penalty for complexity
  - Same problem as AIC

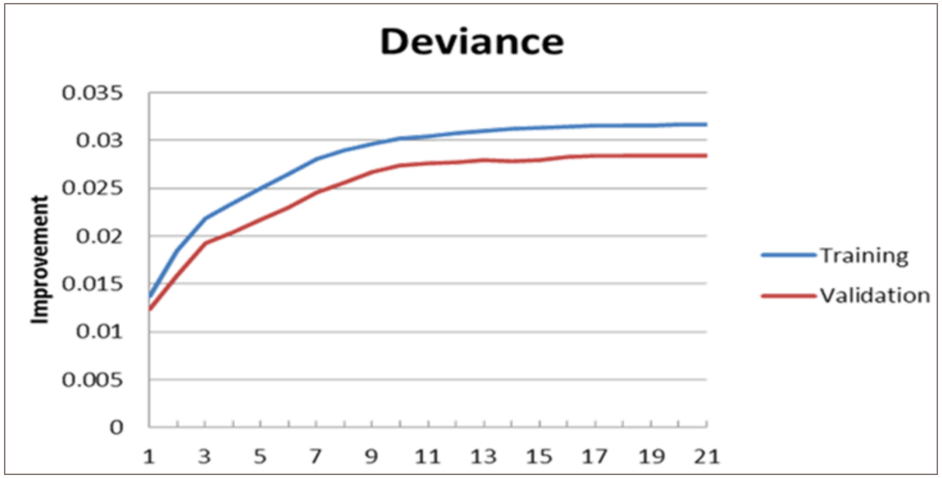
Discussion

- Lower deviance
  - Can’t use on training data
    - Deviance decreases as more variables as included
  - Validation based calculation
    - Also indicated lots of variables
  - Models are not more consistent

Discussion

- Better P values (What are P values?)
  - Use splines
    - Increases beta values
    - Allows better fit to training data
    - Models are not more consistent
- Smaller residuals
  - Proxy for deviance
- Better likelihood
  - Proxy for deviance

Could We See This Coming?



Discussion

- What happens if we average the models
  - Average of 5 sample 1 models v sample 2 models
    - A naïve ensemble

Dispersion			
	Claim Frequency	Severity	Loss Cost
-0.5	0%	0%	0%
-0.2	1%	0%	3%
-0.1	9%	1%	8%
-0.05	13%	12%	11%
-0.02	11%	22%	9%
-0.01	4%	9%	3%
0	8%	19%	8%
0.01	4%	9%	4%
0.02	12%	18%	12%
0.05	18%	10%	18%
0.1	16%	1%	20%
0.2	3%	0%	4%
0.5	0%	0%	0%
Mean	7.71%	3.40%	8.82%
Original Mean	8.86%-11.19%	4.03%-5.52%	11.17%-17.63%

Linear – as in GLM

- Generalised LINEAR Model
  - What does that mean?
- Regression on Claim Frequency using age and car age
  - Observed data

Table 7. Initial data

age	carage	claim
Old	Old	0.2
Old	New	0.3
Young	Old	0.4
Young	New	0.7

Generalised Linear models for Non-life Pricing – Overlooked Facts and Implications
A Report from GIRO Advanced Pricing Techniques Working Party

Linear

- Using Poisson and log link

Table 8. GLM prediction of initial data

age	carage	claim	pred
Old	Old	0.2	0.1875
Old	New	0.3	0.3125
Young	Old	0.4	0.4125
Young	New	0.7	0.6875

- Old-Old now Increased to 0.25

Table 9. GLM prediction of updated data

Age	carage	claim	pred	Change
Old	Old	0.25	0.21667	0.02917
Old	New	0.3	0.33333	0.02083
Young	Old	0.4	0.43333	0.02083
Young	New	0.7	0.66667	-0.02083

- Why does Young New change?

