

# 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 I: Introduction to Generalized Linear Models

Ernesto Schirmacher

Bentley University

Casualty Actuarial Society  
Ratemaking and Product Management Seminar  
March 25–27, 2017  
Boston, MA

# Overview

Overview of GLMs

Personal Injury Claims

Intercept Only Models

One Continuous Predictor

One Discrete Predictor

Many Predictors

Key Concepts

# Standard Linear Model Specification

$$y = \beta_0 + x_1\beta_1 + \cdots + x_k\beta_k + \epsilon \quad \text{with } \epsilon \in N(0, \sigma^2)$$

# Standard Linear Model Specification

$$y = \beta_0 + x_1\beta_1 + \cdots + x_k\beta_k + \epsilon \quad \text{with } \epsilon \in N(0, \sigma^2)$$

A better way to think about this would be

$$\mathbb{E}[y] = \beta_0 + x_1\beta_1 + \cdots + x_k\beta_k$$

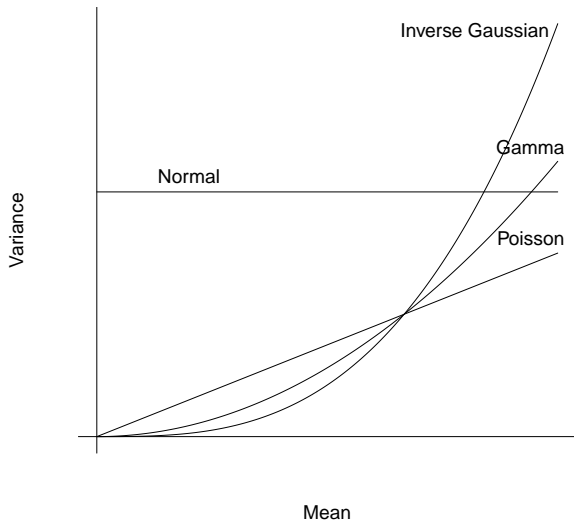
where  $y \in N(\mu, \sigma^2)$  and  $\mu = \beta_0 + x_1\beta_1 + \cdots + x_k\beta_k$  is the linear predictor.

# Generalized Linear Model 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. Offset terms have a known coefficient of 1 in the linear predictor

# Mean–Variance Relationship



# Personal Injury Dataset

The dataset contains 22,036 settled personal injury claims. These claims arose from accidents occurring from July 1989 through January 1999. This is the `persinj.xls` dataset featured in the book by de Jong & Heller [2].

I have taken a random sample of 200 claims.

The variables are:

1. Settled Amount
2. Injury codes
3. Legal representation
4. Accident month
5. Report month
6. Finalization month
7. Operational time

Derived variables:

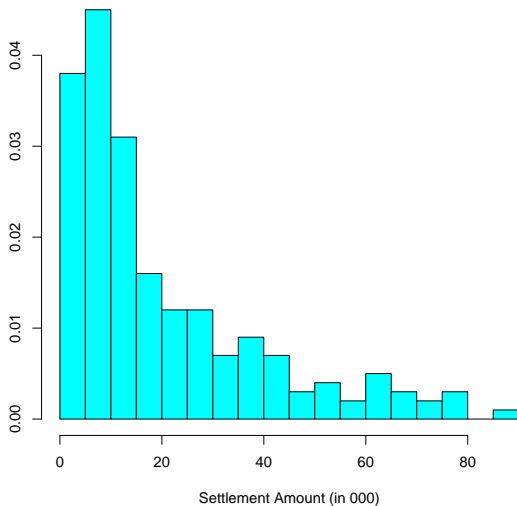
1. Injured count
2. Accident injury code
3. Report delay
4. Settlement delay



