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

Bentley University

Casualty Actuarial Society
Ratemaking and Product Management Seminar
March 25–27, 2019
Boston, MA

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

Basic GLM Specification

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

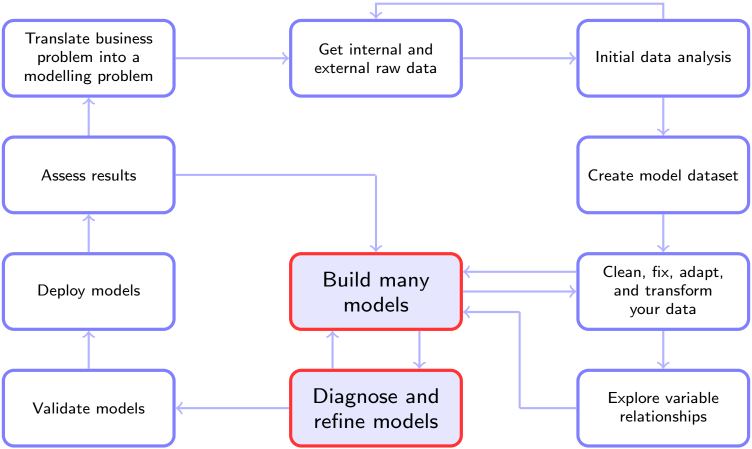
1. The link function is g
2. The distribution of y is a member of the exponential family
3. The explanatory variables x_i may be continuous or discrete
4. The offset term can be used to adjust for exposure or to introduce known restrictions

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

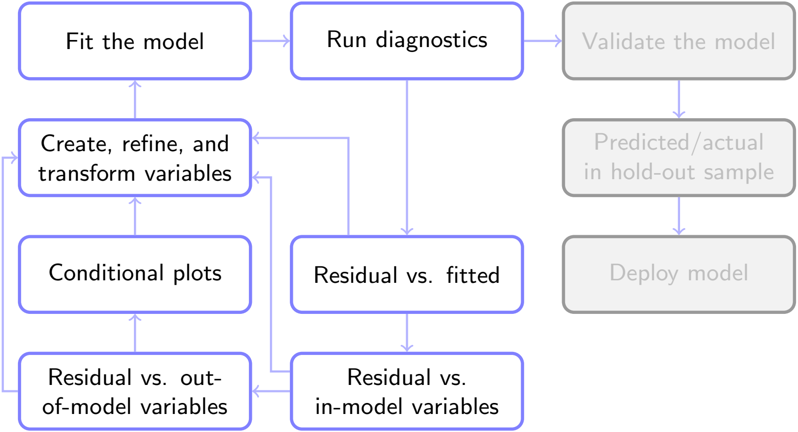
Common Model Forms

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

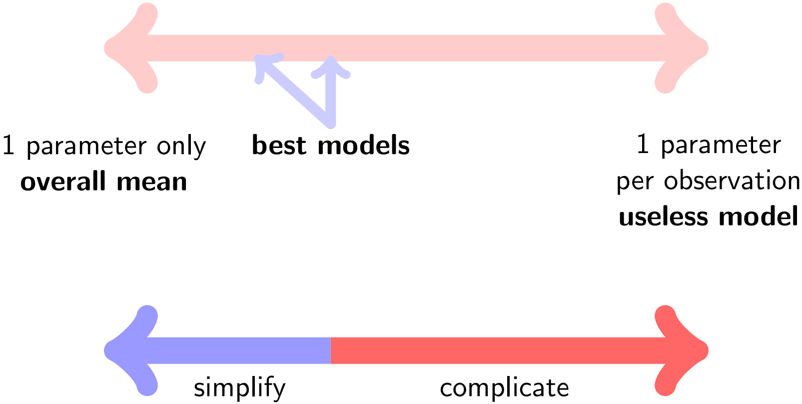
Overall Project Cycle



Model Building Cycle



Model Complexity



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

Exploratory Analysis

- Tabular summaries
- Univariate exploration (along with exposure)
- Bivariate relationships
- Correlations

Preparing to Stay Honest

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

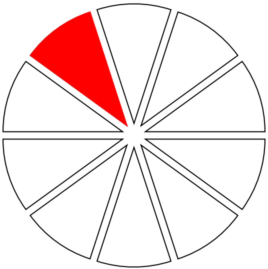
One rule of thumb: $(50\%, 25\%, 25\%)$.

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
```

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?

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
```

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
```

$$e^{-2.61397} = 0.0732$$

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
```

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

Initial Frequency Model

Variables **included** in the model:

1. gender
2. area
3. age category

Variables **not included** in the model:

1. vehicle body
2. vehicle age
3. vehicle value

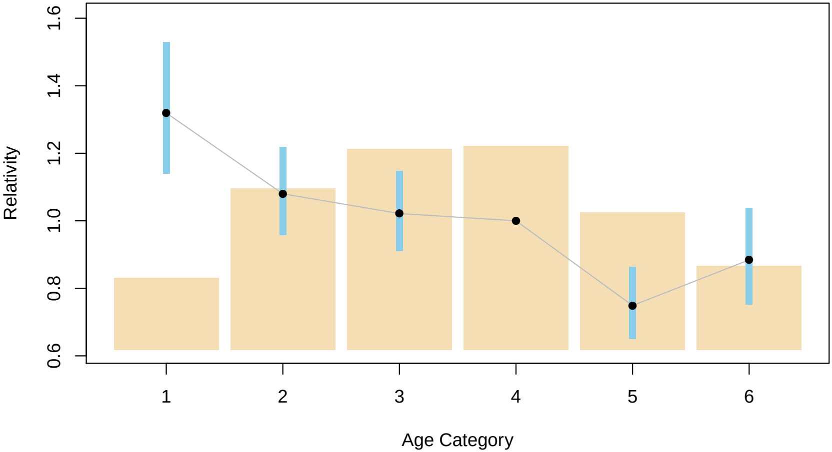
Initial Frequency Model

```
glm(num.claims ~ gender + area + age.cat,

    data = dta, subset = b.idx,
    family = poisson(link = "log"),
    offset = log(exposure))
```