When It's Not Poisson
Create a Poisson Data Generating Function

Stata code to create the simulated data consists of the following:

```
STATA CODE
. clear
. set obs 50000
. set seed 4590
. gen x1 = runiform()
. gen x2 = runiform()
. gen x3 = runiform()
. gen xb = 1 + 0.75*x1 - 1.25*x2 + .5*x3
. gen exb =  exp(xb)
. gen py = rpoisson(exb)
. tab py

         py |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      4,579        9.16        9.16
          1 |      9,081       18.16       27.32
          2 |     10,231       20.46       47.78
          3 |      8,687       17.37       65.16
          4 |      6,481       12.96       78.12
          .            .           .           .
         14 |         16        0.03       99.98
         15 |          7        0.01       99.99
         16 |          3        0.01      100.00
------------+-----------------------------------
      Total |     50,000      100.00
```

Can We Find the Model - Yes

The mean and median of the Poisson response are 3.0. The displayed output has been amended:

```
. sum py, detail
50%      3                  Mean     3.00764
. glm py x1 x2 x3, fam(poi) nolog       // non-numeric mid-header
                                           output deleted
Generalized linear models                  No. of obs        =      50000
Optimization      : ML                     Residual df       =      49996
                                           Scale parameter   =          1
Deviance          =  54917.73016           (1/df) Deviance   =   1.098442
Pearson           =  49942.18916           (1/df) Pearson    =   .9989237
                                           AIC               =   3.744693
Log likelihood    = -93613.32814           BIC               =  -486027.9
-------------------------------------------------------------------------
             |                 OIM
          py |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+-----------------------------------------------------------
          x1 |   .7502913   .0090475    82.93   0.000     .7325586     .768024
          x2 |  -1.240165   .0092747  -133.71   0.000    -1.258343   -1.221987
          x3 |    .504346   .0089983    56.05   0.000     .4867096    .5219825
       _cons |   .9957061     .00835   119.25   0.000     .9793404    1.012072
-------------------------------------------------------------------------
. abic
AIC Statistic    =    3.744693          AIC*n       = 187234.66
BIC Statistic    =    3.744755          BIC(Stata)  = 187269.94
```

Remove a Variable – x2

```
. glm py x1 x3, nolog fam(poi)

Generalized linear models                          No. of obs      =      50000
Optimization     : ML                              Residual df     =      49997
                                                   Scale parameter =          1
Deviance         =  73496.70743                    (1/df) Deviance =   1.470022
Pearson          =  68655.76474                    (1/df) Pearson  =   1.373198
                                                   AIC             =   4.116233
Log likelihood   = -102902.8168                    BIC             =  -467459.7
------------------------------------------------------------------------------
             |                 OIM
          py |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   .7513276   .0090462    83.05   0.000     .7335973    .7690579
          x3 |   .4963757   .0090022    55.14   0.000     .4787317    .5140196
       _cons |    .444481   .0075083    59.20   0.000      .429765    .4591969
------------------------------------------------------------------------------
```

- It’s Over-dispersed – No Longer Poisson
- Can be Modelled Correctly using the Negative Binomial

Implications?

