# 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 Linear & Generalized Linear Models

## Casualty Actuarial Society
## Ratemaking and Product Management Seminar
**March 15—17, 2010**
**Chicago, IL**

**Ashley Lambeth**

**Ernesto Schirmacher**

Safeco Insurance™
Member of Liberty Mutual Group

Liberty Mutual®

# Agenda

**Review of linear regression**

- Concrete example with summary statistics, plots, fitting models, understanding output

**Linear vs. generalized linear models**

- Link function, response distribution, variance as a function of the mean, deviance
- Model output, diagnostics, and interpretation

FICTITIOUS INSURANCE COMPANY

Project: New rating plan

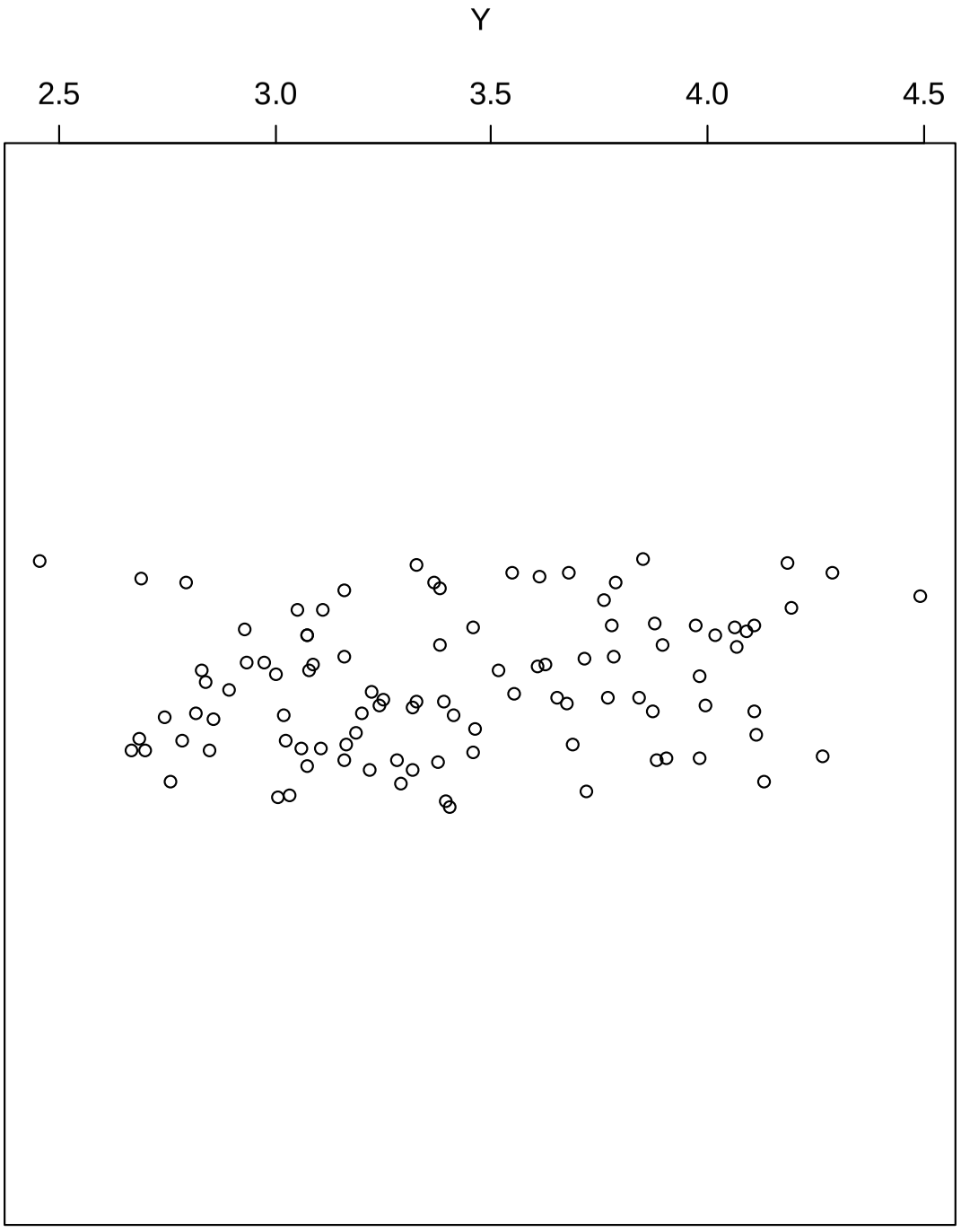
Variables available:

1. Garage location
2. Driver age
3. Gender (M, F, UNKNOWN)
4. Vehicle body(s)
5. Vehicle use
6. Vehicle
7. Loss

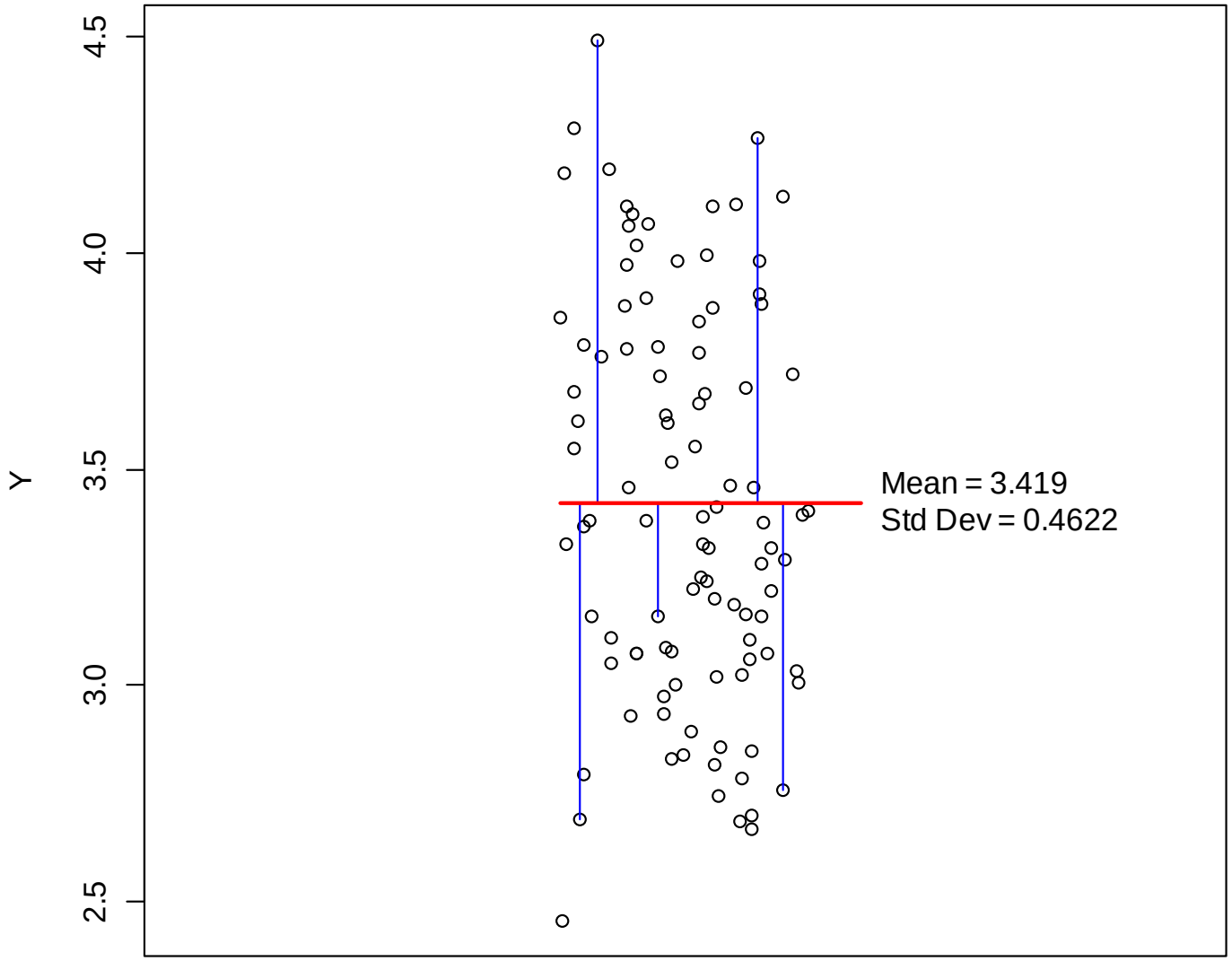
FICTITIOUS INSURANCE COMPANY

Project: New rating factors for auto (sample only).