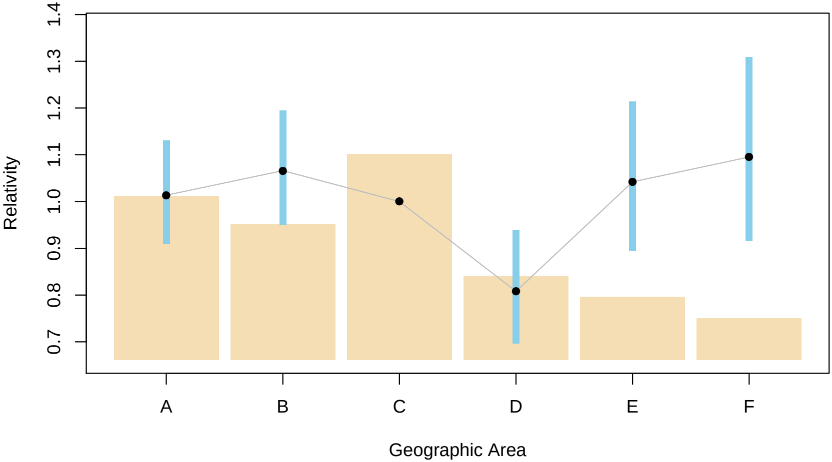
FQ Estimated Coefficients

```
             Estimate Std. Error z value  Pr(>|z|)
(Intercept) -1.849782   0.053461 -34.601   < 2e-16 ***

gender M    -0.008281   0.040575  -0.204  0.838279

area A       0.013438   0.054652   0.246  0.805779
area B       0.064011   0.057184   1.119  0.262975
area D      -0.213205   0.074610  -2.858  0.004269 **
area E       0.041201   0.076134   0.541  0.588394
area F       0.091002   0.089094   1.021  0.307058

age.cat 1    0.277251   0.073570   3.769  0.000164 ***
age.cat 2    0.077001   0.060403   1.275  0.202387
age.cat 3    0.021269   0.057824   0.368  0.713002
age.cat 5   -0.288950   0.070964  -4.072  4.67e-05 ***
age.cat 6   -0.123044   0.080516  -1.528  0.126464
```