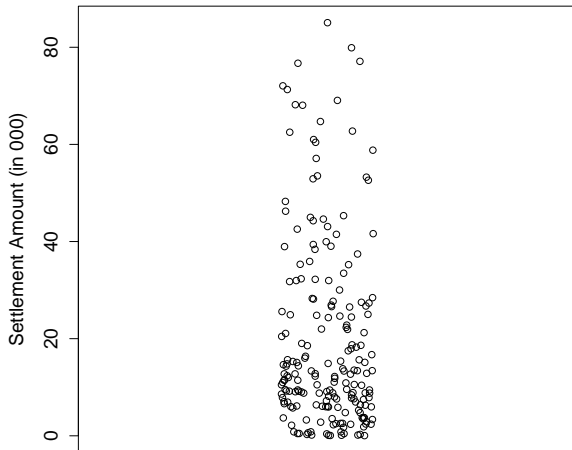
## Variable Descriptions

Variable	Type	Comments
Settled Amount	Cont	range: \$40 to \$85,000
Injury Codes	Cat	Injury level: 1, 2, . . . , 6 = death, 9 = missing
Legal Rep.	Bin	Attorney involved? 1 = Yes, 0 = No
Accident Month	Coded	1 = July 1989, 120 = June 1999
Report Month	Coded	same as accident month
Fin. Month	Coded	same as accident month
Injured Count	Count	Number of persons injured: 1, 2, . . . , 5
Acc. Injury	Cat	Highest injury code among those injured
Report Delay	Cont	# months between accident and report
Settle. Delay	Cont	# months between report and settlement

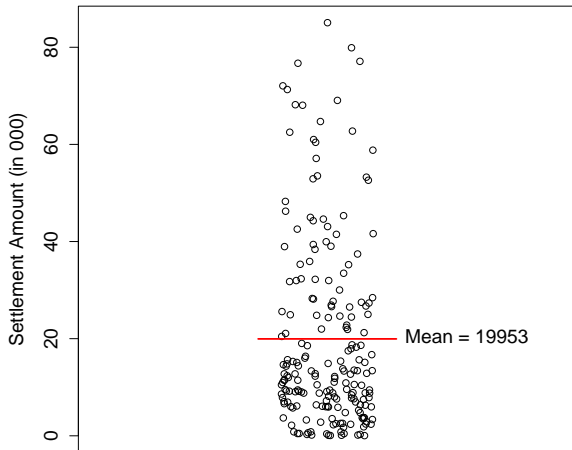
# Histogram of Settlement Amount



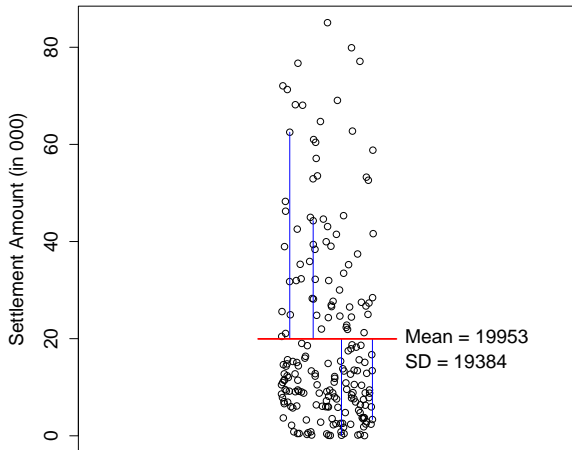
# Distribution of Settlement Amount



## Settlement Amount: mean



## Settlement Amount: mean & standard deviation



## Linear Model—Intercept only

Call:

```
lm(formula = total ~ 1, data = spinj)
```

Residuals:

Min	1Q	Median	3Q	Max
-19913	-13570	-7199	7591	65110

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	19953	1371	14.56	<2e-16 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19380 on 199 degrees of freedom

## Generalized Linear Model—Normal Id—Intercept only

```
Call: glm(formula = total ~ 1,
          family = gaussian(link = identity), data = spinj)
```

```
Deviance Residuals:
```

Min	1Q	Median	3Q	Max
-19913	-13570	-7199	7591	65110

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	19953	1371	14.56	<2e-16 ***

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for gaussian family taken to be 375744867)
```

```
Null deviance: 7.4773e+10  on 199  degrees of freedom  
Residual deviance: 7.4773e+10  on 199  degrees of freedom  
AIC: 4519.5
```

```
Number of Fisher Scoring iterations: 2
```

## Generalized Linear Model—Gamma Id—Intercept only

```
Call: glm(formula = total ~ 1,
           family = Gamma(link = identity), data = spinj)
```

