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

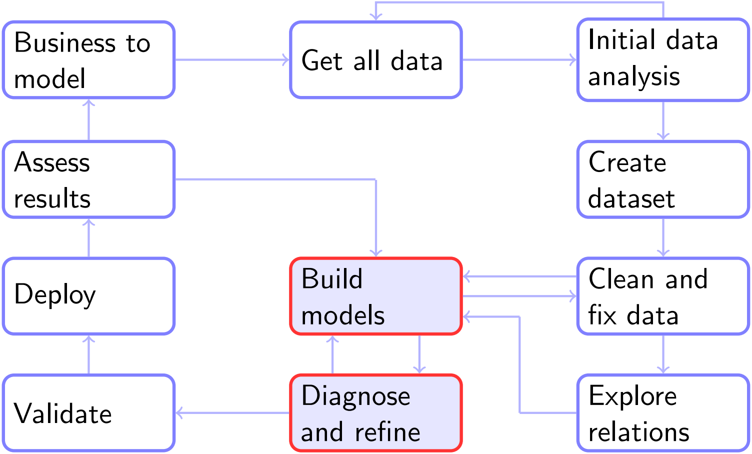
Ernesto Schirmacher

Liberty Mutual Group

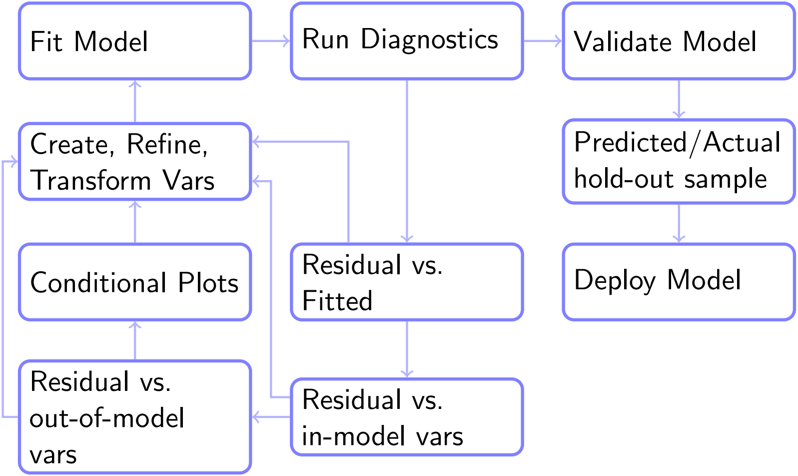
Casualty Actuarial Society
Ratemaking and Product Development Seminar
March 15–17, 2010
Chicago, IL

Overview

Overall Project Cycle



Model Building Cycle



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 indicator
4. The offset term can be used to adjust for exposure or to introduce known restrictions

Common Model Forms

	Target Variable			
	Claim Frequencies	Claim Counts	Average Claim Amount	Probability
Link $g(\mu)$	$\log(\mu)$	$\log(\mu)$	$\log(\mu)$	$\text{logit}(\mu)$
Error	Poisson	Poisson	Gamma	Binomial
Variance $V(\mu)$	μ	μ	μ^2	$\mu(1-\mu)$
Weights	Exposure	1	# claims	1
Offset	0	log(Exposure)	0	0

Personal Auto Claims

The dataset contains $59,876$ policies taken out in 2004 or 2005. This dataset is a subset of the `car.csv` dataset featured in the book by de Jong & Heller [3]. I removed all records were vehicle body was not for personal use.

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

Possible Models?

1. Binary model: will a claim be made?
2. Count model: how many claims?
3. Severity model: how costly will a claim be?
4. Conditional severity model: how costly will a claim be given that at least one claim is filed?

Build, Test, Validate

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: $(50\%, 25\%, 25\%)$.

Let us take three independent samples of 3 000 records each.

Logistic Model

Target variable: Did a policy file a claim?
$\texttt{claim} = 0$; no claim reported
$\texttt{claim} = 1$; one or more claims reported

We want to model the expected value of filing a claim as a linear combination of predictors.

$$\mathbb{E}[\texttt{claim}] = \text{linear combination of predictors}$$

Logistic Model

Target variable: Did a policy file a claim?
$\texttt{claim} = 0$; no claim reported
$\texttt{claim} = 1$; one or more claims reported

We want to model the expected value of filing a claim as a linear combination of predictors.

$$\mathbb{E}[\texttt{claim}] = \text{linear combination of predictors}$$

$$\log\left(\frac{\mathbb{E}[\texttt{claim}]}{1 - \mathbb{E}[\texttt{claim}]}\right) = \text{linear combination of predictors}$$

$$\text{logit}(\mu) = \log\left(\frac{\mu}{1-\mu}\right) \qquad \text{logit}^{-1}(\eta) = \frac{e^{\eta}}{1+e^{\eta}} = \frac{1}{1+e^{-\eta}}$$

Summary Statistics for Build dataset

Continuous Variables

	claim	exposure	veh.value
Min.	:0.000	0.003	0.220
1st Qu.	:0.000	0.224	0.980
Median	:0.000	0.441	1.470
Mean	:0.075	0.467	1.753
3rd Qu.	:0.000	0.706	2.070
Max.	:1.000	0.999	12.090

Vehicle value is in units of $10,000.