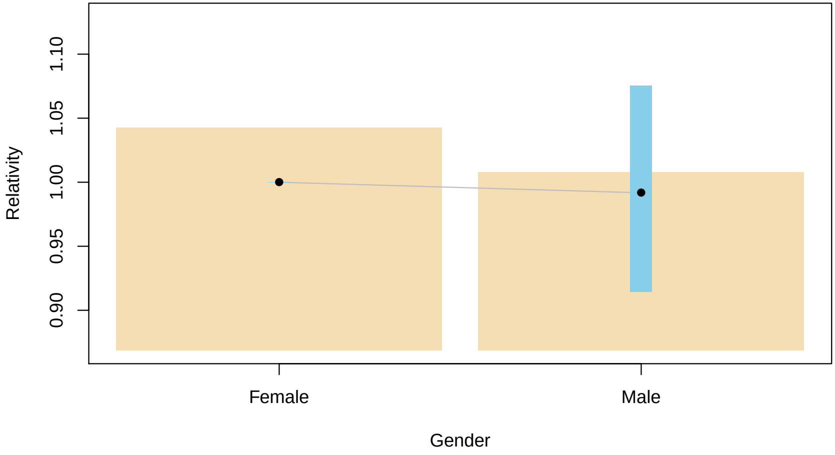
FQ Estimated Age Parameters



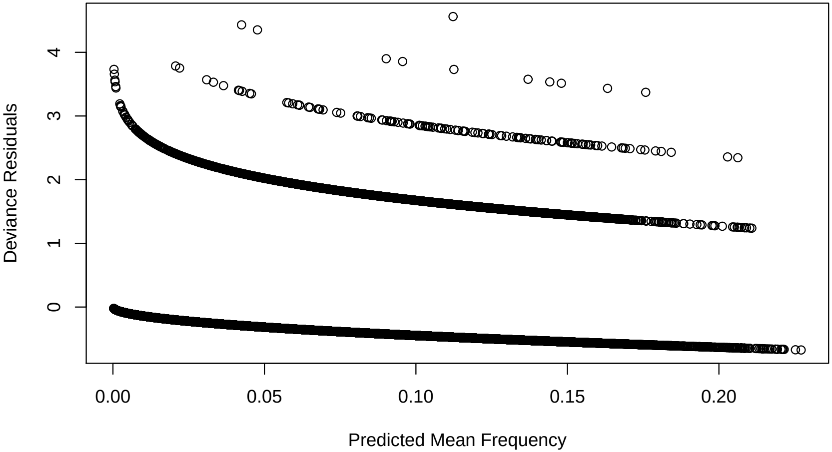
FQ Estimated Geographic Area Parameters



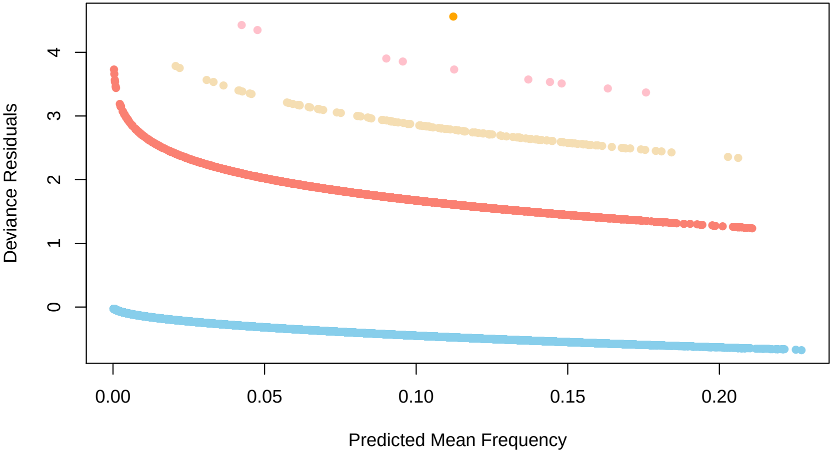
FQ Estimated Gender Parameters



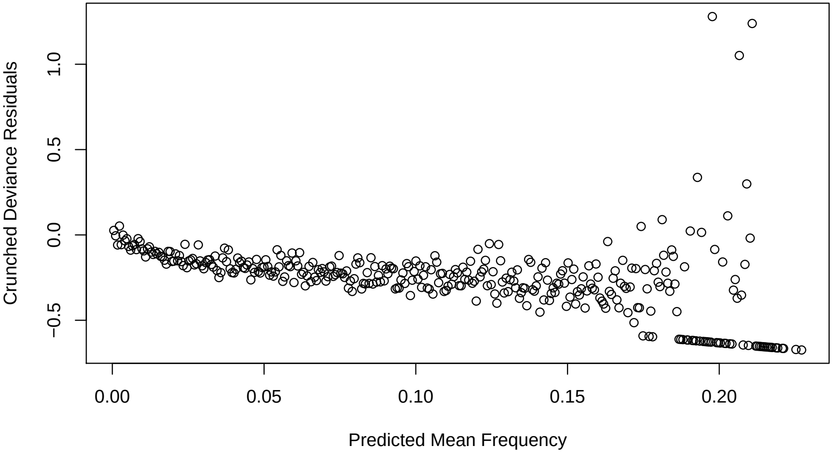
FQ Deviance Residuals



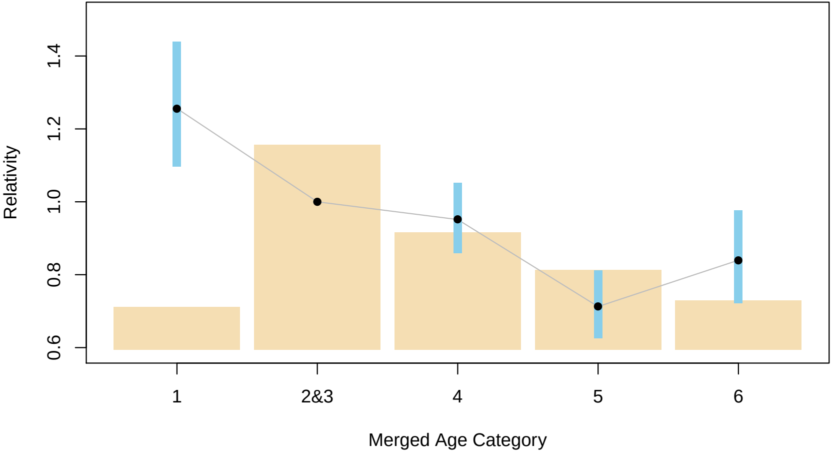
FQ Deviance Residuals



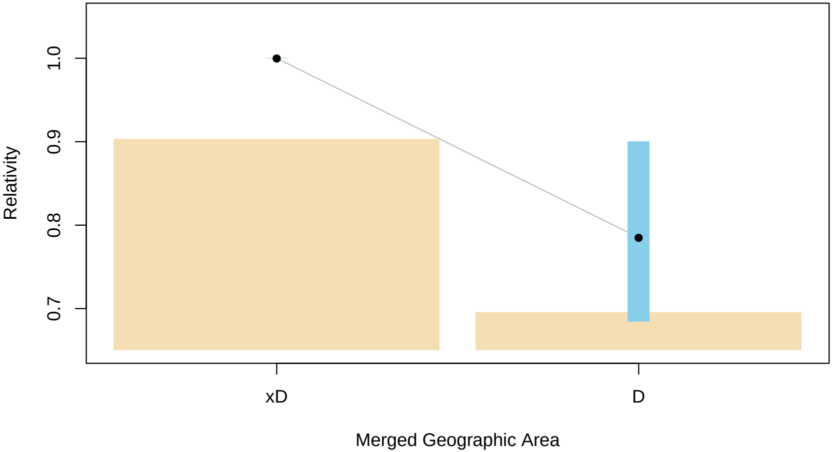
FQ Crunched Deviance Residuals



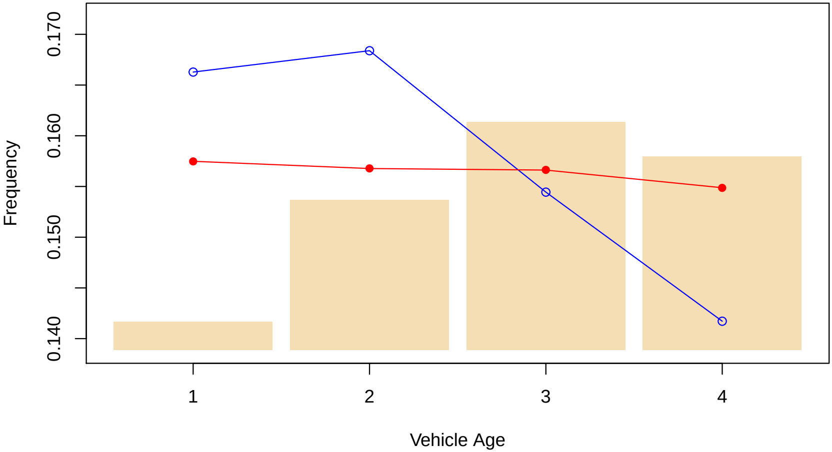
FQ Estimated Merged Age Parameters