```
Deviance Residuals:
```

Min	1Q	Median	3Q	Max
-3.2293	-0.9588	-0.4165	0.3407	1.9043

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	19953	1371	14.56	<2e-16 ***

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for Gamma family taken to be 0.9438079)
```

```
Null deviance: 252.05  on 199  degrees of freedom  
Residual deviance: 252.05  on 199  degrees of freedom  
AIC: 4366.6
```

```
Number of Fisher Scoring iterations: 3
```



## Generalized Linear Model—Gamma Log—Intercept only

```
Call: glm(formula = total ~ 1,
          family = Gamma(link = "log"), data = spinj)
```

```
Deviance Residuals:
```

Min	1Q	Median	3Q	Max
-3.2293	-0.9588	-0.4165	0.3407	1.9043

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	9.9011	0.0687	144.1	<2e-16 ***

```
---
```

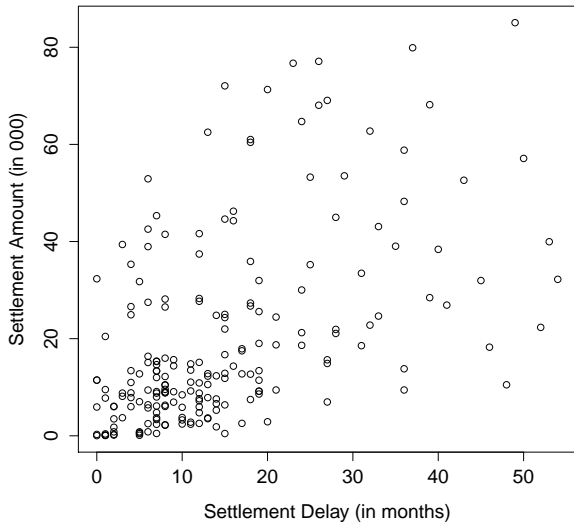
```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for Gamma family taken to be 0.9438079)
```

```
Null deviance: 252.05  on 199  degrees of freedom  
Residual deviance: 252.05  on 199  degrees of freedom  
AIC: 4366.6
```

```
Number of Fisher Scoring iterations: 6
```

## Settlement Amount vs. Settlement Delay



## Linear Model—Intercept and Slope

Call:

```
lm(formula = total ~ settle.delay, data = spinj)
```

Residuals:

Min	1Q	Median	3Q	Max
-37059	-10395	-5085	4366	51957

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	7614.05	1861.85	4.089	6.28e-05 ***
settle.delay	832.30	97.44	8.542	3.50e-15 ***

---

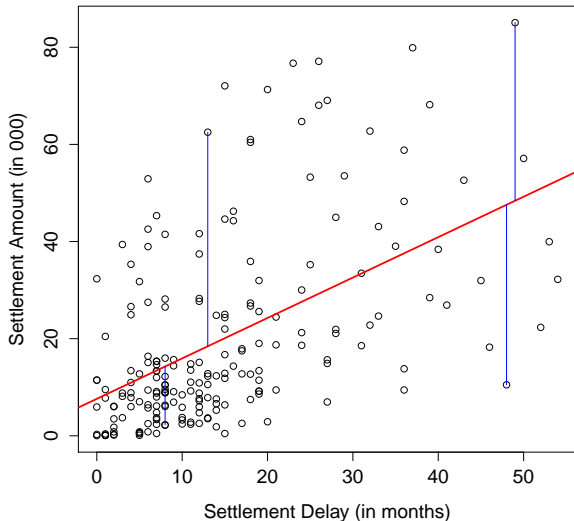
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16610 on 198 degrees of freedom

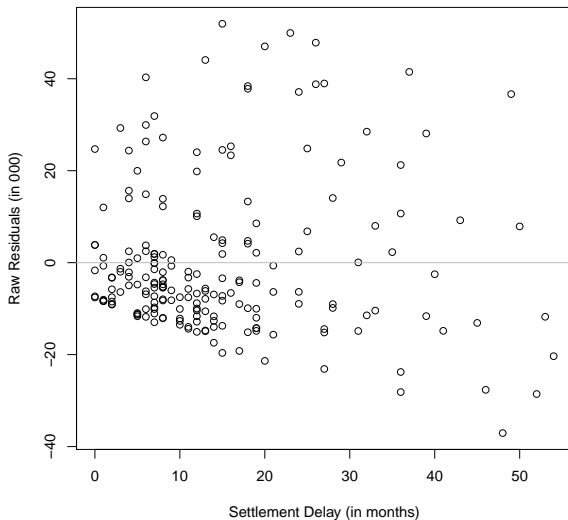
Multiple R-squared: 0.2693, Adjusted R-squared: 0.2656