Summary Statistics for Build dataset

Categorical Variables (record counts)

```
  veh.body   veh.age   area  age.cat  gender
SEDAN:1137     3:892  C:902    4:727  F:1808
HBACK: 940     4:806  A:757    3:679  M:1192
STNWG: 805     2:744  B:621    2:566
HDTOP:   87    1:558  D:350    5:509   claim
COUPE:   28           E:217    6:273  0:2775
CONVT:    2           F:153    1:246  1: 225
RDSTR:    1
```

We know that

$$\mathbb{E}[\text{claim}] = \frac{225}{3000} = 7.5\%$$

Summary Statistics for Build dataset

Categorical Variables (record counts)

```
  veh.body   veh.age   area  age.cat  gender
SEDAN:1137    3:892   C:902    4:727  F:1808
HBACK: 940    4:806   A:757    3:679  M:1192
STNWG: 805    2:744   B:621    2:566
HDTOP:   87   1:558   D:350    5:509   claim
COUPE:   28           E:217    6:273  0:2775
CONVT:    2           F:153    1:246  1: 225
RDSTR:    1
```

We know that

$$\mathbb{E}[\text{claim}] = \frac{225}{3000} = 7.5\%$$

Is this a good estimate?

What is the correct value for $\mathbb{E}[\text{claim}]$?

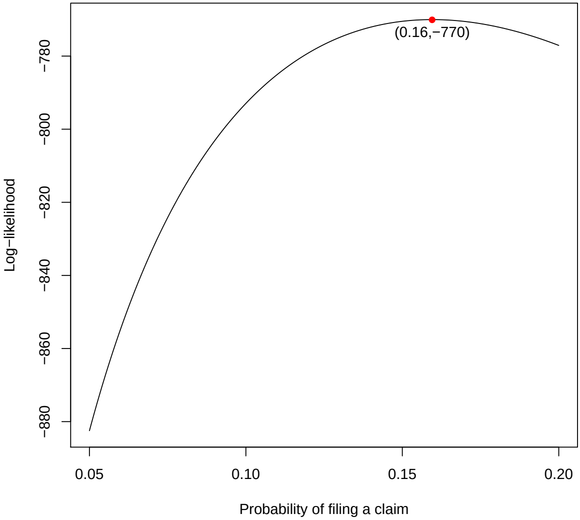
Let π be the probability of filing a claim over a one year period and let t_i be the exposure amount over the year.
Let's estimate π via maximum likelihood.

$$\max_{\pi \in (0,1)} \left\{ \prod_{i=1}^{3000} (t_i \pi)^{\text{claim}_i} (1 - t_i \pi)^{1-\text{claim}_i} \right\}$$

The log-likelihood is

$$\max_{\pi \in (0,1)} \left\{ \sum_{i=1}^{3000} \text{claim}_i \log(t_i \pi) + (1 - \text{claim}_i) \log(1 - t_i \pi) \right\}$$

Log-likelihood as a function of π



Logistic Regression—Null Model

```
Call:
glm(formula = claim ~ 1,
    family  = binomial(link = "logit"), ...)

Deviance Residuals: [...omited...]

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.51231    0.06932  -36.24   <2e-16 ***

    Null deviance: 1598.3  on 2999  degrees of freedom
Residual deviance: 1598.3  on 2999  degrees of freedom
```

How to adjust for Exposure?

How do we do it when modeling claim counts (Poisson, log-link)?

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

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

How should we do it in logistic regression?

How to adjust for Exposure?

How do we do it when modeling claim counts (Poisson, log-link)?

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

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

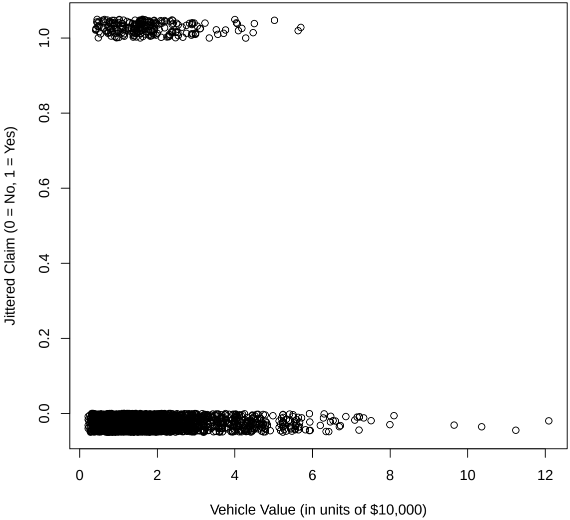
How should we do it in logistic regression?

