GLM II: Basic Modeling Strategy

Ernesto Schirmacher

Liberty Mutual Group

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.

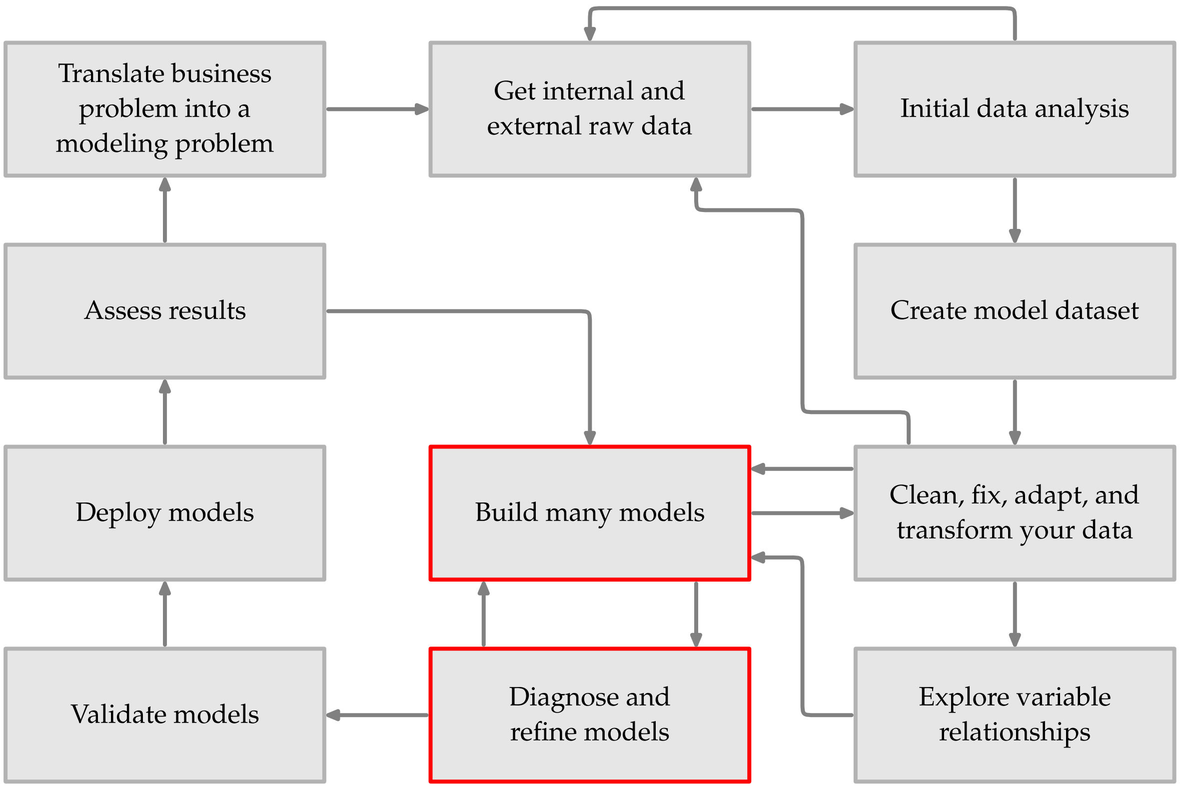
Under no circumstances shall CAS seminars be used as a means for competing companies or firms to reach any understanding – expressed or implied – that restricts competition or in any way impairs the ability of members to exercise independent business judgment regarding matters affecting competition.

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.

Overview

1. The modeling cycle
2. Quick review of GLMs
3. Concrete example
   - Summary statistics
   - Exploratory plots
   - Fitting models and parameter estimates
   - Diagnosing the fit and corrective measures
   - Interactions
4. Validation
5. Model building summary

Basic Modeling Cycle



Basic Model Form

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

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

Common Model Forms

Link functions: identity (additive effects), logarithm (multiplicative effects), reciprocal, log odds, probit, etc ...