F-statistic: 72.96 on 1 and 198 DF, p-value: 3.504e-15

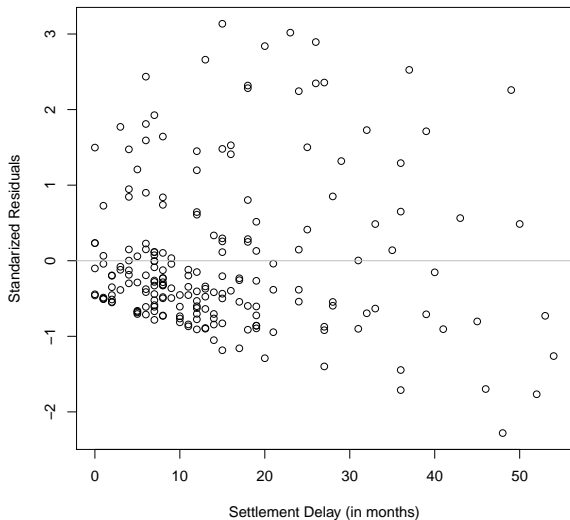
## Settlement Amount vs. Delay: Least Squares Line



## Raw Residuals vs. Settlement Delay



## Standardized Residuals vs. Settlement Delay



# Many Flavors of Residuals

Raw  $y - \hat{y}$  or  $y - \mu$  or  $y - \mathbb{E}[y]$

Pearson  $(y - \mu) / \sqrt{V}$

Deviance  $\text{sgn}(y - \mu) \sqrt{\text{deviance}}$

Standardized Divide residual by  $\sqrt{1 - h}$ , which aims to make its variance constant; where  $h$  are the diagonal elements of the projection ('hat') matrix,  $H = X(X^t X)^{-1} X^t$ , which maps  $y$  into  $\hat{y}$

Studentized Divide residual by  $\sqrt{\phi}$ ; where  $\phi$  is the scale parameter

Stan & Stud Divide residual by both standardized and studentized adjustments

# Deviance

Distribution	Contribution to Squared Deviance
--------------	----------------------------------

---

Normal	$(y_i - \mu_i)^2$
--------	-------------------

Poisson	$2\{y_i \log(y_i/\mu_i) - y_i + \mu_i\}$
---------	--

Gamma	$2\{-\log(y_i/\mu_i) + (y_i - \mu_i)/\mu_i\}$
-------	---

Inverse Gaussian	$(y_i - \mu_i)^2 / (\mu_i^2 y_i)$
------------------	-----------------------------------



## Gamma Log GLM–Intercept and Slope

```
Call: glm(formula = total ~ settle.delay,
          family = Gamma(link = "log"), data = spinj)
```

```
Deviance Residuals:
```

	Min	1Q	Median	3Q	Max
	-3.0008	-0.8017	-0.3145	0.1991	1.8982

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	9.187173	0.102174	89.917	< 2e-16 ***
settle.delay	0.040473	0.005347	7.569	1.39e-12 ***

```
---
```

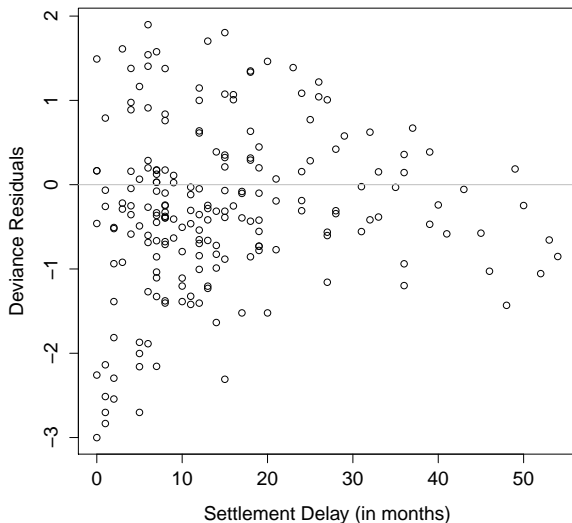
```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for Gamma family taken to be 0.8310652)
```

```
Null deviance: 252.05  on 199  degrees of freedom  
Residual deviance: 206.47  on 198  degrees of freedom  
AIC: 4321.8
```

```
Number of Fisher Scoring iterations: 7
```