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

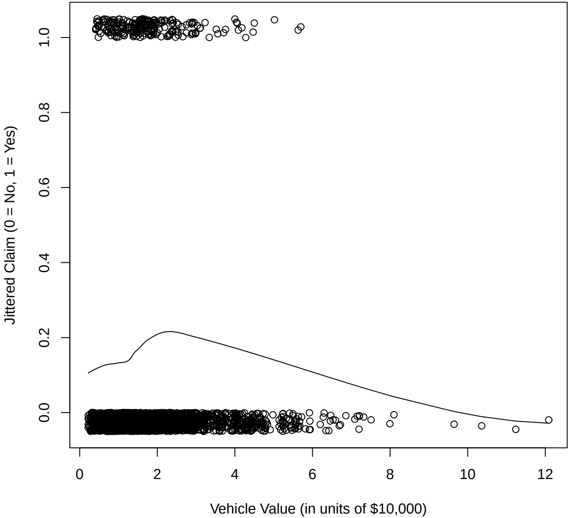
Logistic Regression—Null Model *Exposure Adjusted*

```
Call:
glm(formula = claim ~ 1,
    family  = binomial(link = logitexp(expo)), ...)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.66145    0.07532  -22.06   <2e-16 ***

    Null deviance: 1598.3  on 2999  degrees of freedom
Residual deviance: 1540.0  on 2999  degrees of freedom
```

Occurrence of a claim vs. vehicle value



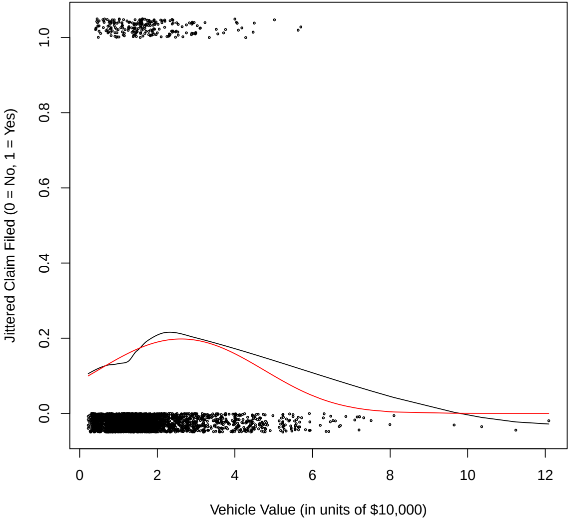
Occurrence of a claim vs. vehicle value with smoother



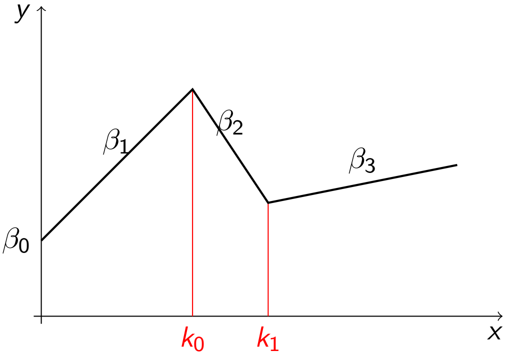
Logistic model with quadratic vehicle value

```
Call:
glm(formula = claim ~ veh.value + I(veh.value^2),
    family  = binomial(link = logitexp(expo)), ...)

Deviance Residuals: [...omited...]
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -2.35514    0.26906  -8.753  < 2e-16 ***
veh.value       0.73131    0.25412   2.878  0.00400 **
I(veh.value^2) -0.13991    0.05037  -2.778  0.00547 **

    Null deviance: 1598.3  on 2999  degrees of freedom
Residual deviance: 1529.0  on 2997  degrees of freedom
```

Claim vs vehicle value with quadratic fit



How to create piecewise linear fits?



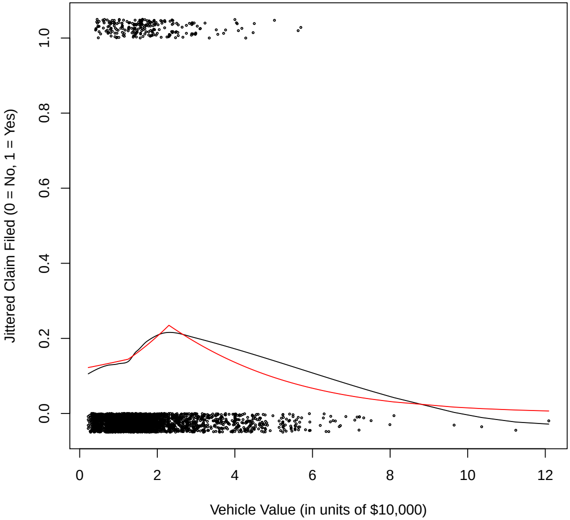
x	x^*	x^{**}
1.2	0	0
1.8	0	0
2.3	0.3	0
2.7	0.7	0
2.9	0.9	0
3.5	1.5	0.5
3.7	1.7	0.7
4.6	2.6	1.6

$$\mathbb{E}[y] = \beta_0 + \beta_1 x + \beta_2 \underbrace{\max(0, x - k_0)}_{\text{new variable } x^*} + \beta_3 \underbrace{\max(0, x - k_1)}_{\text{another variable } x^{**}}$$