Response distributions: normal, gamma, inverse gaussian, Tweedie, binomial, poisson, negative binomial

Offset: to adjust for exposure or to incorporate known effects

Personal Injury Claims

The dataset (see [4]) contains 22,036 claims arising from accidents between July 1989 and January 1999. Claims settled with zero payment are not included. The variables in the dataset are:

1. Settlement amount (range: $10 to $4.5M)
2. Injury type (codes: 1, 2, 3, 4, 5, 6, 9)
3. Legal representation (codes: 1–Yes, 0–No)
4. Accident, reporting, and settlement month
5. Operational time

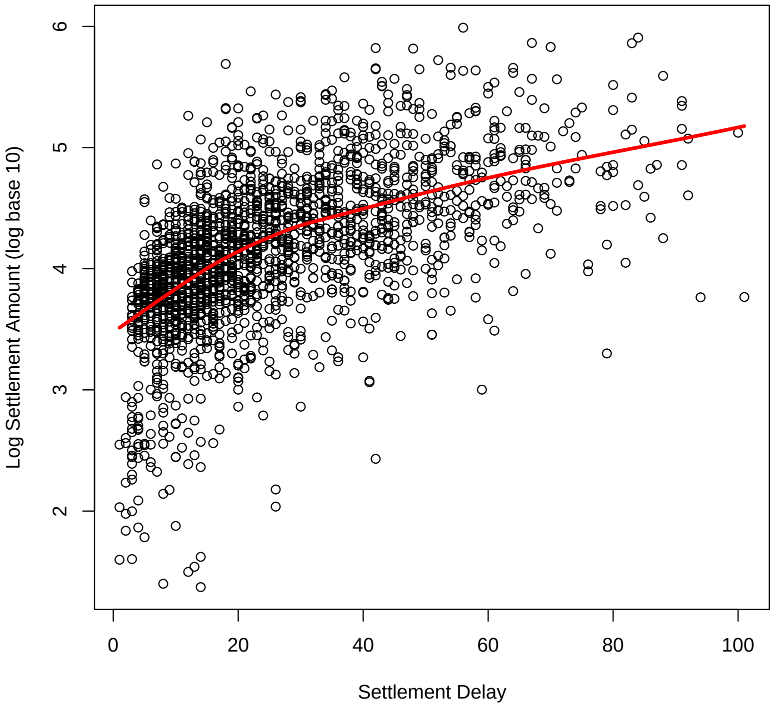
We will work with a random sample of 2,000 claims.

Summary Statistics (for random sample)

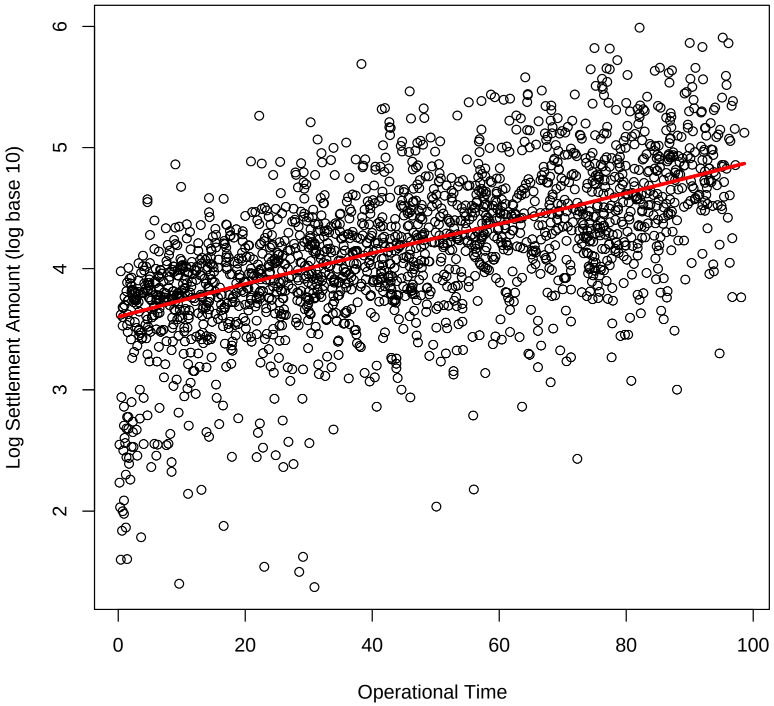
	Claim Amount
Minimum	24
1st Quartile	6,144
Median	14,222
Mean	37,525
3rd Quartile	35,435
Maximum	976,379

There are 172 records ($\approx$ 8.5%) with claim amounts greater than 100,000.