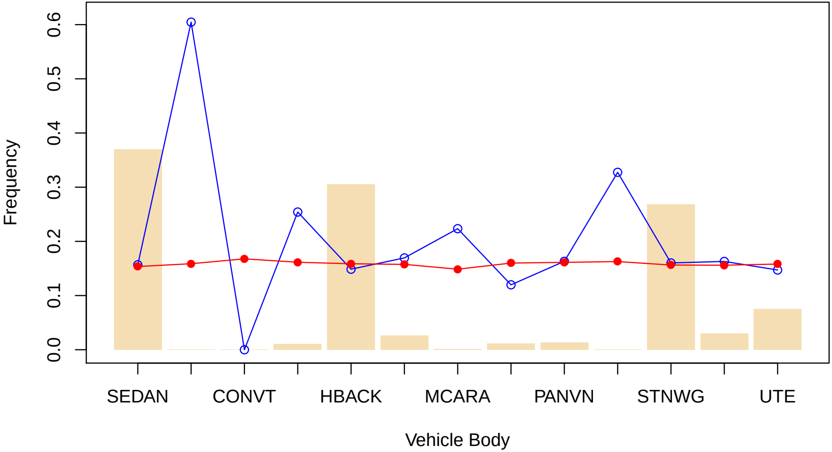
FQ Estimated Merged Geographic Area Parameters



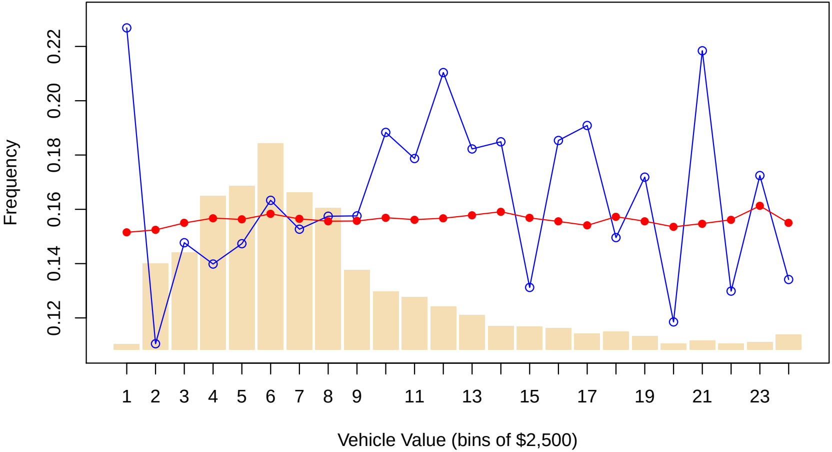
FQ Actual vs. Expected Vehicle Age



FQ Actual vs. Expected Vehicle Body



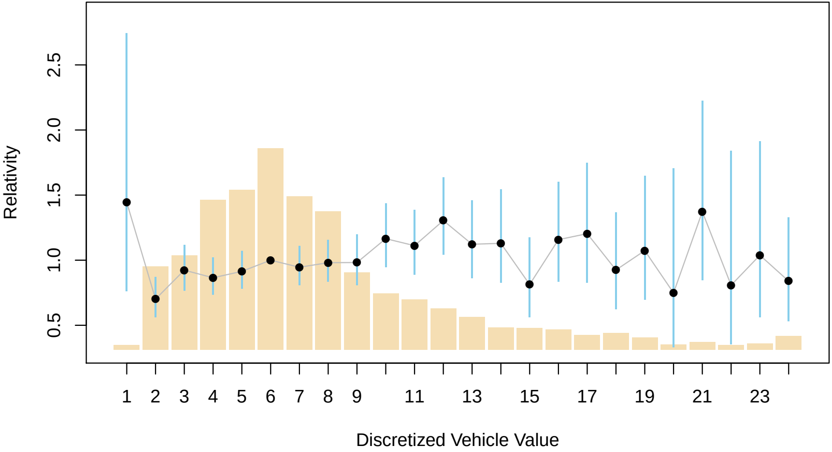
FQ Actual vs. Expected Vehicle Value



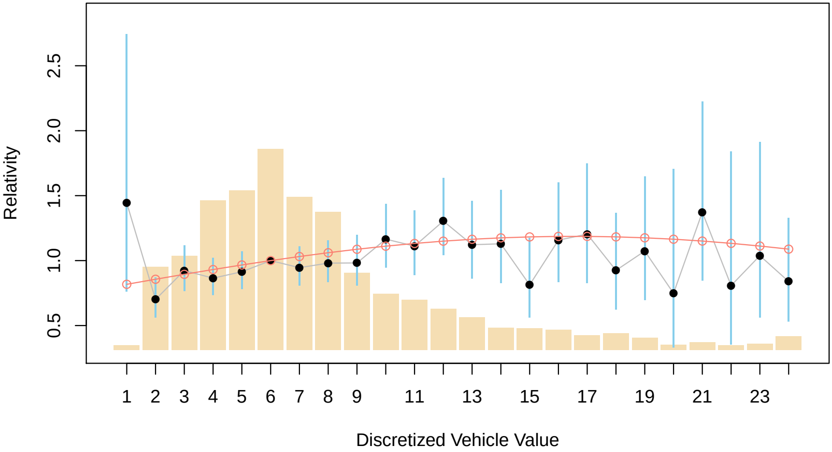
New Frequency Model with Vehicle Value

```
glm(num.claims ~ area.2 + age.cat.2 + veh.val.cat,

    data = dta, subset = b.idx,
    family = poisson(link = "log"),
    offset = log(exposure))
```