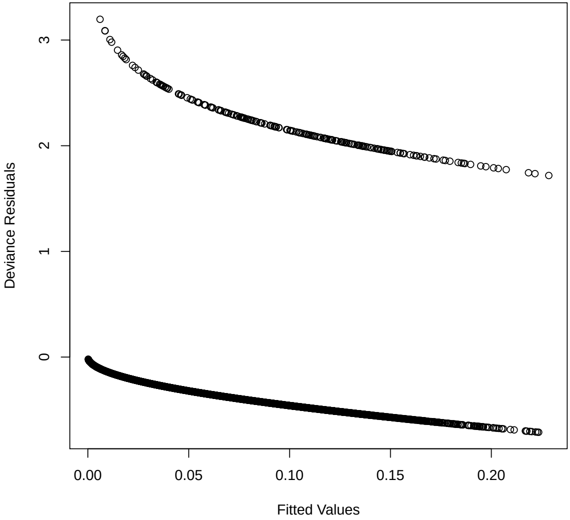
Logistic model with piecewise linear fit

```
Call:
glm(formula = claim ~ vv + vv2 + vv3,
    family  = binomial(link = logitexp(expo)), ...)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.0121     0.3762  -5.348  8.9e-08 ***
veh.value     0.1874     0.3730   0.503  0.61530
veh.value*    0.3812     0.5528   0.690  0.49045
veh.value**  -0.9606     0.3556  -2.701  0.00691 **

    Null deviance: 1598.3  on 2999  degrees of freedom
Residual deviance: 1528.1  on 2996  degrees of freedom
```

Claim vs Vehicle Value with Piecewise Linear Fits



Residual Plot: deviance residuals vs. fitted values



Many flavors of residuals

Response	$y - \hat{\mu}$
Working	$(y - \hat{\mu})(\partial\eta/\partial\hat{\mu})$
Partial	$(y - \hat{\mu})(\partial\eta/\partial\hat{\mu}) + x_k\hat{\beta}_k$
Pearson	$(y - \hat{\mu})/\sqrt{V(\hat{\mu})}$ the variance.
Anscombe	Transformed response residuals towards normality.
Deviance	Signed square root contribution to the Deviance from each observation.
Quantile	For each response variable find the equivalent standard normal deviate. Use randomization for discrete distributions.

Toppings can vary too!

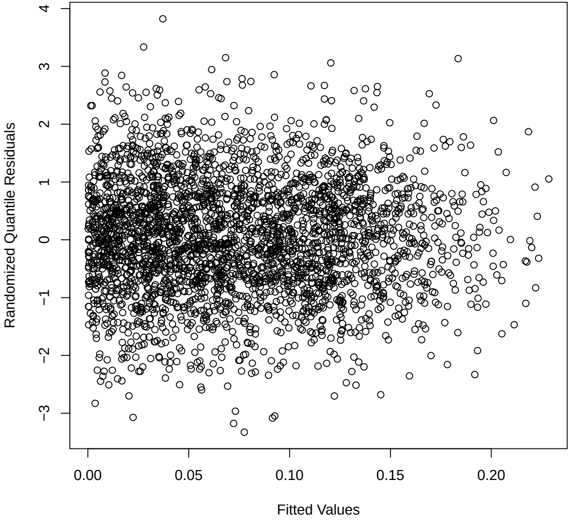
Modified The denominator of the residual has been modified to be a reasonable estimate of the response variance.

Standarized The variance of the residual has been standarized to take into account the correlation between the response and the fitted value.

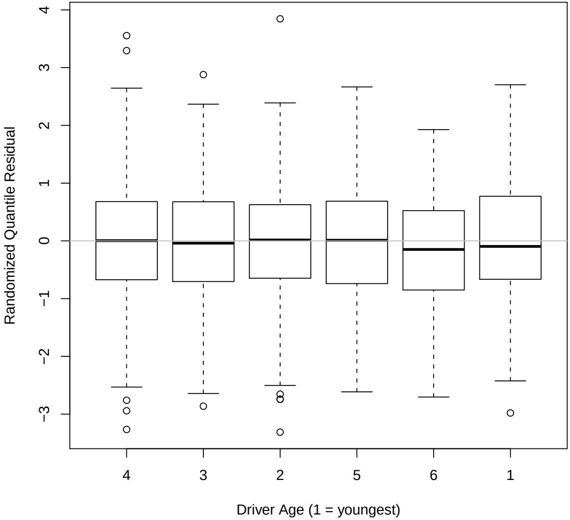
Studentized Residuals have been scaled by an estimate of the unknown scale parameter.

Adjusted The residual has been adjusted from its original definition to bring higher moments in line with the normal distribution.