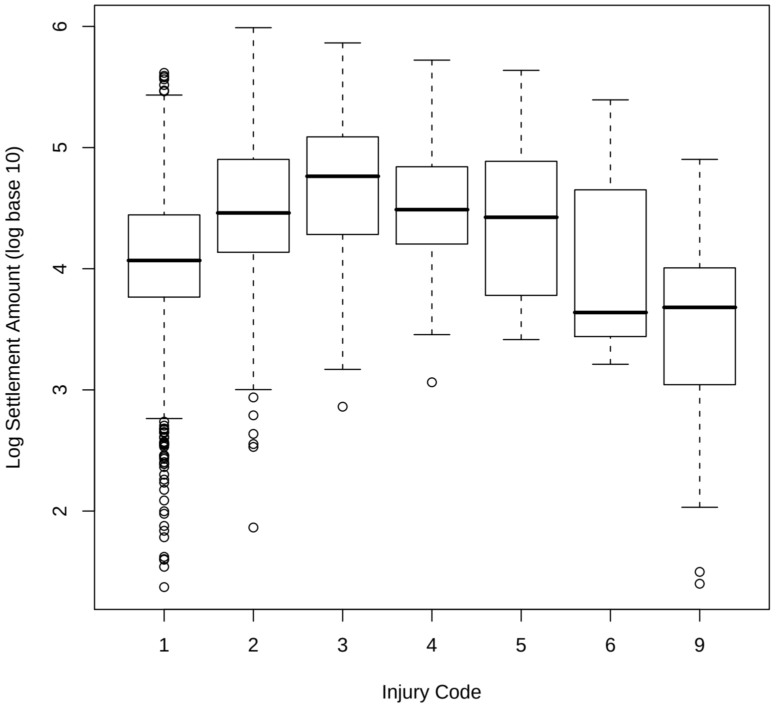
Exploratory Plots I



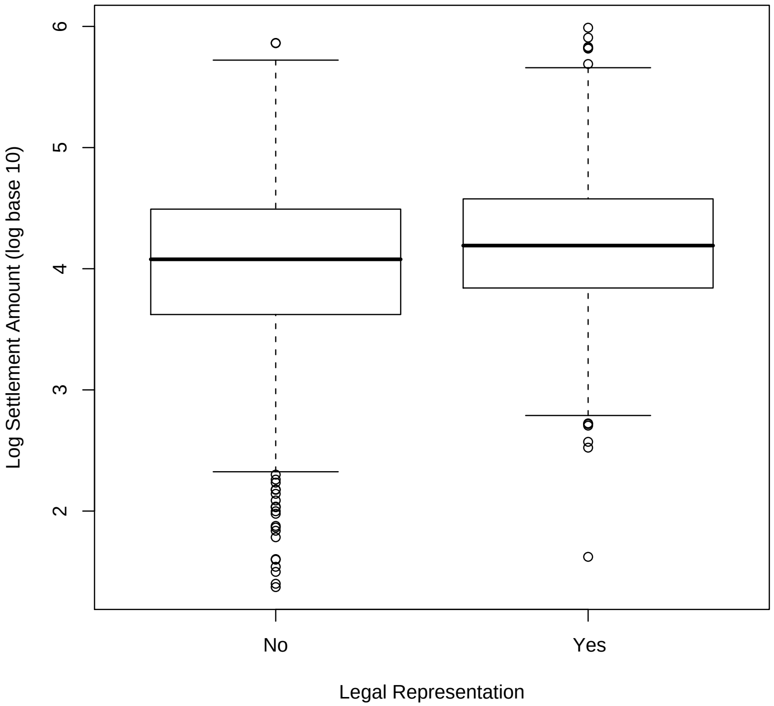
Exploratory Plots II



Exploratory Plots III



Exploratory Plots IV



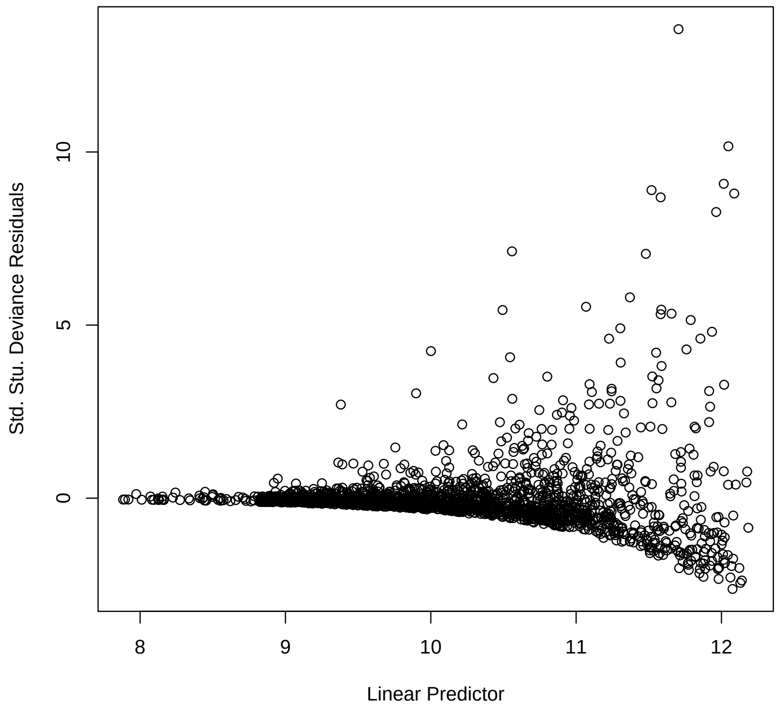
Normal log-link model

log(Settlement Amount) = Op.Time + Injury + Attorney

	Estimate	Std. Error	t value	Pr(>t)
(Intercept)	8.817	0.138	63.99	< 2e-16
Op.Time	0.026	0.002	15.82	< 2e-16
injury 2	0.757	0.067	11.31	< 2e-16
injury 3	0.844	0.079	10.75	< 2e-16
injury 4	0.607	0.182	3.33	0.0009
injury 5	0.505	0.199	2.54	0.0113
injury 6	0.645	0.245	2.63	0.0086
injury 9	-0.942	0.554	-1.70	0.0892
attorney Yes	-0.017	0.057	-0.29	0.7705

