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.

GLM II: Basic Modeling Strategy

Ernesto Schirmacher

Liberty Mutual Insurance

Casualty Actuarial Society
Ratemaking and Product Management Seminar
March 30–April 1, 2014
Washington, DC

Overview

Basic GLM Specification

$$g(\mathbb{E}[y]) = \beta_0 + x_1\beta_1 + \cdots + x_k\beta_k + \text{offset}$$

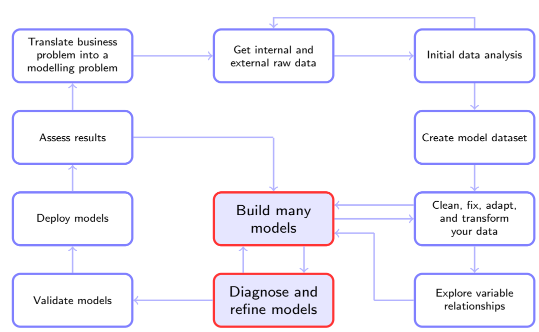
1. The link function is g
2. The distribution of y is a member of the exponential family
3. The explanatory variables x_i may be continuous or discrete
4. The offset term can be used to adjust for exposure or to introduce known restrictions

$$\mathbb{E}[y] = g^{-1}\left(\beta_0 + x_1\beta_1 + \cdots + x_k\beta_k + \text{offset}\right)$$

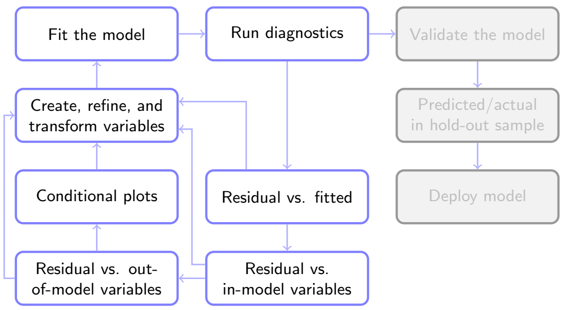
Common Model Forms

	Freq	Counts	Severity	Prob
Link	$\log(\mu)$	$\log(\mu)$	$\log(\mu)$	$\text{logit}(\mu)$
Error	Poisson	Poisson	Gamma	Binomial
Variance	μ	μ	μ^2	$\mu(1-\mu)$
Weights	Exposure	1	# claims	1
Offset	0	log(Exposure)	0	0

Overall Project Cycle



Model Building Cycle



Personal Auto Claims

The dataset contains 67, 856 policies taken out in 2004 or 2005. This is the `car.csv` dataset featured in the book by de Jong & Heller [3].

The available variables are:

1. Driver age
2. Gender
3. Garage location
4. Vehicle body
5. Vehicle age
6. Vehicle value (∞)
7. Exposure (∞)
8. Claim?
9. Number of claims
10. Total claim cost (∞)

(∞) denotes a continuous variable. All other variables are categorical or counts.

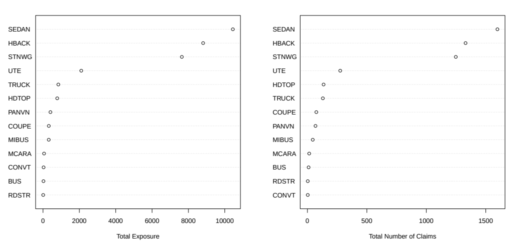
Variable Descriptions

Variable	Type	Comments
Driver Age	Cat	$1 =$ youngest, $2, \ldots, 6 =$ oldest
Gender	Cat	F = Female, M = Male
Garage Location	Cat	A, B, C, D, E, F
Vehicle Body	Cat	13 classes
Vehicle Age	Cat	1 to 4 = oldest
Vehicle Value	Cont	range: 0 to 34.56, in units of $10K
Exposure	Cont	range: 0.003 to 0.999
Claim?	Cat	0 = no claim, 1 = claim
Number of Claims	Count	$0, 1, 2, 3, 4$
Total Claim Cost	Cont	range: $0 to $55, 922

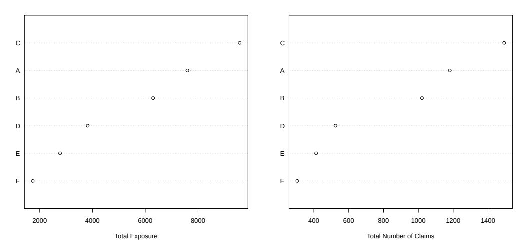
Exploratory Analysis

- Tabular summaries
- Univariate exploration (along with exposure)
- Bivariate relationships
- Correlations
- Missing Value Check Model

Exploratory Analysis: by Vehicle Body



Exploratory Analysis: by Geographic Area



Exploratory Analysis: Linear Correlations

	VV	VB	VA	A	G
Vehicle Value					
Vehicle Body	0.29				
Vehicle Age	−0.54	0.07			
Area	0.10	0.16	0.02		
Gender	0.10	0.19	0.05	0.01	
Age	−0.06	0.00	0.02	−0.05	0.05

Missing Value Check Model

Should be the very first model you build!