Randomized Quantile Residual Plot



Quantile Residuals vs. Driver Age



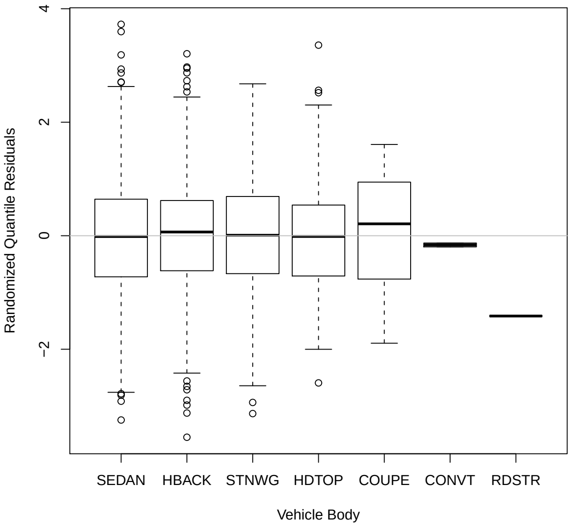
Parameter Estimates Two Variable Model

```
Call:
glm(formula = claim ~ vv + vv2 + vv3 + age.cat,
    family = binomial(link = logitexp(expo)), ...)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.6205     0.3981  -4.070  4.7e-05 ***
vv            0.1149     0.3797   0.303 0.762232
vv2           0.4440     0.5617   0.790 0.429285
vv3          -0.9598     0.3608  -2.660 0.007808 **
age.cat 3    -0.4237     0.2177  -1.946 0.051602 .
age.cat 2    -0.2388     0.2264  -1.055 0.291546
age.cat 5    -0.4734     0.2370  -1.998 0.045758 *
age.cat 6    -1.6637     0.4467  -3.724 0.000196 ***
age.cat 1     0.1398     0.2791   0.501 0.616583
    Null deviance: 1598.3  on 2999  degrees of freedom
Residual deviance: 1503.4  on 2991  degrees of freedom
```

New Estimates with Merged Levels

```
Call:
glm(formula = claim ~ vv + vv2 + vv3 + age.cat2,
    family = binomial(link = logitexp(expo)), ...)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.86257    0.38295  -4.864 1.15e-06 ***
vv           0.09459    0.37896   0.250  0.80289
vv2          0.47624    0.56028   0.850  0.39532
vv3         -0.98212    0.35880  -2.737  0.00620 **
age.cat2 5  -1.40331    0.43102  -3.256  0.00113 **
age.cat2 6   0.40216    0.25279   1.591  0.11163

    Null deviance: 1598.3  on 2999  degrees of freedom
Residual deviance: 1509.0  on 2994  degrees of freedom
```

Quantile Residuals vs. Vehicle Body



How well does our model predict?

Let the threshold probability be 15%. Then, on our build dataset, we have

Build Dataset		Predicted Claim		
		No	Yes	Total
Actual	No	2 568	207	2 775
Claim	Yes	192	33	225
	Total	2 760	240	3 000

$$\text{Error rate} = (207 + 192)/3000 = 0.133$$

$$\text{Sensitivity} = 33/225 = 0.147$$

$$\text{Specificity} = 2\,568/2\,775 = 0.925$$

Predictions against test and validate datasets?

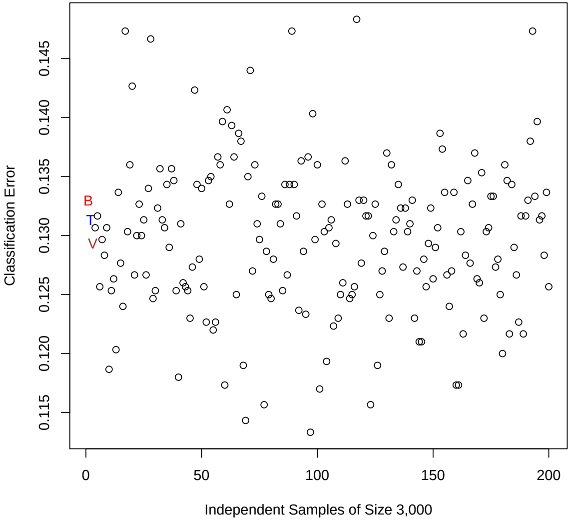
Test Dataset		Predicted Claim		
		No	Yes	Total
Actual	No	2 573	220	2 793
Claim	Yes	174	33	207
	Total	2 747	253	3 000

Error rate $= (220 + 174)/3000 = 0.131$

Validate Dataset		Predicted Claim		
		No	Yes	Total
Actual	No	2 584	231	2 815
Claim	Yes	157	28	185
	Total	2 741	259	3 000