FQ Estimated Vehicle Value Parameters



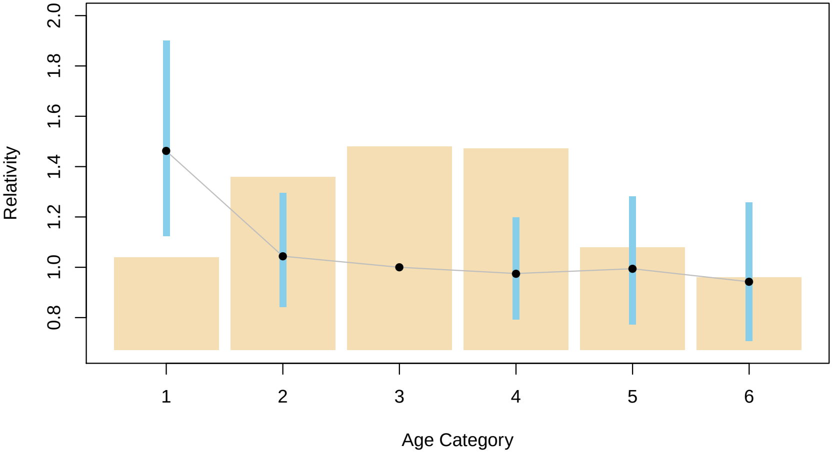
FQ Estimated Polynomial Vehicle Value Parameters



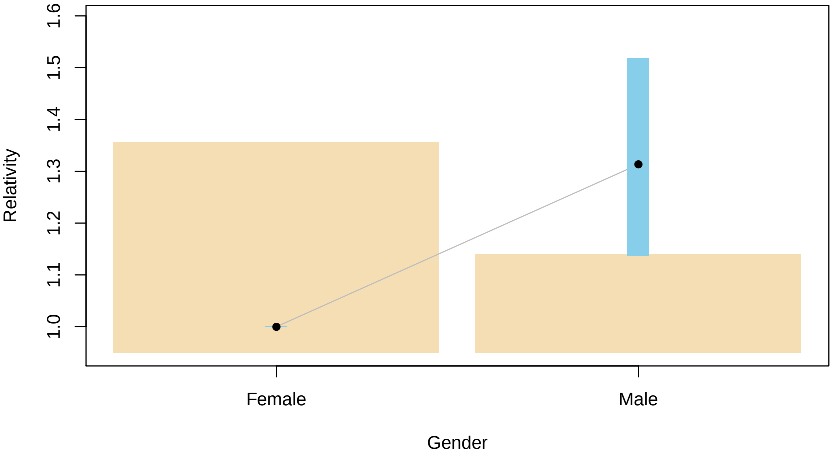
Severity Modeling

```
glm(avg.cost ~ age.cat + gender,

    data = dtb, subset = b.idx,
    family = Gamma(link = "log"),
    weights = num.claims)
```