1. Make a copy of you dataset
2. Place a 1 if a predictor variable's value is *not missing*
3. Place a 0 if a predictor variable's value is missing
4. Leave all the response variables untouched!

The only information that remains in the input dataset is whether or not there is something entered for a predictor variable's value.

Create a predictive model that attempts to predict the value of the response variables.

Preparing to Stay Honest

Take precautions to make sure that the results achieved are actually worth having. To this end split your data into three sets:

1. *Build*: used to create many models
2. *Test*: used to check intermediate models
3. *Validate*: used only once to check your final model

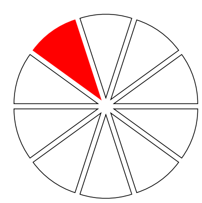
One rule of thumb: (50%, 25%, 25%).

Set	Records
Build	33,928
Test	16,964
Validate	16,964
Total	67,856

Preparing to Stay Honest

What if you don't have a large dataset that would allow you to split it in three segments (Build, Test, Validate)?

Use Cross-Validation!



Summary Statistics for Build Dataset

Continuous Variables

```
          total
          claim
           cost  exposure  veh.value
 Min.   :    0.0    0.003     0.000
 1st Qu.:    0.0    0.219     1.010
 Median :    0.0    0.446     1.500
 Mean   :  143.4    0.469     1.777
 3rd Qu.:    0.0    0.709     2.150
 Max.   :55920.0    0.999    34.560
```

Vehicle value is in units of $10,000.

Summary Statistics for Build Dataset

Categorical Variables (record counts)

```
    veh.body   veh.age   area
 SEDAN:11149   1: 6017   A: 8216
 HBACK: 9372   2: 8332   B: 6603
 STNWG: 8114   3:10126   C:10344
 UTE  : 2351   4: 9453   D: 4035
 TRUCK:  886             E: 2971
 HDTOP:  770             F: 1759
 COUPE:  396
 PANVN:  378
 MIBUS:  373
 MCARA:   60
 CONVT:   37
 BUS  :   27
 RDSTR:   15
```

Summary Statistics for Build Dataset

Categorical Variables (record counts)

```
                                  claim
 age.cat   gender     claim?      count
  1:2852  F:19264  No :31599    0:31599
  2:6501  M:14664  Yes: 2329    1: 2185
  3:7971                        2:   133
  4:8086                        3:    10
  5:5290                        4:     1
  6:3228
```

What is the claim frequency?

$$\text{frequency} \stackrel{?}{=} \frac{2329}{2329 + 31599} = 6.86\%$$

A naive GLM model for Claim Counts

```
Call: glm(formula = num.claims ~ 1,
          family = poisson(link = "log"),
            data = car[b.idx, ])

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.61397    0.02006  -130.3   <2e-16 ***

    Null deviance: 13437 on 33927 degrees of freedom
Residual deviance: 13437 on 33927 degrees of freedom
```

$$e^{-2.61397} = 0.0732 = \frac{2485}{33928}$$

How to adjust for Exposure?

For a frequency model with a log-link we have

$$\log\left(\frac{\mathbb{E}[\text{counts}]}{\text{exposure}}\right) = \text{linear predictor}$$

$$\log\left(\mathbb{E}[\text{counts}]\right) = \text{linear predictor} + \underbrace{\log\left(\text{exposure}\right)}_{\text{offset term}}$$

A simple GLM model for Claim Counts

```
Call: glm(formula = num.claims ~ 1,
          family = poisson(link = "log"),
            data = car[b.idx, ],
          offset = log(exposure))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.85591    0.02006  -92.52   <2e-16 ***

    Null deviance: 12864 on 33927 degrees of freedom
Residual deviance: 12864 on 33927 degrees of freedom
```

$$e^{-1.85591} = 0.1563 = \frac{2485}{15897.84}$$

Continue with Brent's presentation

References

- John M. Chambers, William S. Cleveland, Beat Kleiner, and Paul A. Tukey.
  Graphical Methods for Data Analysis.
  The Wadsworth Statistics/Probability Series. Wadsworth International Group, Belmont, California, 1983.
- W.S. Cleveland.
  Visualizing Data.
  Hobart Press, 1993.
- Piet De Jong and Gillian Z. Heller.
  Generalized Linear Models for Insurance Data.
  Cambridge University Press, 2008.

References

- Peter K. Dunn and Gordon K. Smyth.
  Randomized quantile residuals.
  Journal of Computational and Graphical Statistics, 5(3):236–244, 1996.
- L. Fahrmeir and G. Tutz.
  Multivariate Statistical Modelling Based on Generalized Linear Models.
  Springer, 2001.
- James Hardin and Joseph Hilbe.
  Generalized Linear Models and Extensions.
  Stata Press, College Station, Texas, 2001.
- W.N. Venables and B.D. Ripley.
  Modern Applied Statistics with S.
  Springer New York, 2002.