Residual deviance: 7.9e+12 on 1991 degrees of freedom

Residual Check: Normal error



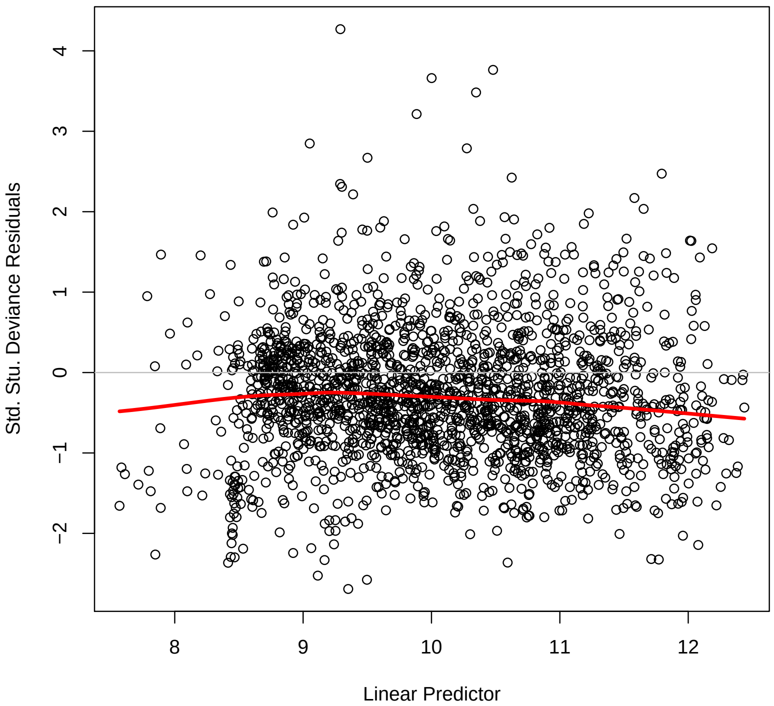
Gamma log-link model

log(Settlement Amount) = Op.Time + Injury + Attorney

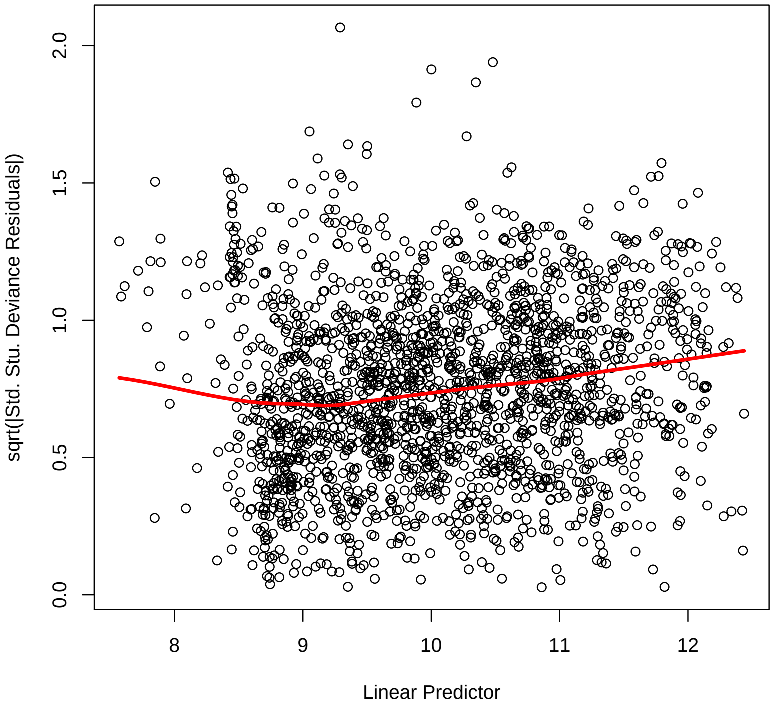
	Estimate	Std. Error	t value	Pr(>t)
(Intercept)	8.425	0.064	130.69	< 2e-16
Op.Time	0.030	0.001	29.67	< 2e-16
injury 2	0.707	0.074	9.49	< 2e-16
injury 3	0.900	0.116	7.75	1.46e-14
injury 4	1.045	0.271	3.85	0.0001
injury 5	0.279	0.323	0.86	0.39
injury 6	0.199	0.247	0.80	0.42
injury 9	-0.864	0.129	-6.68	3.00e-11
attorney Yes	0.200	0.057	3.52	0.0004

Residual deviance: 2072.0 on 1991 degrees of freedom

Residual Check: Gamma error



Location–Spread Plot for Gamma Model

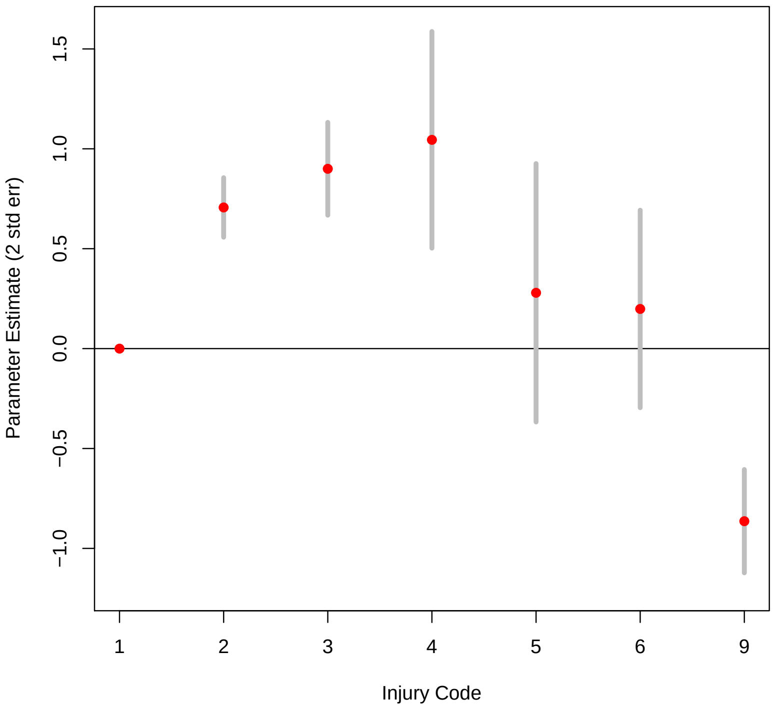


Analysis of Deviance Table

```
Model: Gamma, link: log
Response: settlement amount
Terms added sequentially (first to last)
```

	Df	Change in Deviance	Resid. Deviance	Resid. Df
(Intercept)			3894	1999
Op.Time	1	1502	2392	1998
injury	6	303	2089	1992
attorney	1	17	2072	1991

Injury Parameter Estimates



Grouping Injury Levels

Model	Injury levels	Deviance	Diff	q	Crit.Val.
1	1 2 3 4 5 6 9	2072			
2	1 (2 3 4) 5 6 9	2077	5	2	5.9
3	1 (2 3 4) (5 6) 9	2077	5	3	7.8
4	(1 5 6) (2 3 4) 9	2079	7	4	9.5
5	1 (2 3 4 5 6) 9	2086	14	4	9.5

Diff: is the difference between the current model and model 1.
q: is the number of restrictions in the current model compared to model 1.
Crit.Val.: is the 0.95 quantile of the chi-squared distribution with q degrees of freedom.