| AREA | DRIVER'S AGE 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| A | 2 500 | 2 050 | 1 600 | 1800 / 1200 |
| B | 2 400 | 2 000 | 1 600 | 1 600 |
| C | 2 300 | 2 300 | 1 800 | 1 [illegible] |
| D | 2 [illegible] | 2 [illegible] | 2 100 | 2 300 |
| E | 3 450 | 1 700 | 2 [illegible] | 2 450 |
| F | 5 200 / 3000 | 2 [illegible] | 400 | |

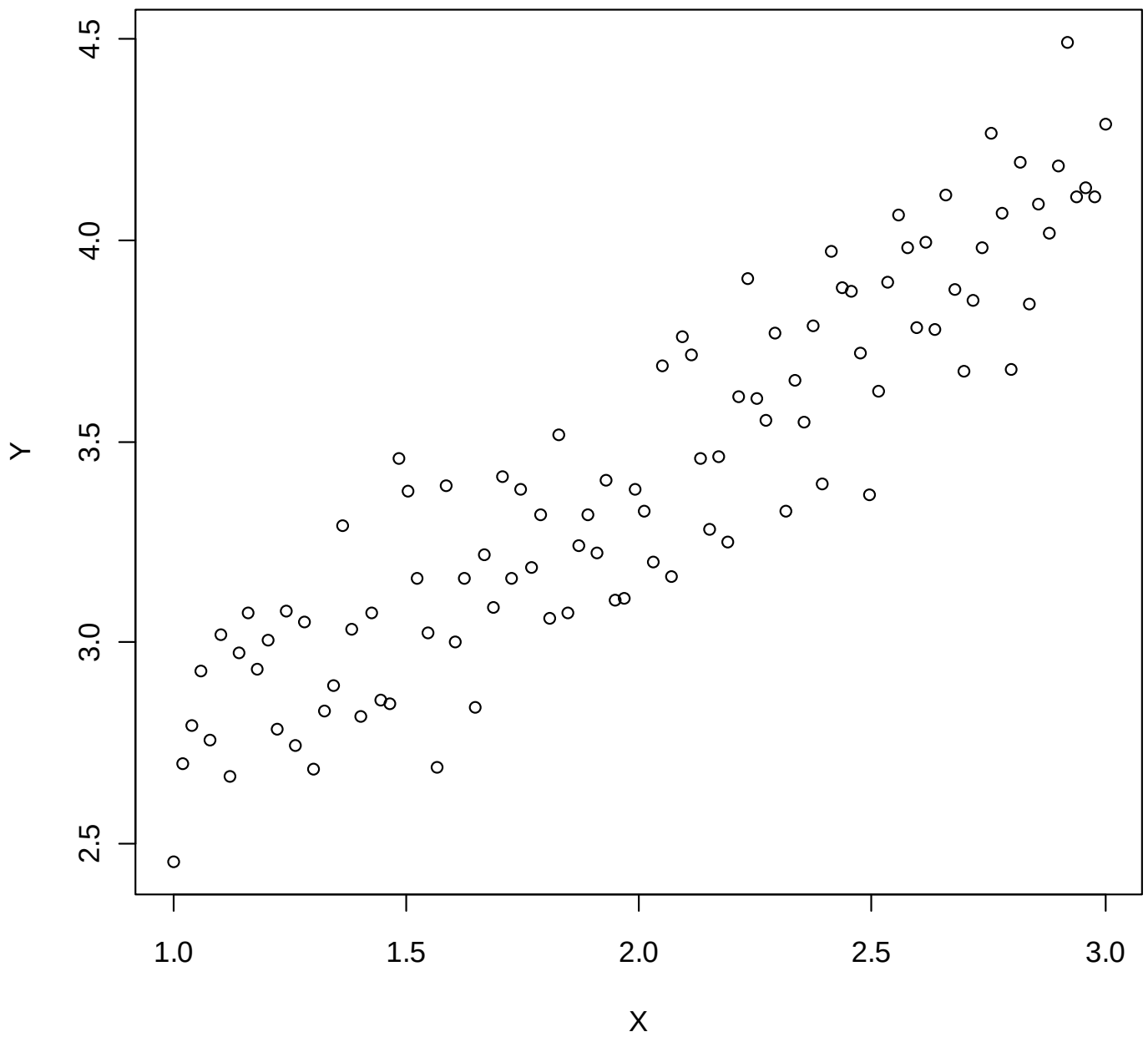

Y
2.5
3.0
3.5
4.0
4.5

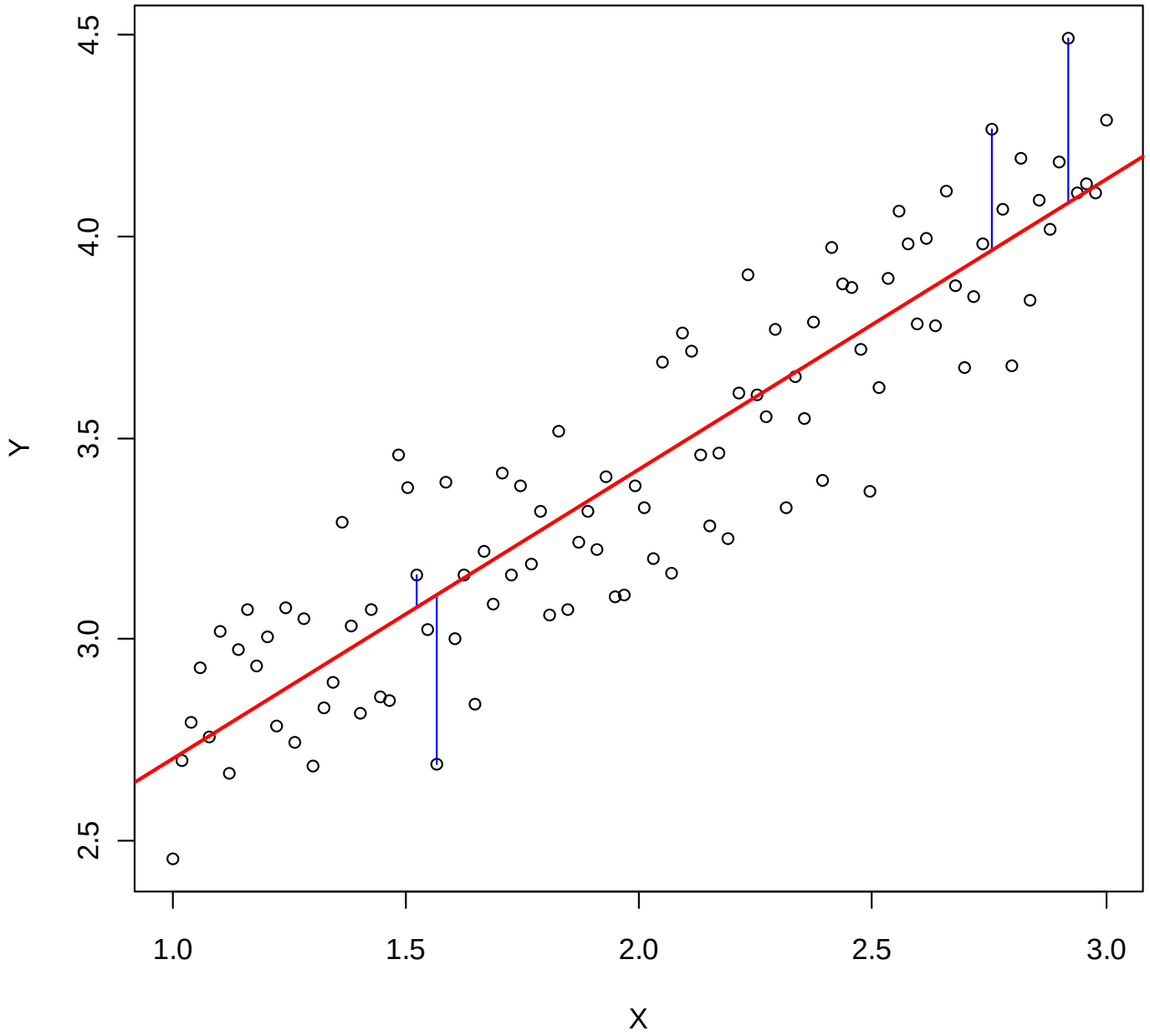
Y
4.5
4.0
3.5
3.0
2.5
Mean = 3.419


Y
4.5
4.0
3.5
3.0
2.5
Mean = 3.419
Std Dev = 0.4622

# Simple linear regression (intercept only)

```
Call:
lm(formula = y ~ 1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.96133 -0.34756 -0.04019  0.36406  1.06999

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.41912    0.04622   73.98   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4622 on 99 degrees of freedom
```


Y = 2 + 0.7 X + N(0, 0.2)
Y
X
2.5
3.0
3.5
4.0
4.5
1.0
1.5
2.0
2.5
3.0


Y
4.5
4.0
3.5
3.0
2.5
1.0
1.5
2.0
2.5
3.0
X

# Simple linear regression

```
Call:
lm(formula = y ~ x)

Residuals:
      Min        1Q    Median        3Q       Max
-0.418728 -0.128286 -0.003599  0.146999  0.409867