Error rate $= (231 + 157)/3000 = 0.129$

Classification error across many samples



Personal Injury Claims

The dataset (see [3]) 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

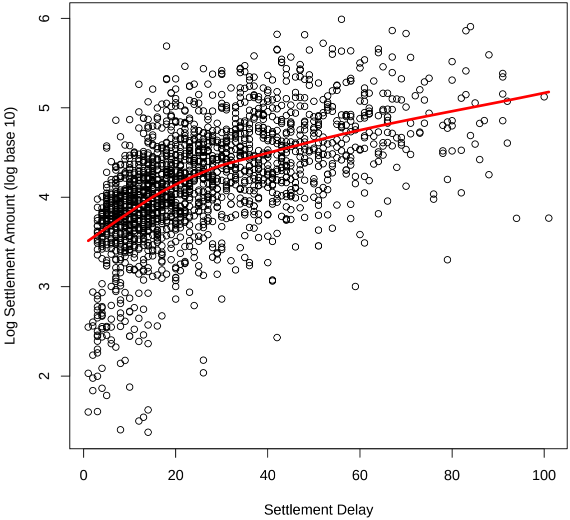
We will work with a random sample of 2,000 claims.

Summary Statistics (for random sample)

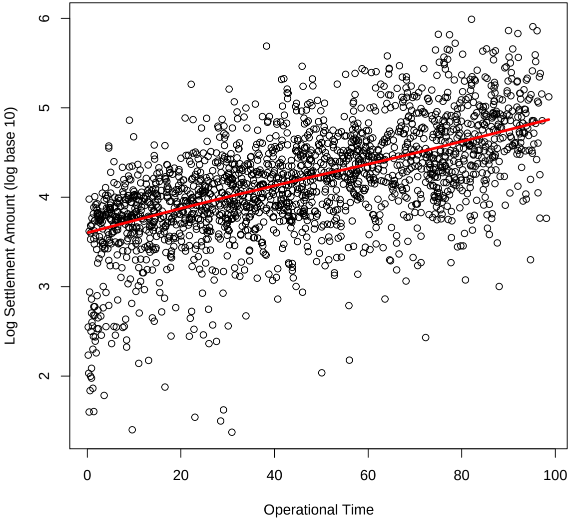
	Claim Amount
Minimum	24
1st Quartile	6,144
Median	14,222
Mean	37,525
3rd Quartile	35,435
Maximum	976,379

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

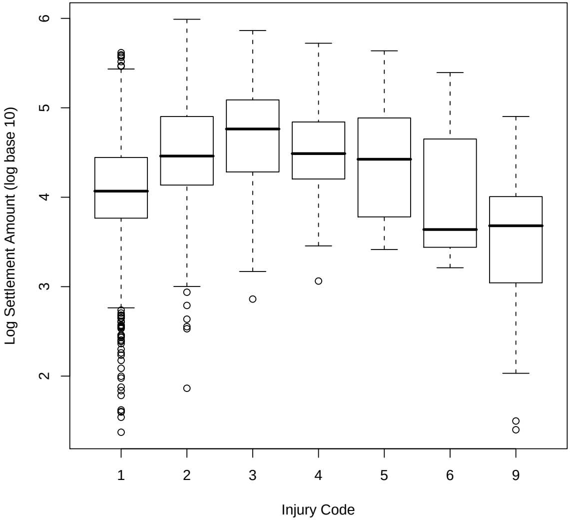
Exploratory Plots I



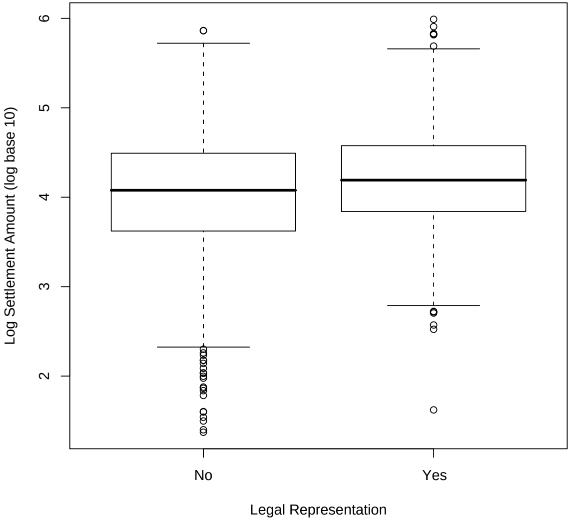
Exploratory Plots II



Exploratory Plots III



Exploratory Plots IV



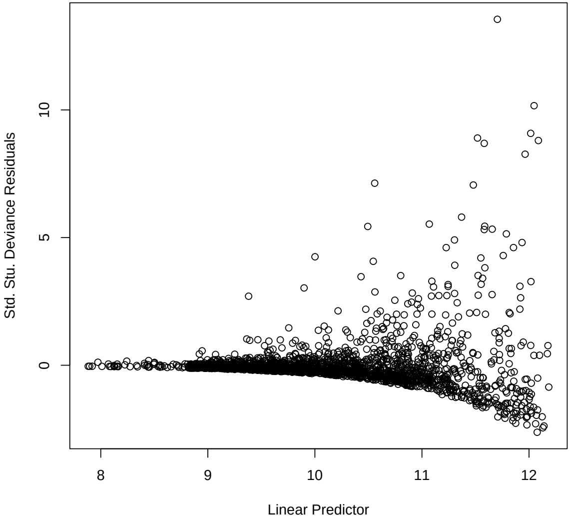
Normal log-link model

$$\log(\text{Settlement Amount}) = \text{Op.Time} + \text{Injury} + \text{Attorney}$$