Checking the Link Function

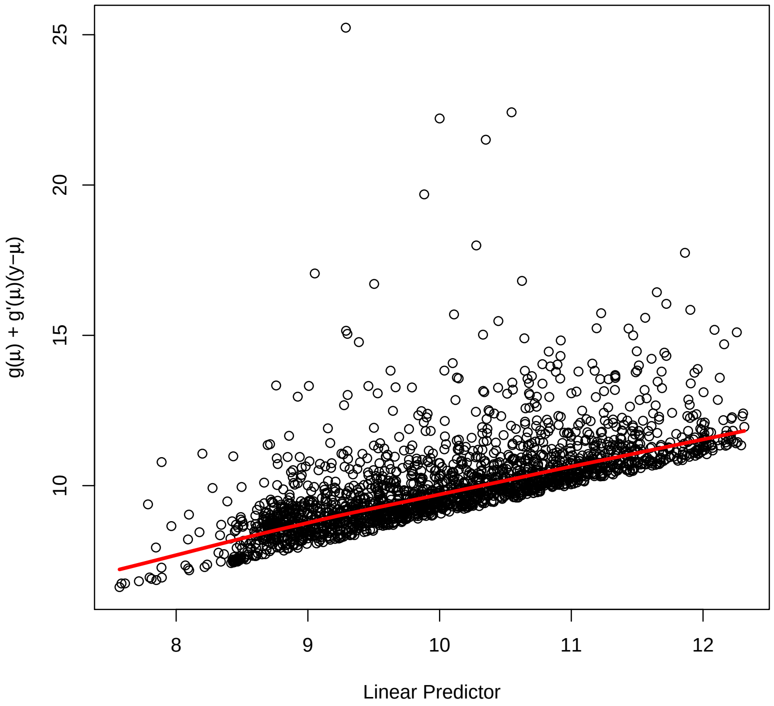
Two ways to assess the link function:

1. Embed the link function in a parametric family and compare model fit at various points.
2. We know that

$$x_i'\beta = g(y_i) \approx g(\mu_i) + g'(\mu_i)(y_i - \mu_i)$$

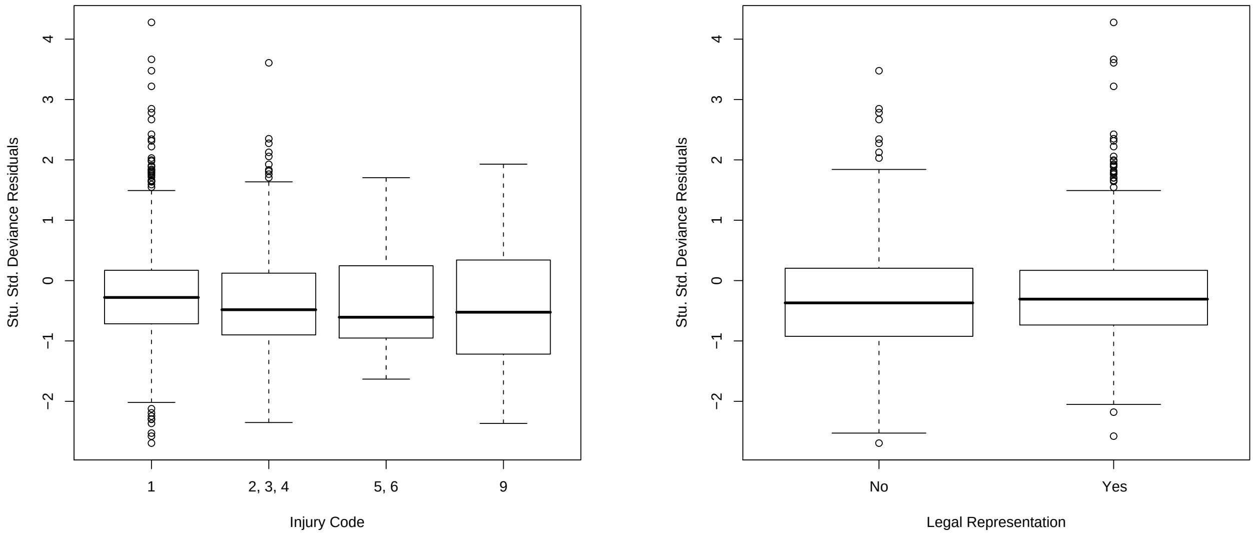
   So plotting the linear predictor against the right-hand side of the above equation should give us a straight line.