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.98280    0.06856   28.92   <2e-16 ***
x            0.71816    0.03291   21.82   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1919 on 98 degrees of freedom
Multiple R-squared: 0.8293, Adjusted R-squared: 0.8276
F-statistic: 476.2 on 1 and 98 DF,  p-value: < 2.2e-16
```

# Vehicle insurance claims

| Variable | Type | Comments |
|---|---|---|
| Loss | Continuous | range: $200 to $56,000 |
| Driver Age | Categorical | 6 levels |
| Gender | Categorical | 2 levels |
| Area | Categorical | 6 levels |
| Vehicle Body | Categorical | 13 levels |
| Vehicle Age | Categorical | 4 levels |
| Vehicle Value | Continuous | range: $0 to $140,000 |

Source: dataset `car.csv` from "Generalized Linear Models for Insurance Data."

loss
drv.age
gender
area
veh.body
veh.age
veh.value

# Summary statistics

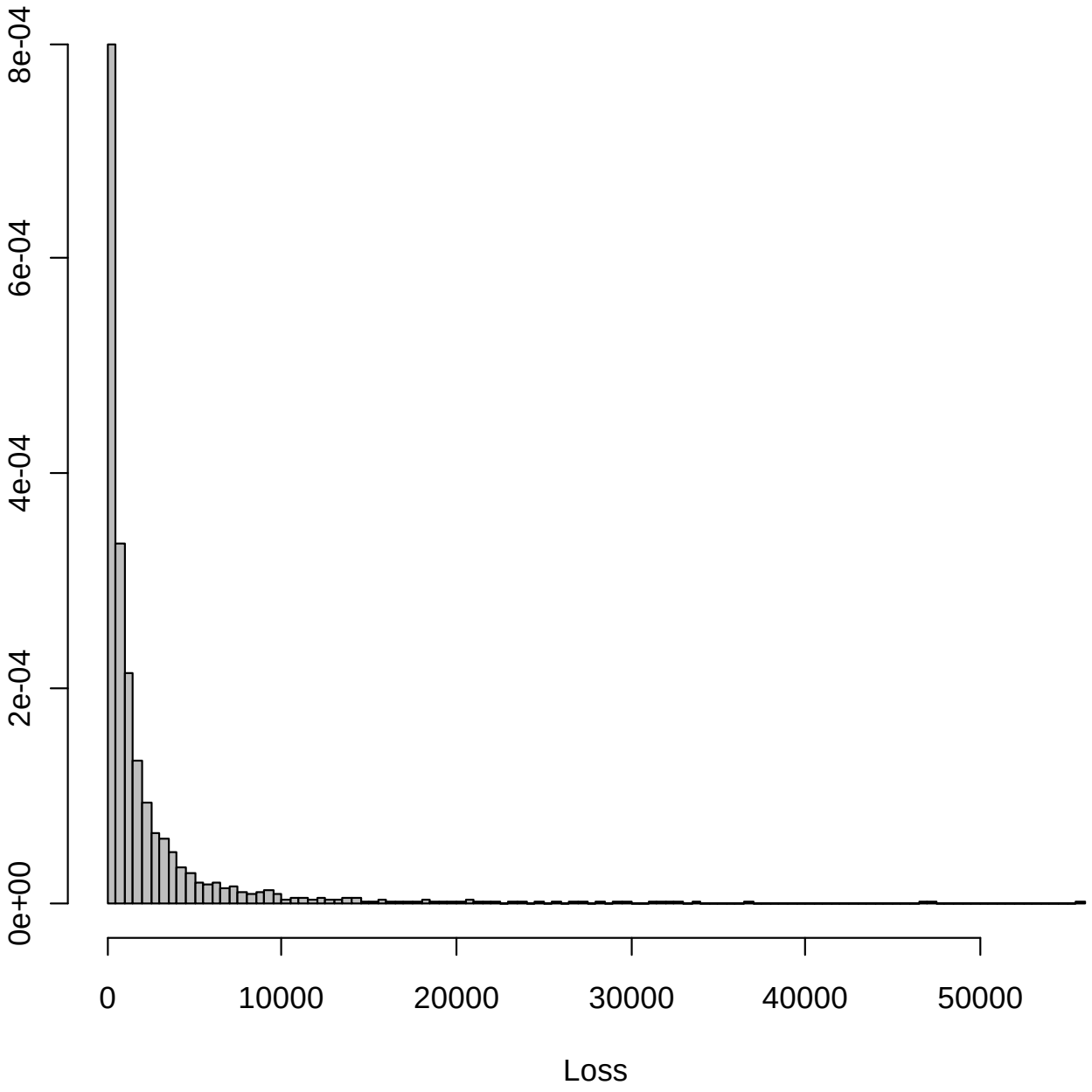
**Continuous Variables**

```
      loss                veh.value
 Min.   :   200      Min.   : 0.000
 1st Qu.:   353      1st Qu.: 1.100
 Median :   761      Median : 1.570
 Mean   : 2,014      Mean   : 1.859
 3rd Qu.: 2,091      3rd Qu.: 2.310
 Max.   :55,922      Max.   :13.900
```