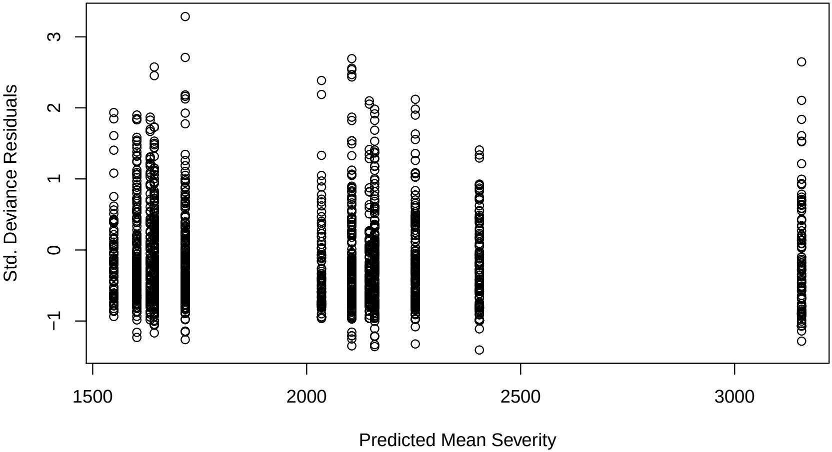
SV Estimated Age Parameters



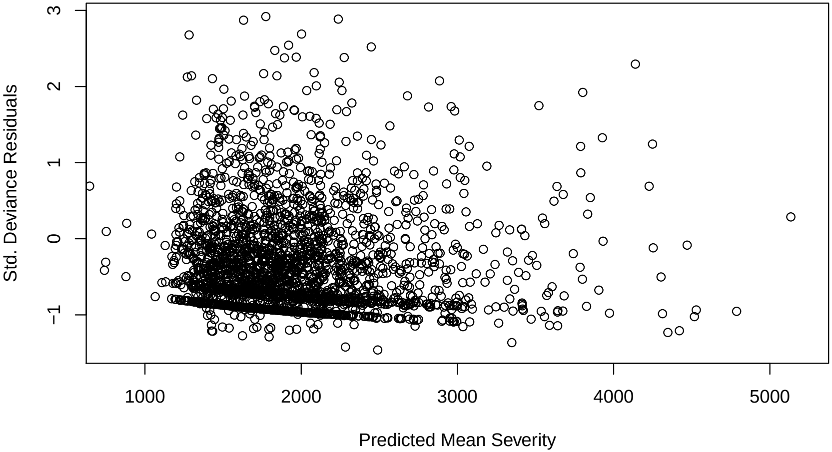
SV Estimated Gender Parameters



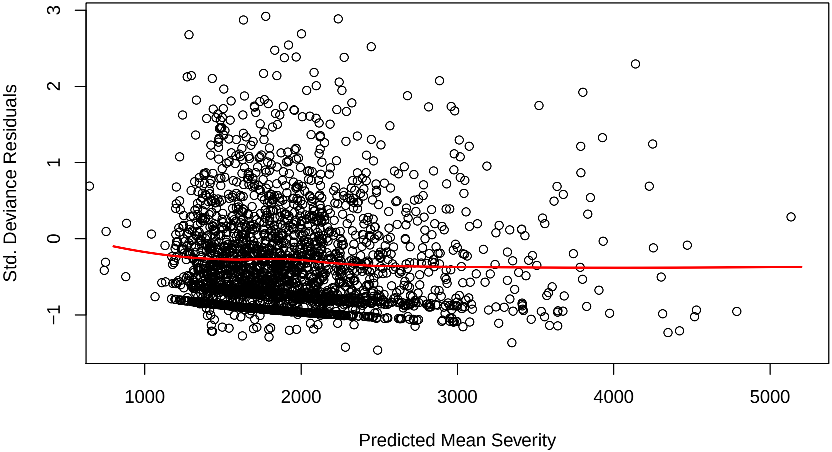
SV Std. Deviance Residuals vs. Predicted Values



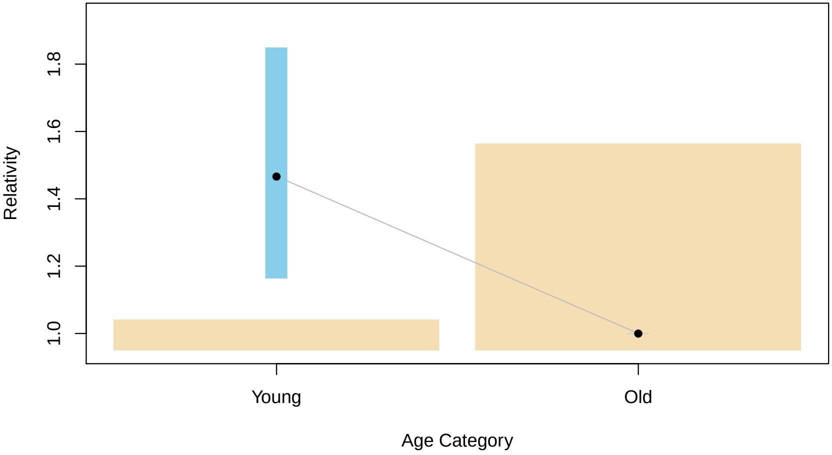
Residuals for All Main Effects Variables



Residuals for All Main Effects with a Smooth



SV Estimated Merged Age Parameters



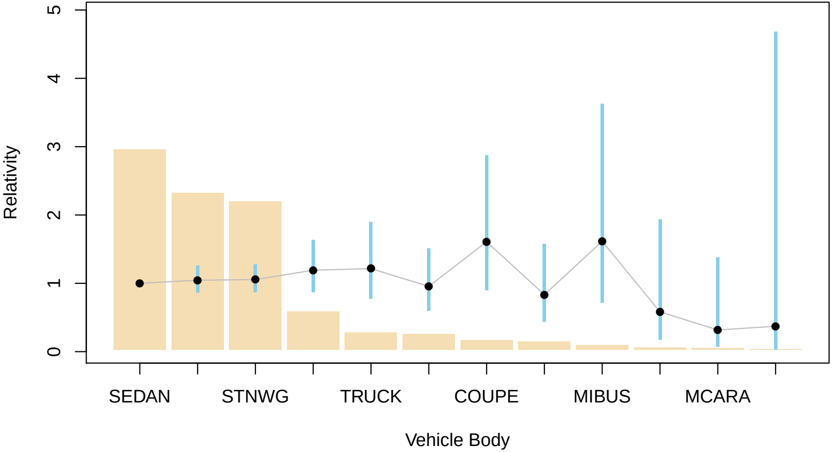
SV Include Vehicle Body

```
glm(avg.cost ~ age.cat.2 + gender + veh.body,

    data = dtb, subset = b.idx,
    family = Gamma(link = "log"),
    weights = num.claims)
```

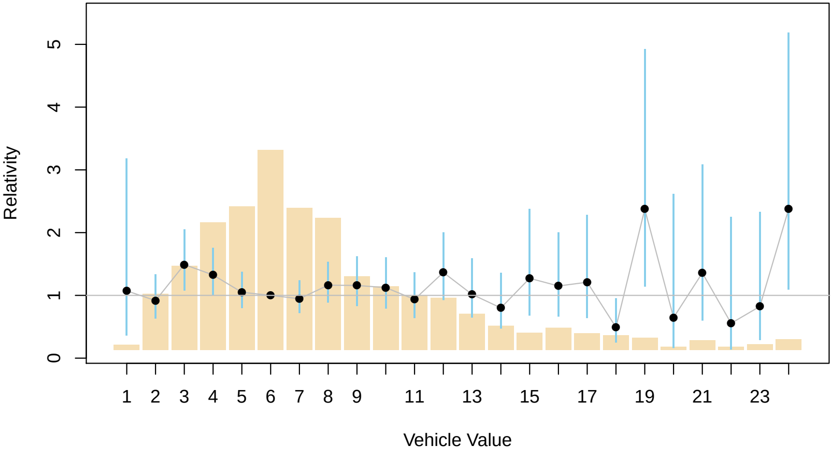
Analysis of Deviance Table

```
Response: avg.cost

Terms added sequentially (first to last)

           Df  Diff  Resid   Resid
                Dev     Df     Dev      F   Pr(>F)
NULL                  2328  3947.8
age.cat.2   1 45.097  2327  3902.7  14.10  0.00018 ***
gender      1 44.111  2326  3858.6  13.79  0.00021 ***
veh.body   11 31.073  2315  3827.5   0.88  0.55657
```