	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



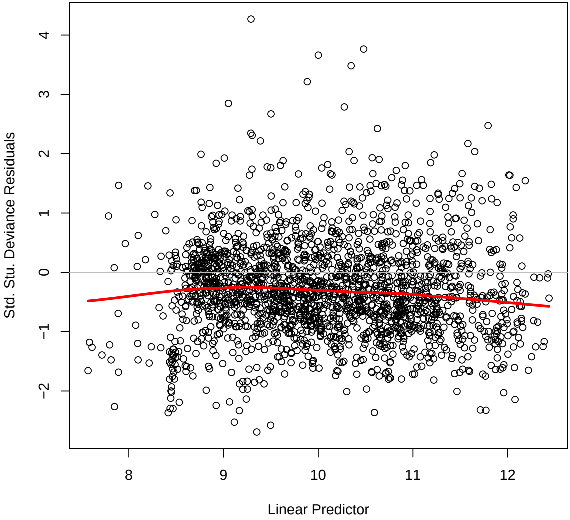
Gamma log-link model

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

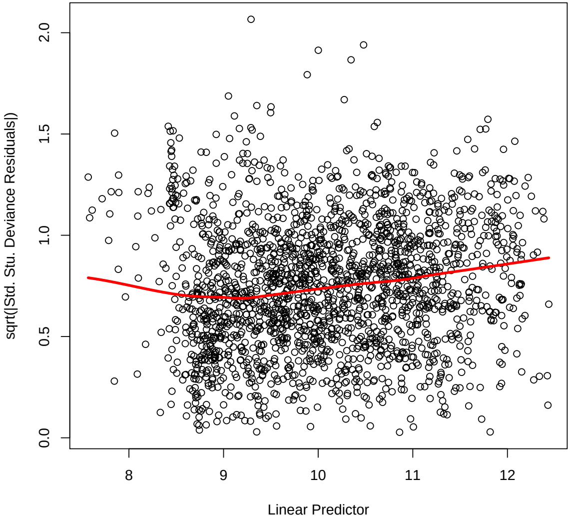
	Estimate	Std. Error	t value	Pr(>|t|)
Intercept)	8.425	0.064	130.69	< 2e-16
Op.Time	0.030	0.001	29.67	< 2e-16
injury 2	0.707	0.074	9.49	< 2e-16
injury 3	0.900	0.116	7.75	1.46e-14
injury 4	1.045	0.271	3.85	0.0001
injury 5	0.279	0.323	0.86	0.39
injury 6	0.199	0.247	0.80	0.42
injury 9	-0.864	0.129	-6.68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