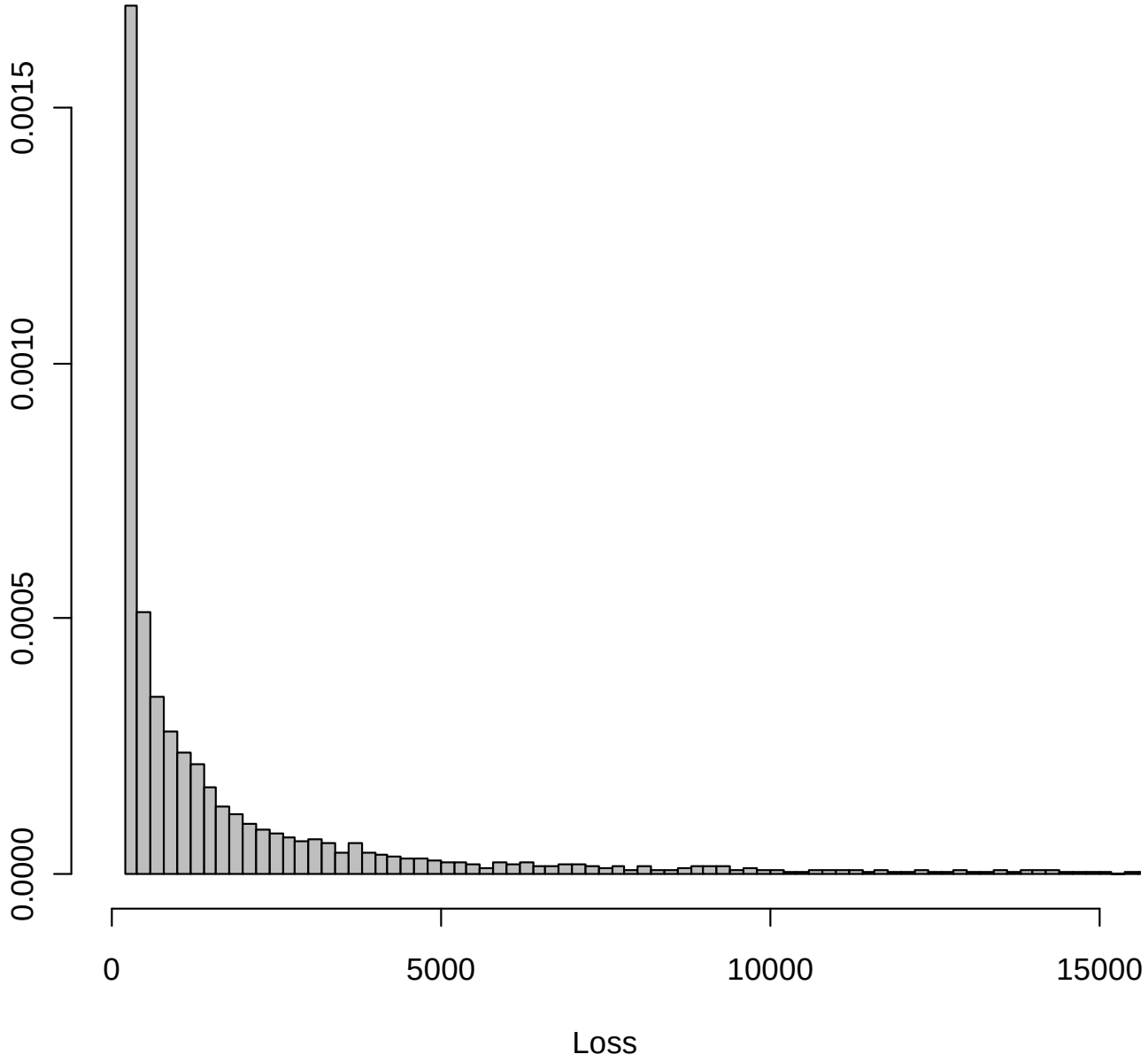
**Categorical Variables**

```
 drv.age gender    area       veh.body  veh.age
 1: 496    F:2648   A:1085   SEDAN  :1476   1: 825
 2: 932    M:1976   B: 965   HBACK  :1264   2:1259
 3:1113             C:1412   STNWG  :1173   3:1362
 4:1104             D: 496   UTE    : 260   4:1178
 5: 614             E: 386   HDTOP  : 130
 6: 365             F: 280   TRUCK  : 120
                             (Other): 201
```