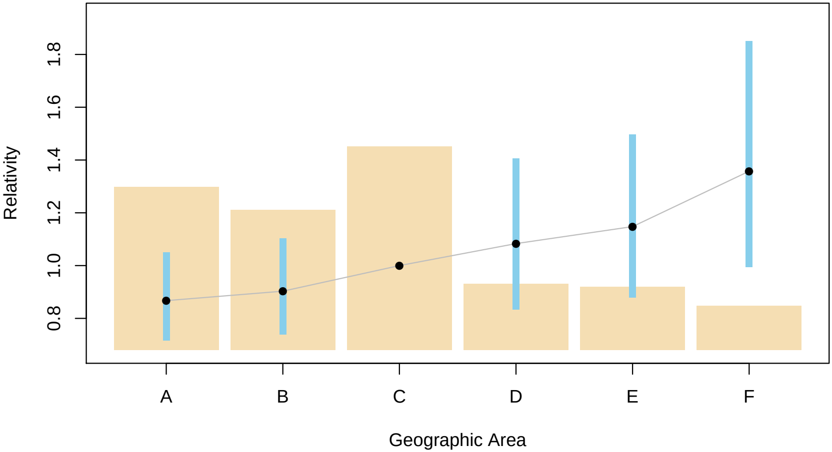
SV Estimated Vehicle Body Parameters



SV Estimated Vehicle Value Parameters

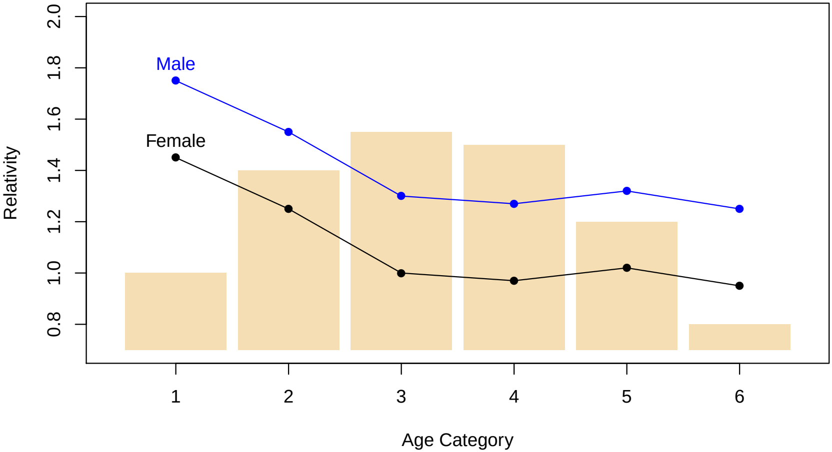


SV Estimated Geographic Area Parameters

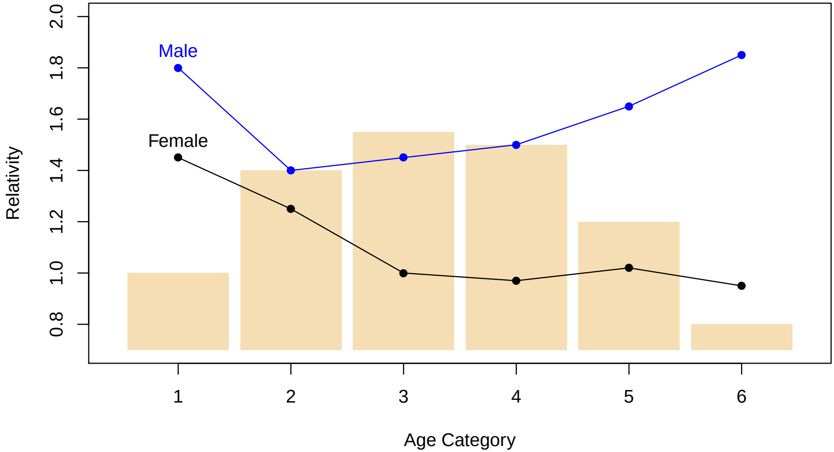


Interactions?

No interaction between Age and Gender



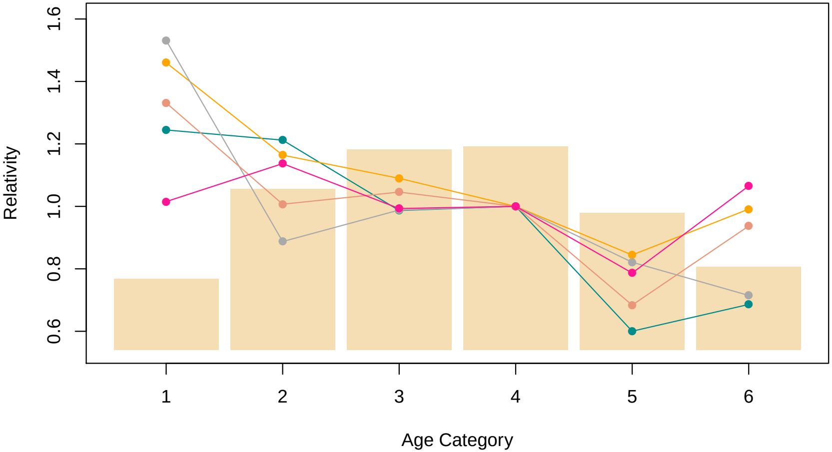
Interaction between Age and Gender



Interaction between Age and Gender?

```
Response: avg.cost

Terms added sequentially (first to last)

              Diff Resid   Resid
          Df   Dev    Df     Dev        F Pr(>F)
NULL                2328  3947.8
age.2      1 45.097 2327  3902.7  14.2497 0.0002 ***
gender     1 44.111 2326  3858.6  13.9380 0.0002 ***
age.2:gender 1 10.971 2325  3847.6   3.4666 0.0627 .
```