## Gamma Model: Deviance Residuals vs. Settlement Delay



## Poisson Log GLM—Intercept and Slope

```
Call: glm(formula = tot.amt ~ settle.delay,
           family = poisson(link = "log"), data = spinj)
```

```
Deviance Residuals:
```

Min	1Q	Median	3Q	Max
-229.41	-92.18	-42.51	35.74	299.99

```
Coefficients:
```

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	9.323e+00	8.583e-04	10862.1	<2e-16 ***
settle.delay	3.280e-02	3.338e-05	982.7	<2e-16 ***

```
---
```

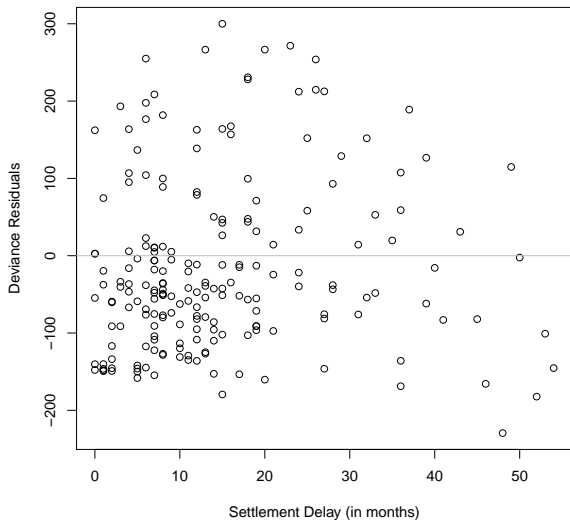
```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for poisson family taken to be 1)
```

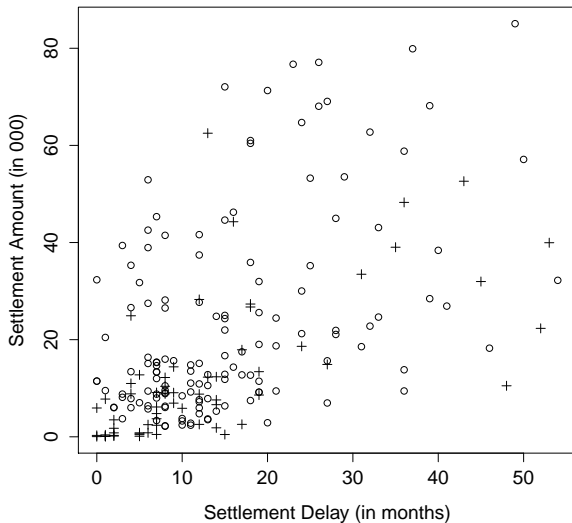
```
Null deviance: 3366902  on 199  degrees of freedom  
Residual deviance: 2515703  on 198  degrees of freedom  
AIC: 2517928
```

```
Number of Fisher Scoring iterations: 5
```

# Poisson Model: Deviance Residuals vs. Settlement Delay



# Legal Representation?



## Gamma Log GLM—Legal Representation?

```
Call: glm(formula = total ~ settle.delay + legrep,  
          family = Gamma(link = "log"), data = spinj)
```

```
Deviance Residuals:
```

	Min	1Q	Median	3Q	Max
	-2.8152	-0.8183	-0.3115	0.2864	2.6778

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	8.64459	0.13476	64.148	< 2e-16 ***
settle.delay	0.04112	0.00539	7.628	9.96e-13 ***
legrep1	0.70702	0.13989	5.054	9.85e-07 ***

```
---
```