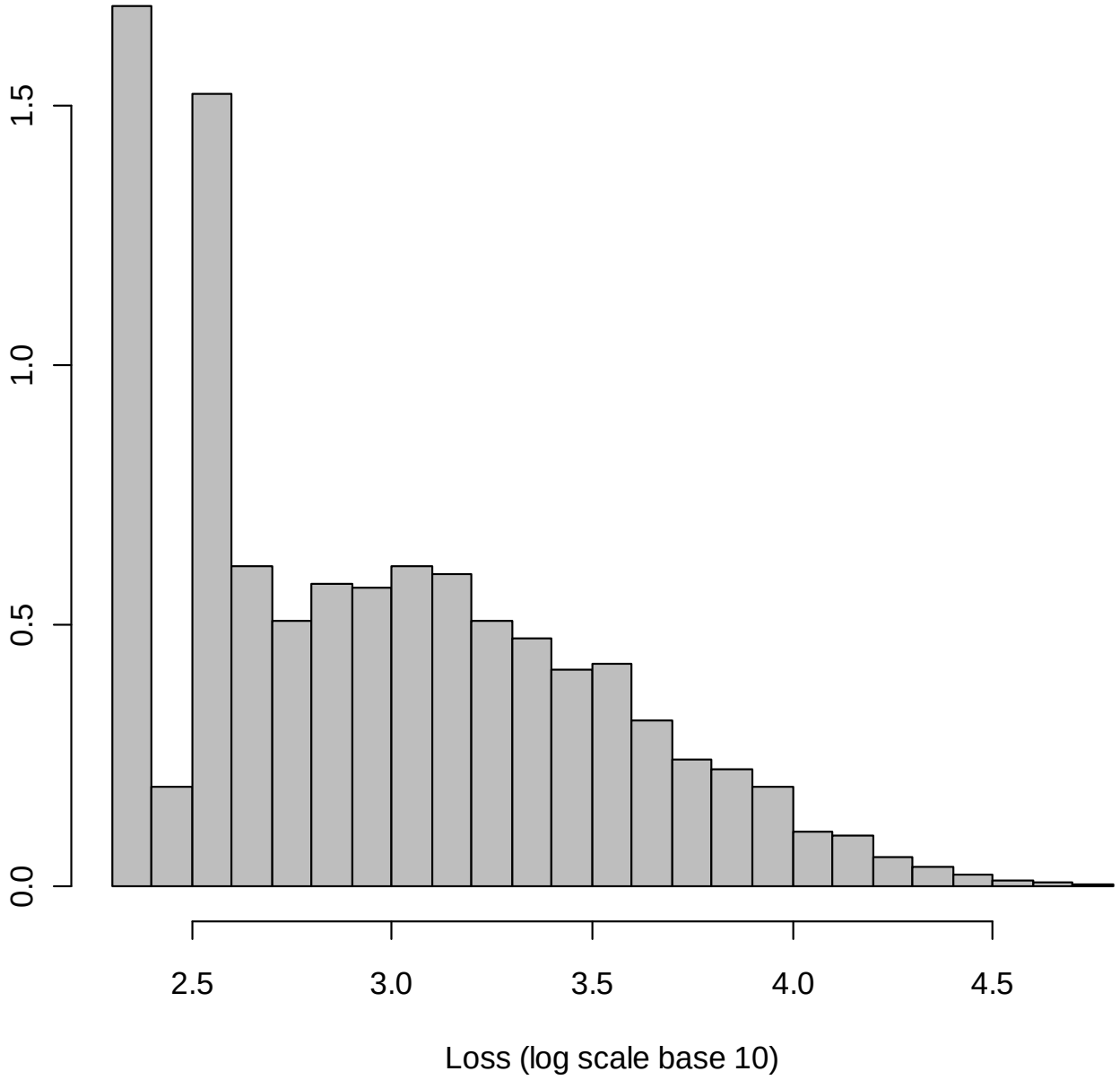
# Histogram of loss



# Histogram of loss (restrict $0 to $15K)



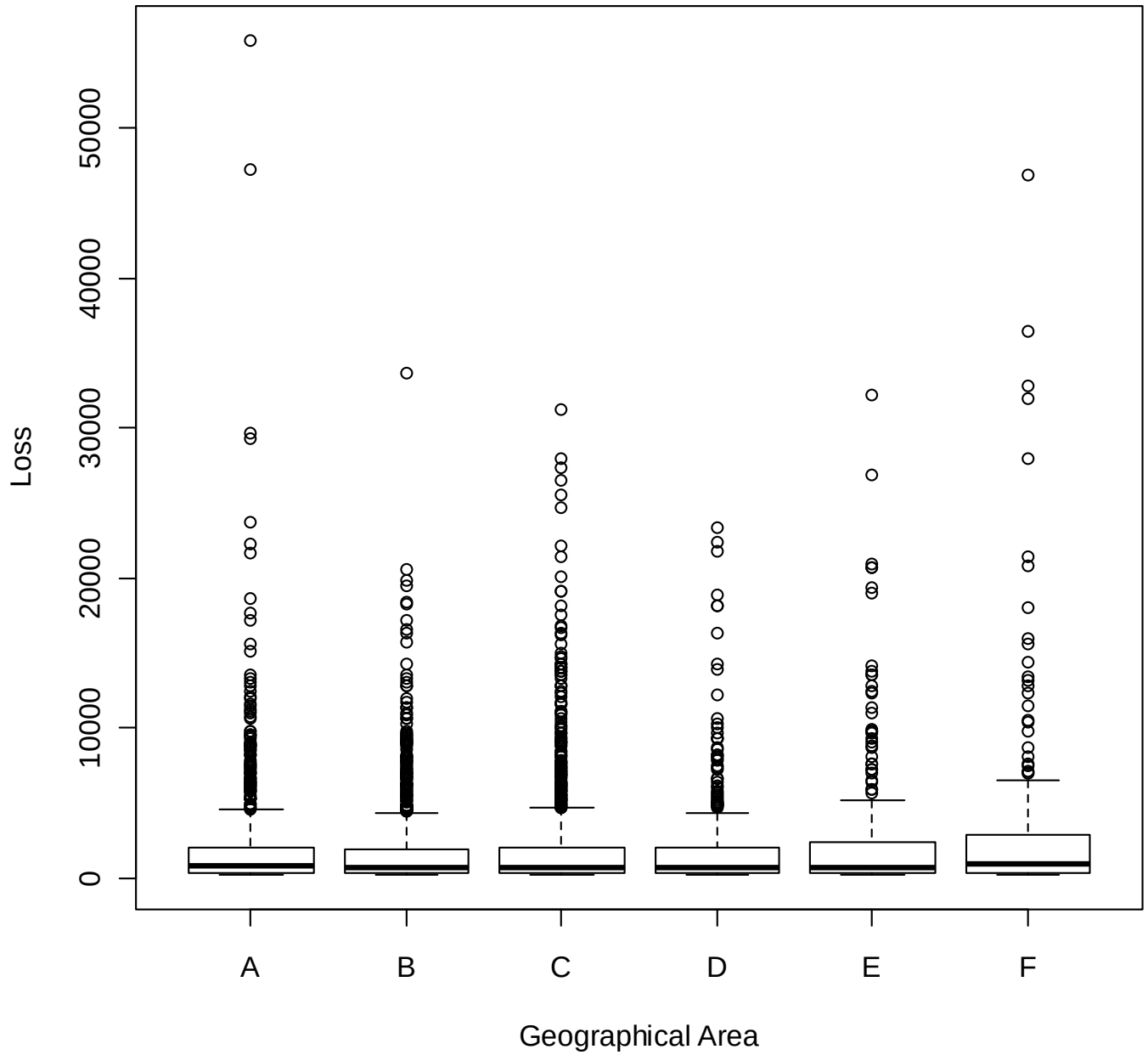
# Histogram of loss (log scale)



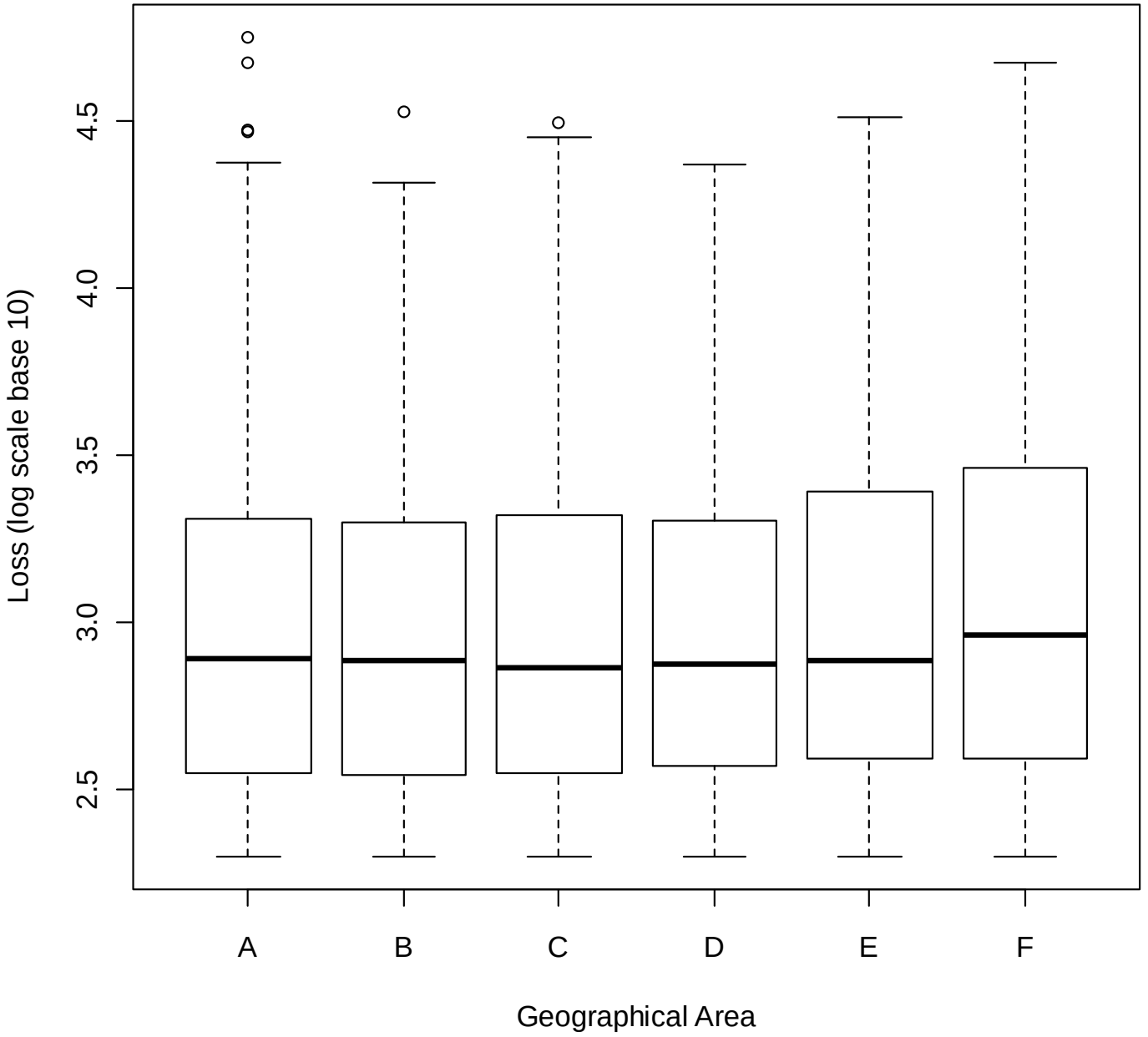
# Linear regression—null model

```
Call:
lm(formula = loss ~ 1, data = sv)

Residuals:
     Min       1Q   Median       3Q      Max 
-1814.40 -1660.63 -1252.84    77.02 53907.73 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2014.40      52.19    38.6   <2e-16 ***
---
Signif. codes:0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3549 on 4623 degrees of freedom
```