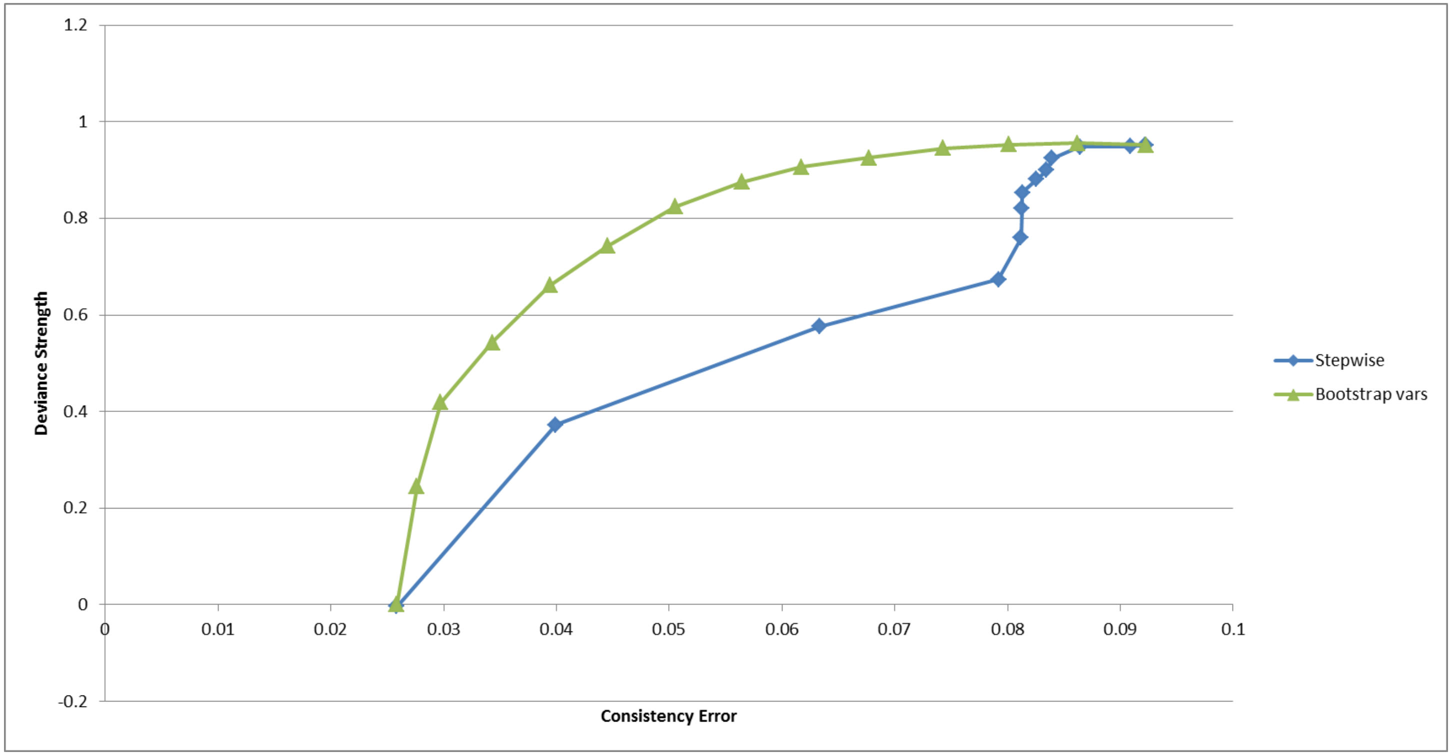
- Imagine Our Data Generating Function is Poisson
- We Only Have Imperfect Variables to Use
  - We Have This Situation
- Depending Which Variables We use
  - Over/Under-Dispersal Changes
  - Different Shape Parameter values for Negative Binomial
  - Shape Affects Deviance and Likelihood
  - Adding Variables Causes Deviance to Increase
    - Nested Models are Not Nested Any More!
    - AIC, BIC - all increase as well

Standard Errors

- Standard Errors Reflect
  - Distribution of Predictor Variables - YES
  - Variance of the Dependent Variable – NO
  - Only Depend on X and W
- Assume that
  - Model Structure is Ideal
  - Transformed Errors are Normally Distributed
  - And So Do P Values

Ensembles of GLMs

- 100 Iteration GLMs on Bootstrap Samples
  - Tweedie Example



Conclusion

- Systematic/random conjecture
  - Doesn't look very real
    - Systematic polluted by randomness
  - Model statistics are of very limited help
  - Statistical inference - uncertain!
- Linearity
  - The irrelevant influences all the model
- Negative Binomial Problem
  - Nested models exhibit increasing deviance
- Standard Errors
  - Not so good news