Consistency Across Time or Random Subsets

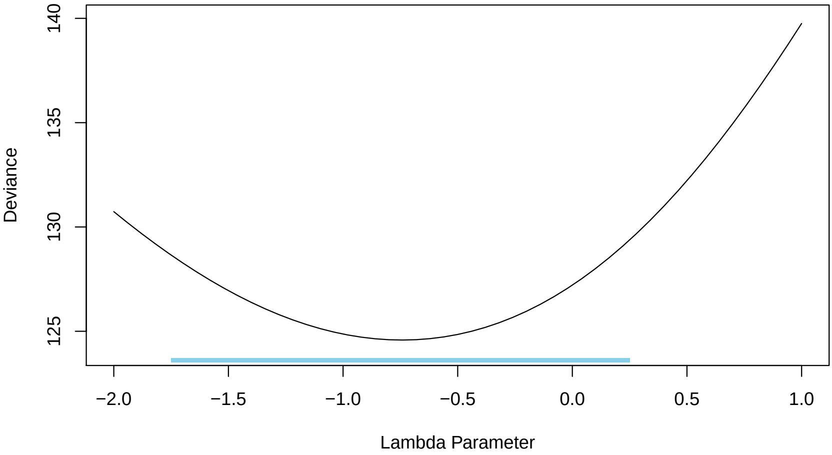


Checking the Link Function

Embed the link function in a family of functions. For example,

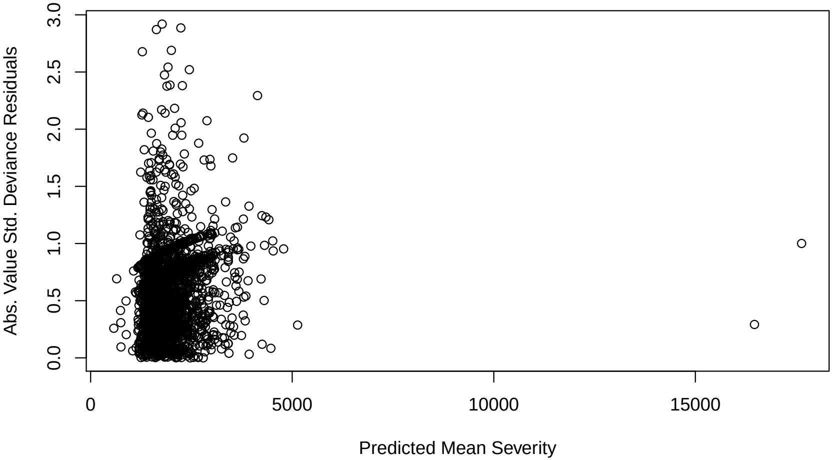
$$\text{link}(\mu) = \begin{cases} \mu^{\lambda} & \text{for } \lambda \neq 0, \\ \log \mu & \text{for } \lambda = 0. \end{cases}$$

Deviance as λ varies

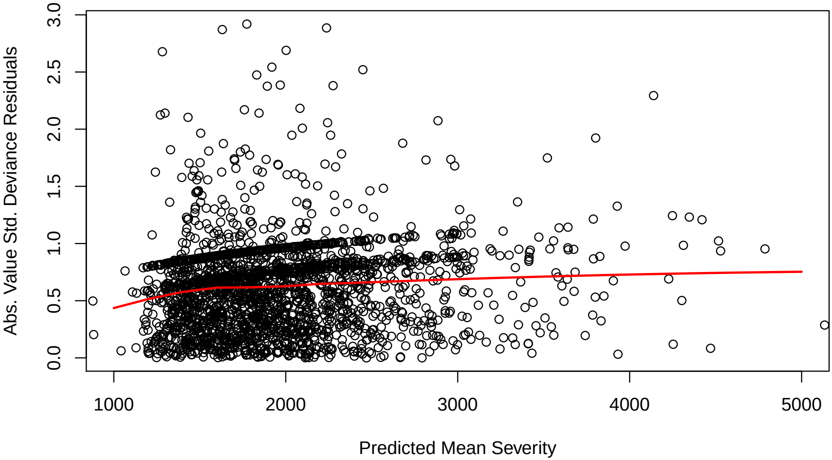


Source: Modified version of Fig. 11.1 in McCullagh & Nelder p. 377.

Checking the Variance Function



Checking the Variance Function



Constraints via the Offset

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

1. Regulatory constraints
2. Own-company constraints

Refitting causes correlated variables to partially adjust.

Summary

- Exploratory analysis
- Build, test, validate
- Cross-validation
- Start with simple models
- Simplify
- Complicate
- Analysis of deviance table
- Residual plots
- Embed link/variance in a family
- Lift curves
- Other graphical methods

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.

P. McCullagh and J. A. Nelder.
Generalized Linear Models.
2nd Ed. Chapman & Hall, 1989.

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.