# Loss vs. geographical area



# log(loss) vs. geographical area



# Least squares fit: loss vs. area

```
Call:
lm(formula = loss ~ area, data = sv)

Residuals:
    Min      1Q  Median      3Q     Max
-2664.1 -1660.4 -1220.0   106.2 54012.7

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1909.46     107.54  17.755  < 2e-16 ***
area B         -49.05     156.75  -0.313    0.754
area C         120.08     143.01   0.840    0.401
area D         -72.65     192.01  -0.378    0.705
area E         341.38     209.94   1.626    0.104
area F         954.66     237.45   4.020  5.9e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3542 on 4618 degrees of freedom
Multiple R-squared: 0.004715,     Adjusted R-squared: 0.003638
F-statistic: 4.376 on 5 and 4618 DF,  p-value: 0.0005632
```

# Why so many parameters?

**formula = loss ~ area**

**Original Data**

| Row ID | loss | area |
|---|---|---|
| 4369 | 353 | A |
| 894 | 721 | B |
| 590 | 249 | B |
| 3877 | 389 | E |
| 3770 | 1,275 | C |
| 1695 | 1,301 | A |
| 3875 | 620 | D |
| 4411 | 2,073 | F |
| 599 | 2,114 | F |
| 623 | 359 | C |
| 1033 | 31,974 | F |

**Design Matrix**

| Y | $X_0$ | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ |
|---|---|---|---|---|---|---|---|
| 353 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 721 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| 249 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| 389 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| 1,275 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| 1,301 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 620 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| 2,073 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 2,114 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 359 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| 31,974 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |

$$Y = X_0 \beta_0 + X_2 \beta_1 + X_3 \beta_2 + X_4 \beta_3 + X_5 \beta_4 + X_6 \beta_5 + \varepsilon$$

**In Matrix Notation $Y = X \beta + \varepsilon$**

$$E[Y] = X \beta$$

# Value of the parameters

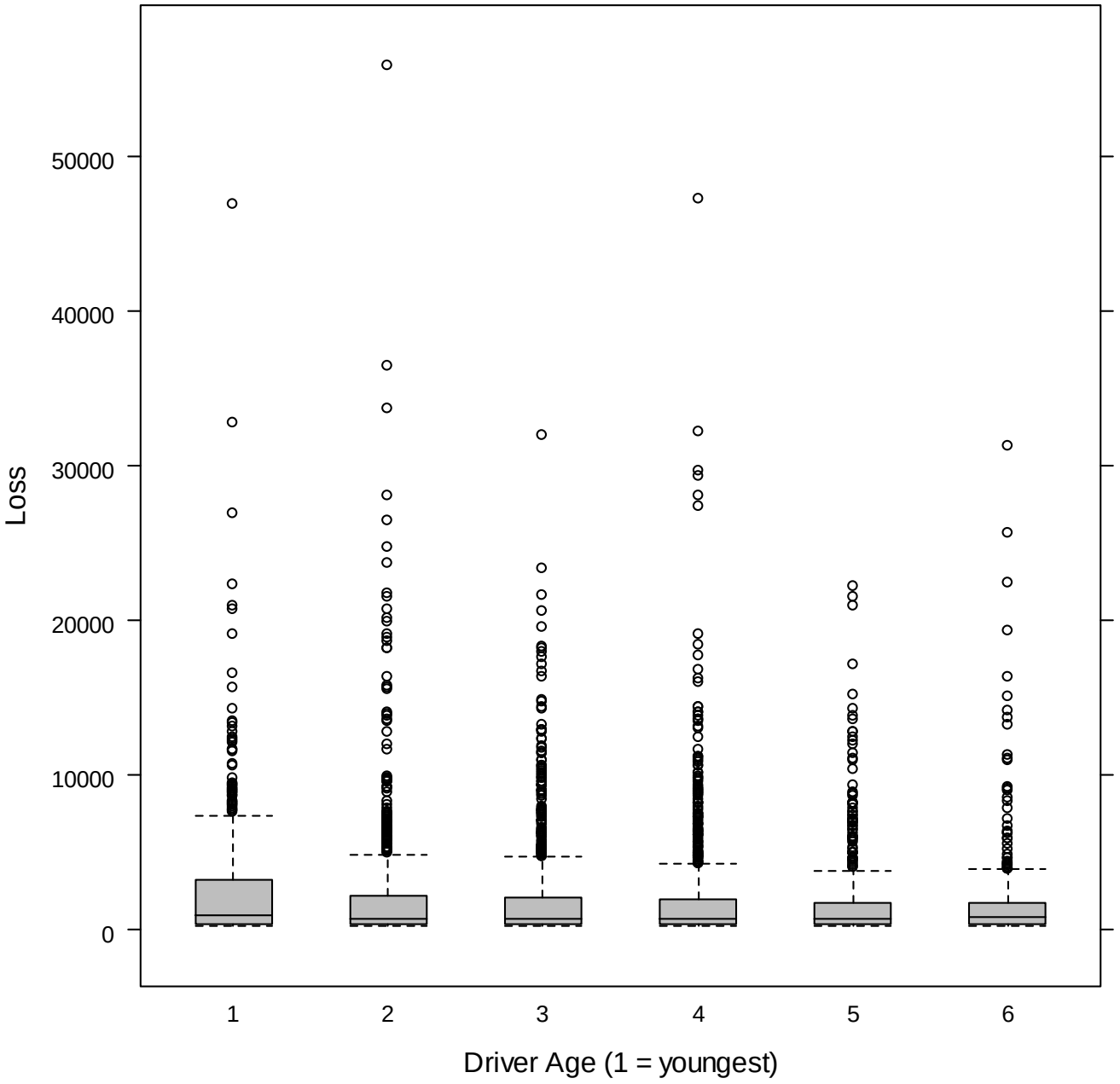
```
Call:
lm(formula = loss ~ area, data = sv)

Coefficients:
             Estimate
(Intercept)   1909.46
area B         -49.05
area C         120.08
area D         -72.65
area E         341.38
area F         954.66
```

| area | mean loss |
|---|---|
| A | 1,909.46 |
| B | 1,860.40 |
| C | 2,029.53 |
| D | 1,836.81 |
| E | 2,250.83 |
| F | 2,864.12 |