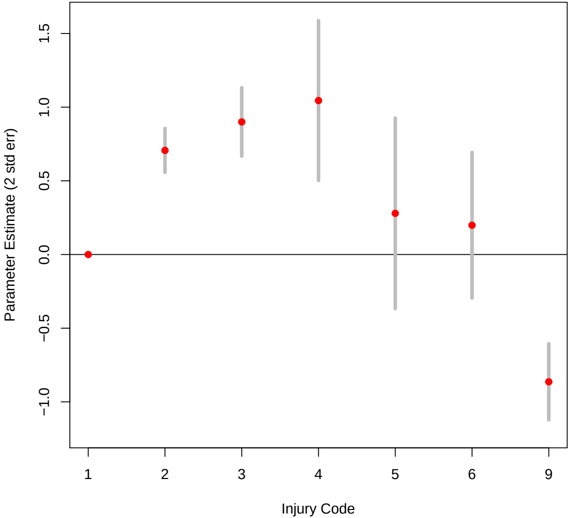


Analysis of Deviance Table

```
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	2 072			
2	1 $\underline{2\ 3\ 4}$ 5 6 9	2 077	5	2	5.9
3	1 $\underline{2\ 3\ 4}$ $\underline{5\ 6}$ 9	2 077	5	3	7.8
4	$\underline{1\ 5\ 6}$ $\underline{2\ 3\ 4}$ 9	2 079	7	4	9.5
5	1 $\underline{2\ 3\ 4\ 5\ 6}$ 9	2 086	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

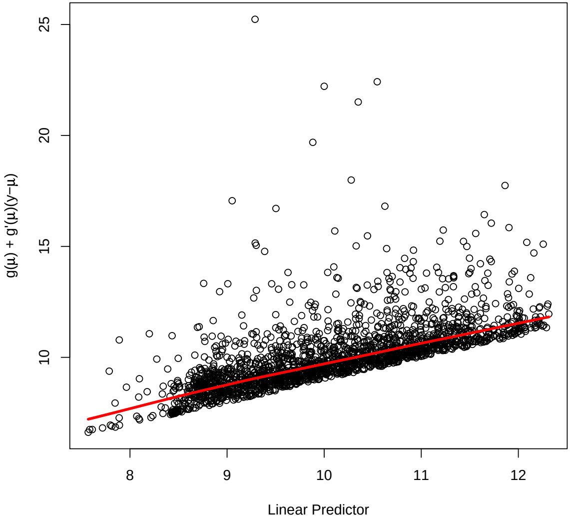
Two ways to assess the link function:

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

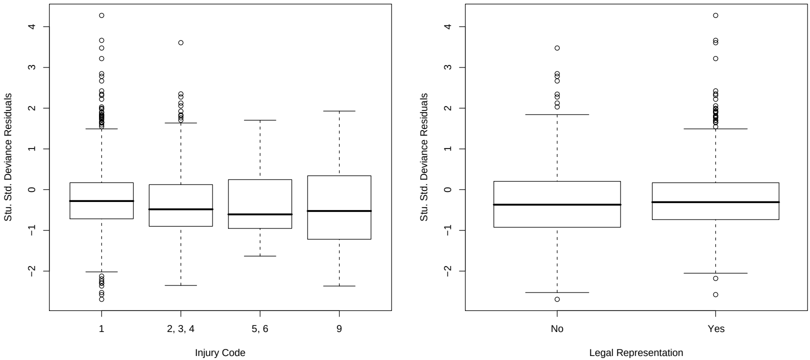
$$x_i\beta = g(y_i) \approx g(\mu_i) + g'(\mu_i)(y_i - \mu_i)$$

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

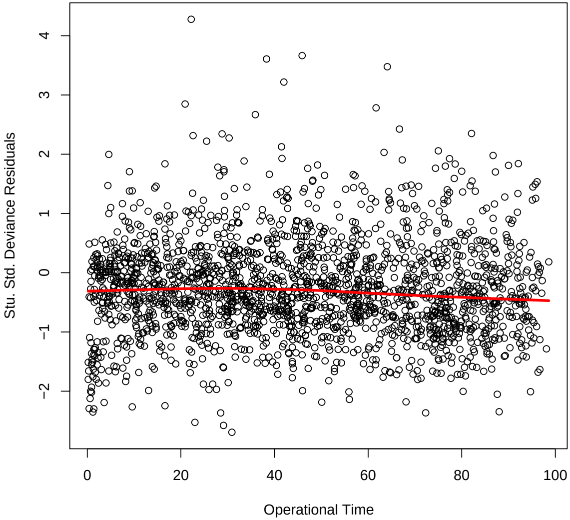
Checking the Link Function



Checking Explanatory Variables



Checking Explanatory Variables

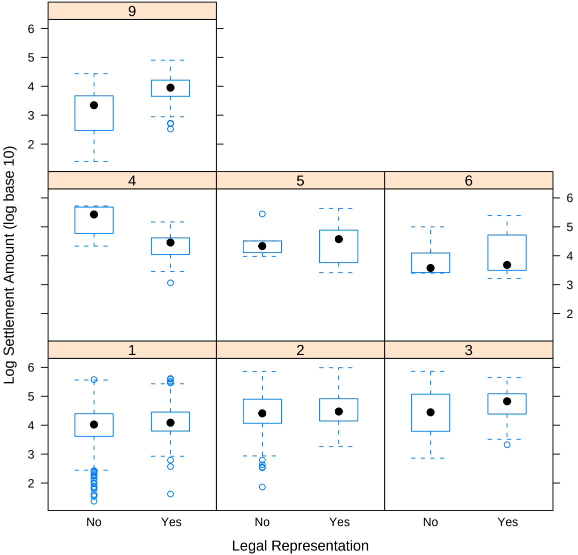


Interactions

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

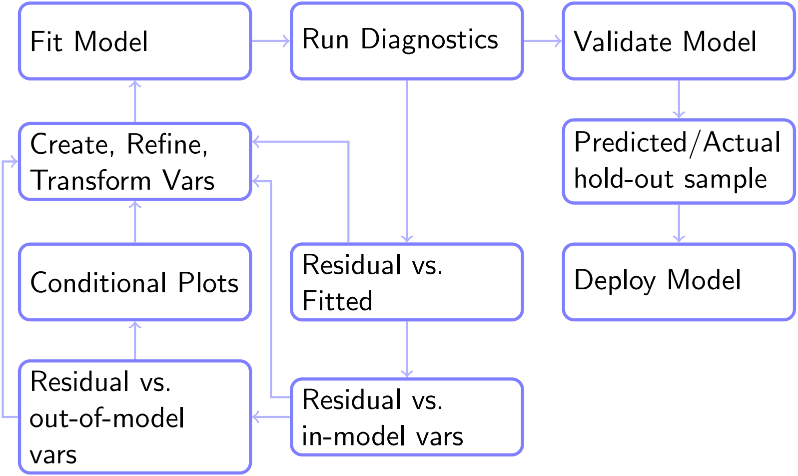
Out-of-Sample Validation

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 Cycle



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.