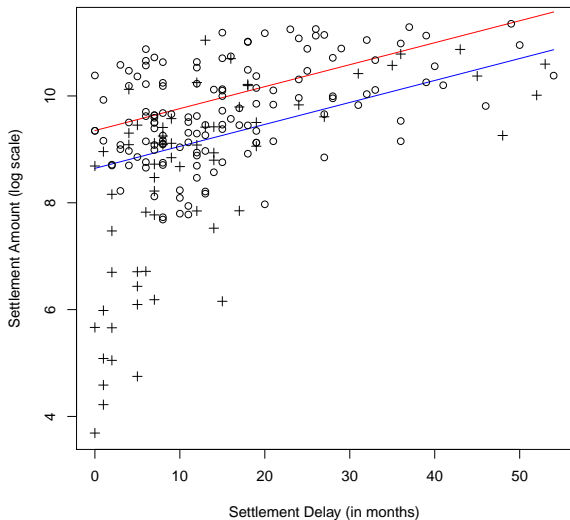
```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for Gamma family taken to be 0.8354751)
```

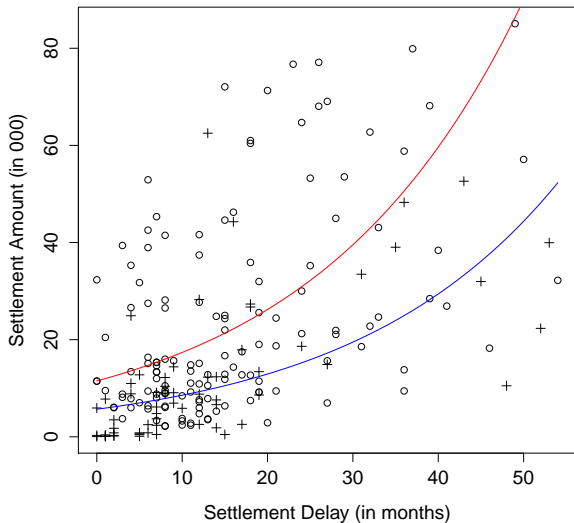
```
Null deviance: 252.05 on 199 degrees of freedom  
Residual deviance: 186.98 on 197 degrees of freedom  
AIC: 4300.9
```

```
Number of Fisher Scoring iterations: 8
```