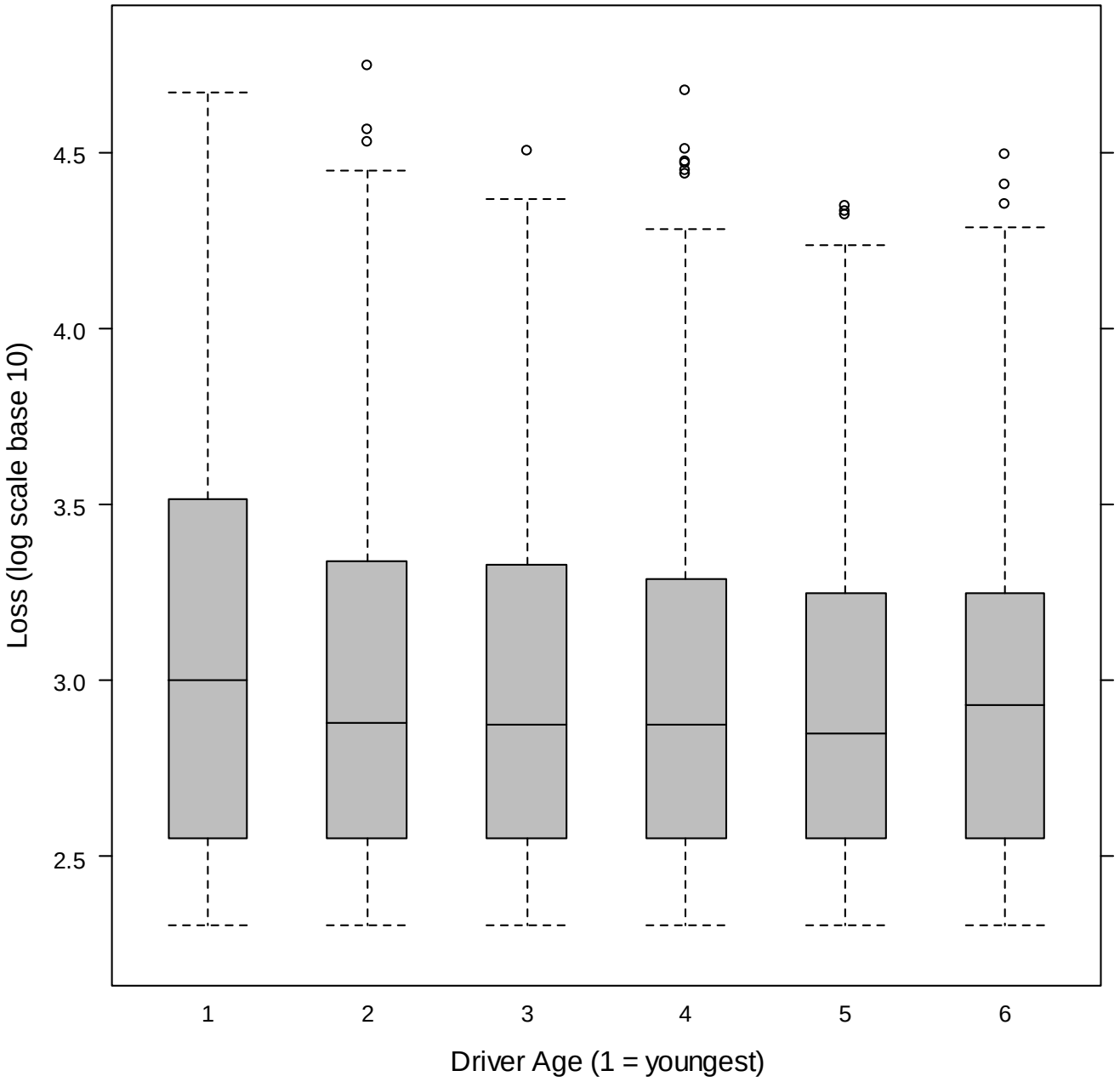
# Is the "loss ~ area" model any good?

```
Call:
lm(formula = loss ~ area, data = sv)

Residuals:
    Min      1Q  Median      3Q     Max
-2664.1 -1660.4 -1220.0   106.2 54012.7

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1909.46     107.54  17.755  < 2e-16 ***
area B         -49.05     156.75  -0.313    0.754
area C         120.08     143.01   0.840    0.401
area D         -72.65     192.01  -0.378    0.705
area E         341.38     209.94   1.626    0.104
area F         954.66     237.45   4.020  5.9e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3542 on 4618 degrees of freedom
Multiple R-squared: 0.004715,     Adjusted R-squared: 0.003638
F-statistic: 4.376 on 5 and 4618 DF,  p-value: 0.0005632
```

# Loss vs. driver age



# log(loss) vs. driver age



# Adding one more variable

```
Call:
lm(formula = loss ~ area + drv.age, data = sv)

Residuals:
    Min      1Q  Median      3Q     Max
-3088.0 -1593.8 -1169.2   111.8 53925.4

Coefficients:
             Estimate      |                Estimate
(Intercept)   2516.66 ***  |
area B         -42.77       |  drv.age 2     -519.95 **
area C         122.20       |  drv.age 3     -722.84 ***
area D         -49.79       |  drv.age 4     -667.12 ***
area E         338.64       |  drv.age 5     -875.63 ***
area F         925.07 ***   |  drv.age 6     -699.93 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3537 on 4613 degrees of freedom
Multiple R-squared: 0.009106,    Adjusted R-squared: 0.006958
F-statistic: 4.239 on 10 and 4613 DF,  p-value: 6.813e-06
```

# Do estimates match the data?

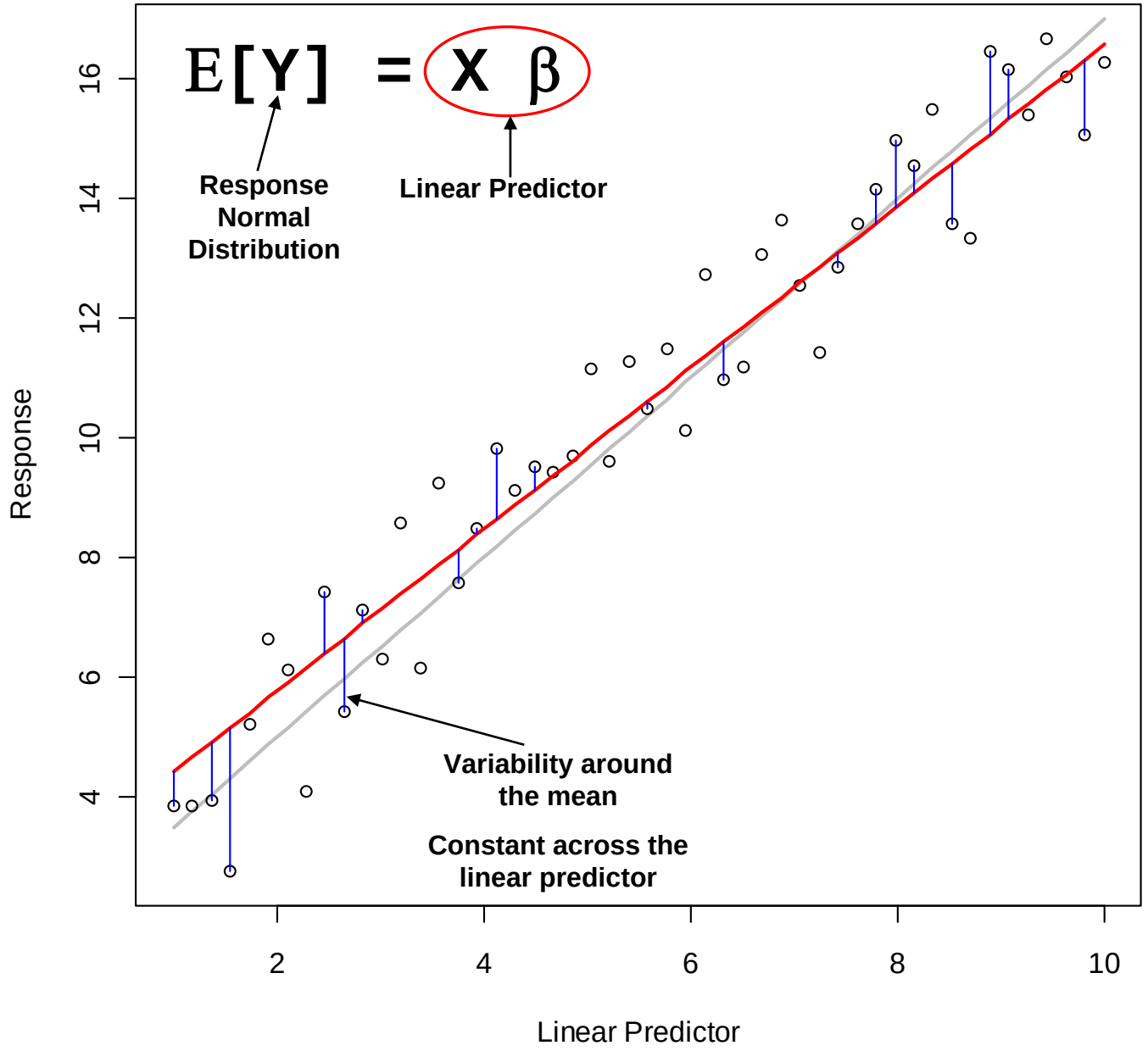
**Average loss by driver's age and area**