Checking the Link Function

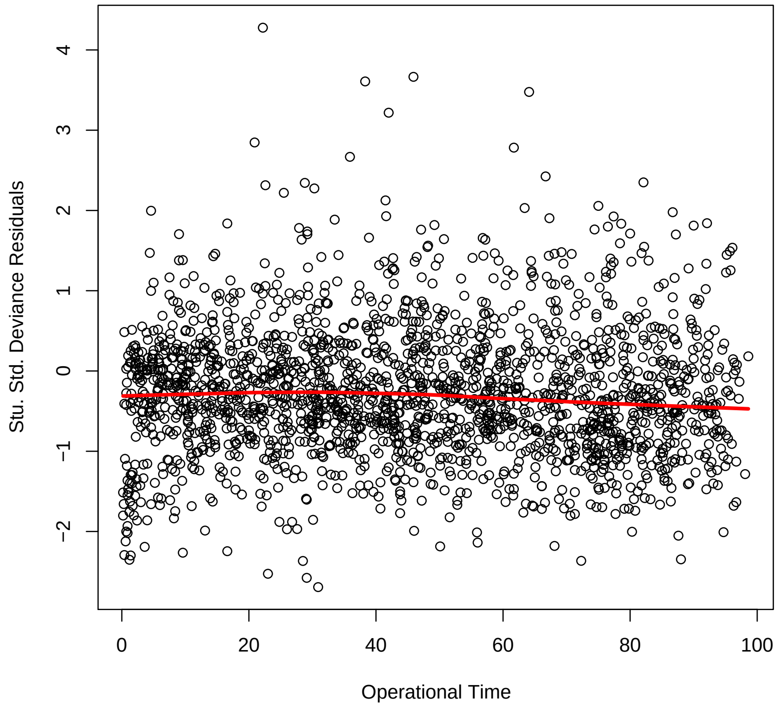


Checking Explanatory Variables

Plot residuals against explanatory variables.



Checking Explanatory Variables

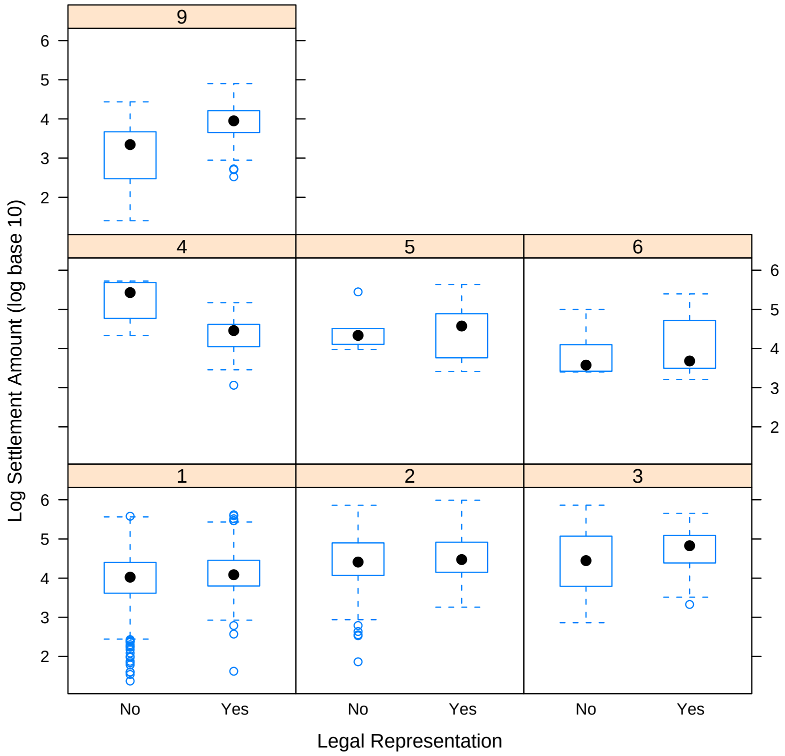


Interactions

We say that two explanatory variables x and z interact if the effect of x on the response variable depends on the values of z.

For our example, does the effect of attorney involvement depend on the type of injury?