# Legal Representation: Linear Predictor

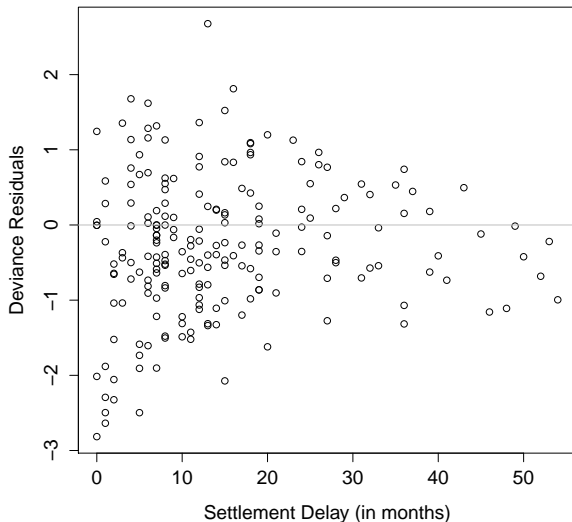


## Legal Representation: Fitted Values

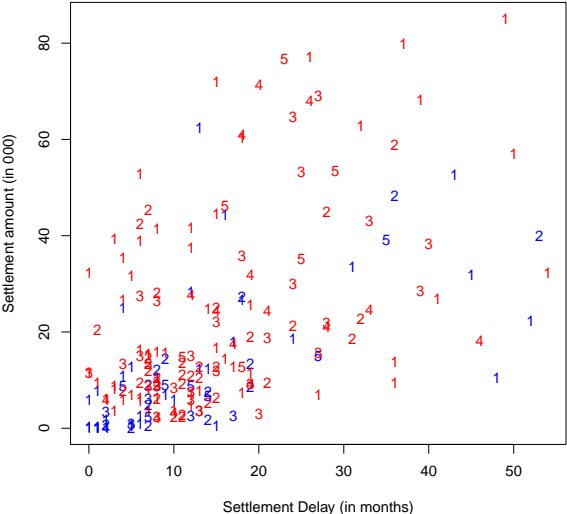




## Legal Representation: Deviance Residuals



# Number of Injured Persons



## Gamma Log GLM—Many Predictors

```
Call: glm(formula = total ~ settle.delay + legrep + inj.count,
          family = Gamma(link = "log"), data = spinj)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	8.722358	0.141721	61.546	< 2e-16	***
settle.delay	0.042138	0.005222	8.069	7.38e-14	***
legrep1	0.786161	0.139411	5.639	6.01e-08	***
inj.count2	-0.300230	0.160788	-1.867	0.0634	.
inj.count3	-0.416338	0.177247	-2.349	0.0198	*
inj.count4	-0.216891	0.244640	-0.887	0.3764	
inj.count5	0.005267	0.254395	0.021	0.9835	

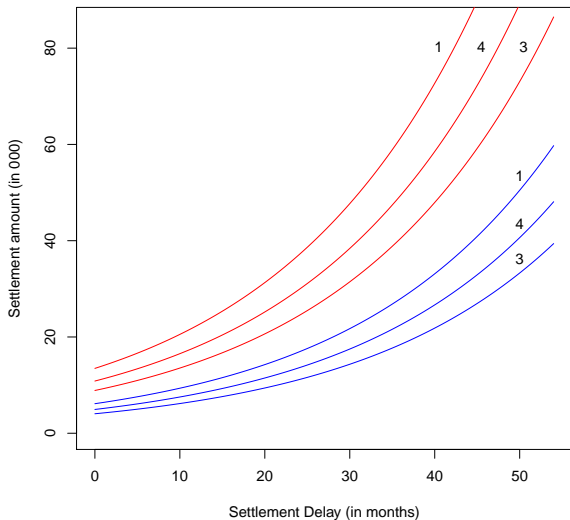
```
Null deviance: 252.05 on 199 degrees of freedom
Residual deviance: 181.44 on 193 degrees of freedom
AIC: 4302
```

```
Number of Fisher Scoring iterations: 9
```

## Predicted Values

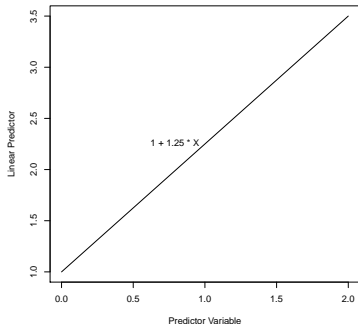
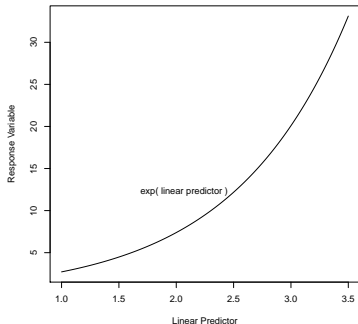
Settle Delay	Legal Rep?	Injured Count	Linear Predictor	Fitted Value
0	No	1	$8.7 + 0 \cdot 0.042 = 8.7$	$e^{8.7} = 6003$
0	Yes	1	$8.7 + 0 \cdot 0.042 + 0.79 = 9.5$	$e^{9.5} = 13360$
10	No	4	$8.7 + 10 \cdot 0.042 - 0.22 = 8.5$	$e^{8.9} = 7332$

## Many Predictors: Fitted Values



# Summary Key Concepts: Link Function

The link function is the bridge between the space of the linear predictor and the space of the response.






## Summary Key Concepts: Deviance

The deviance tells us how to measure the distance between an observation and its fitted value.

Distribution	Contribution to Squared Deviance
Normal	$(y_i - \mu_i)^2$
Poisson	$2\{y_i \log(y_i/\mu_i) - (y_i - \mu_i)\}$
Gamma	$2\{-\log(y_i/\mu_i) + (y_i - \mu_i)/\mu_i\}$
Inverse Gaussian	$(y_i - \mu_i)^2/(\mu_i^2 y_i)$

# 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.
-  Annette J. Dobson.  
*An introduction to Generalized Linear Models.*  
Chapman & Hall, London, 1990.
-  Edward W. Frees.  
*Regression Modeling with Actuarial and Financial Applications.*  
Cambridge University Press, 2010.



# References



James Hardin and Joseph Hilbe.  
*Generalized Linear Models and Extensions.*  
Stata Press, College Station, Texas, 2001.



Piet De Jong and Gillian Z. Heller.  
*Generalized Linear Models for Insurance Data.*  
Cambridge University Press, 2008.



W.N. Venables and B.D. Ripley.  
*Modern Applied Statistics with S.*  
Springer New York, 2002.