Conditional Plot



Model Validation

Several model validation techniques:

1. Out-of-sample
2. Cross-validation
3. Bootstrap estimates of prediction errors

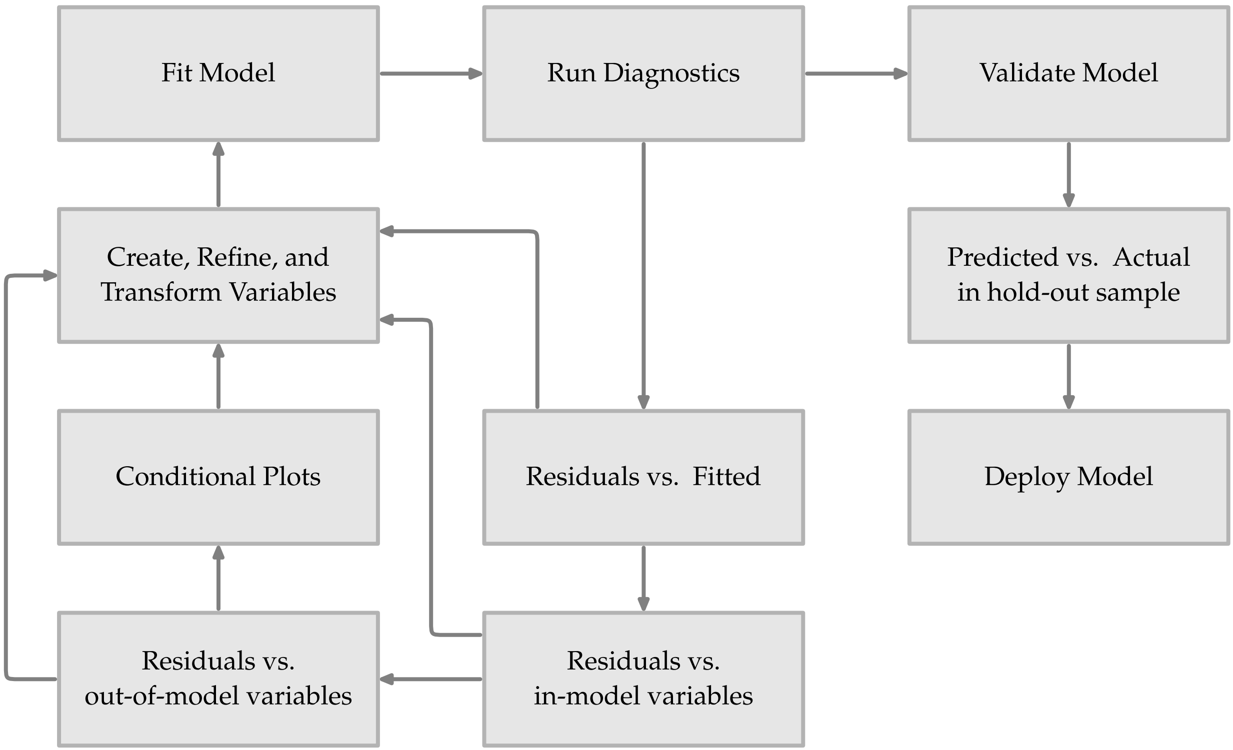
Out-of-Sample Validation

Predicted values compared against actual values for a new sample of 2,000 claims.

Predicted Range	Type	1st Qu.	Mean	Ratio A/P	3rd Qu.
(43800, 61500]	A	14770	45790		52880
	P	48150	52720	0.87	57460
(61500, 91600]	A	22800	77900		85350
	P	67180	74800	1.04	81860
(91600,232000]	A	42680	150700		171700
	P	106300	135000	1.12	156700

Only the last three groups of the table are shown.
The type column refers to actual (A) or predicted (P) values.
The column ratio A/P is the ratio of the actual mean divided by the predicted mean.

Model Building Summary



References

[1] Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. 1983. *Graphical Methods for Data Analysis*. Belmont, California: Wadsworth International Group.

[2] Fahrmeir, L., and Tutz, G. 2001. *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer.

[3] Hardin, J., and Hilbe, J. 2001. *Generalized Linear Models and Extensions*. College Station, Texas: Stata Press.

[4] De Jong, P., and Heller, G. Z. 2008. *Generalized Linear Models for Insurance Data*. Cambridge University Press.

[5] Cleveland, W. 1993. *Visualizing Data*. Hobart Press.

[6] Venables, W., and Ripley, B. 2002. *Modern Applied Statistics with S*. Springer New York.