| | driver's age | | | | | |
|---|---|---|---|---|---|---|
| area | 1 | 2 | 3 | 4 | 5 | 6 |
| A | 2,536 | 2,052 | 1,583 | 2,031 | 1,794 | 1,490 |
| B | 2,181 | 2,066 | 2,081 | 1,652 | 1,580 | 1,367 |
| C | 2,363 | 2,296 | 1,808 | 1,858 | 1,821 | 2,532 |
| D | 2,064 | 2,258 | 2,122 | 1,678 | 1,090 | 1,518 |
| E | 3,415 | 1,766 | 2,012 | 2,708 | 1,719 | 2,351 |
| F | 5,313 | 2,211 | 2,426 | 2,789 | 3,450 | 2,497 |

**Model fitted values**

| | driver' age | | | | | |
|---|---|---|---|---|---|---|
| area | 1 | 2 | 3 | 4 | 5 | 6 |
| A | 2,517 | 1,997 | 1,794 | 1,850 | 1,641 | 1,817 |
| B | 2,474 | 1,954 | 1,751 | 1,807 | 1,598 | 1,774 |
| C | 2,639 | 2,119 | 1,916 | 1,972 | 1,763 | 1,939 |
| D | 2,467 | 1,947 | 1,744 | 1,800 | 1,591 | 1,767 |
| E | 2,855 | 2,335 | 2,132 | 2,188 | 1,980 | 2,155 |
| F | 3,442 | 2,922 | 2,719 | 2,775 | 2,566 | 2,742 |

# Main Ideas in Linear Modeling



# Linear vs. Generalized Linear Model

| Assumption | Linear Regression Model | Generalized Linear Model |
|---|---|---|
| Relationship between X and Y | Y is a linear combination of X | Y is a function of a linear combination of X |
| Distribution of Y | Normal | Any distribution from the Exponential family |
| Variance of Y | Constant | Function of the mean |

# Flexibility of Relationship between X & Y

- Recall that the Multiple Linear Regression Model can be written as:

$$E[\, Y_i \,] = X_i\beta$$

(Or… $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_n X_{in}$ )

- Generalized Linear Models assume a more general relationship between X and Y:

$$E[\, Y_i \,] = h(X_i\beta)$$

# Flexibility of Relationship between X & Y

- Generalized Linear Models assume a more general relationship between X and Y:

$$g(Y_i) = X_i\beta$$

Link Function

Examples:

- $Y_i = X_i\beta$ (Identity Link)
- $\ln(Y_i) = X_i\beta$ (Log Link)
- $\ln(Y_i/(1-Y_i)) = X_i\beta$ (Logit Link)
- $1/Y_i = X_i\beta$ (Reciprocal Link)

# Identity Link vs. Log Link

Identity Link:

- $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_n X_{in}$

Log Link:

- $\ln(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_n X_{in}$
- $Y_i = \exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_n X_{in})$
- $Y_i = \exp(\beta_0) \cdot \exp(\beta_1 X_{i1}) \cdot \exp(\beta_2 X_{i2}) \cdot \ldots \cdot \exp(\beta_n X_{in})$

Identity Link produces additive factors; Log Link produces multiplicative factors

# Why choose log link?

Convenience: match structure of rating plan/scorecard

Intuition: Do rating variables have additive or multiplicative effects on severity?

Evaluation: Test appropriateness of link function

# Model output: normal error / log link

| Parameter Number | Name | Value | Standard Error | Standard Error (%) | Alias Indicator (%) | Weight | Weight (%) | Exp(Value) |
|---|---|---|---|---|---|---|---|---|
| 1 | Mean | 7.4885 | 0.06907 | 0.9 | | 4,937 | 100.0 | 1,787.4562 |
| 2 | area (A) | -0.0841 | 0.07396 | 88.0 | | 1,181 | 23.9 | 0.9194 |
| 3 | area (B) | -0.0840 | 0.07759 | 92.3 | | 1,021 | 20.7 | 0.9194 |
| - | area (C) | | | | | 1,493 | 30.2 | |
| 4 | area (D) | -0.0807 | 0.09882 | 122.5 | | 524 | 10.6 | 0.9225 |
| 5 | area (E) | 0.1103 | 0.09273 | 84.1 | | 413 | 8.4 | 1.1166 |
| 6 | area (F) | 0.3267 | 0.08824 | 27.0 | | 305 | 6.2 | 1.3864 |
| 7 | agecat (1) | 0.3556 | 0.08171 | 23.0 | | 525 | 10.6 | 1.4270 |
| 8 | agecat (2) | 0.0834 | 0.07938 | 95.1 | | 1,000 | 20.3 | 1.0870 |
| - | agecat (3) | | | | | 1,189 | 24.1 | |
| 9 | agecat (4) | 0.0336 | 0.07913 | 235.4 | | 1,185 | 24.0 | 1.0342 |
| 10 | agecat (5) | -0.0613 | 0.09974 | 162.6 | | 648 | 13.1 | 0.9405 |
| 11 | agecat (6) | 0.0259 | 0.11462 | 442.5 | | 390 | 7.9 | 1.0262 |

$$E[Y] = \exp(\beta_1) \cdot \exp(\beta_2 X_2) \cdot \exp(\beta_3 X_3) \cdot \ldots \cdot \exp(\beta_{11} X_{11})$$

# Model output: normal error / log link

| Parameter Number | Name | Value | Standard Error | Standard Error (%) | Alias Indicator (%) | Weight | Weight (%) | Exp(Value) |
|---|---|---|---|---|---|---|---|---|
| 1 | Mean | 7.4885 | 0.06907 | 0.9 | | 4,937 | 100.0 | 1,787.4562 |
| 2 | area (A) | -0.0841 | 0.07396 | 88.0 | | 1,181 | 23.9 | 0.9194 |
| 3 | area (B) | -0.0840 | 0.07759 | 92.3 | | 1,021 | 20.7 | 0.9194 |
| - | area (C) | | | | | 1,493 | 30.2 | |
| 4 | area (D) | -0.0807 | 0.09882 | 122.5 | | 524 | 10.6 | 0.9225 |
| 5 | area (E) | 0.1103 | 0.09273 | 84.1 | | 413 | 8.4 | 1.1166 |
| 6 | area (F) | 0.3267 | 0.08824 | 27.0 | | 305 | 6.2 | 1.3864 |
| 7 | agecat (1) | 0.3556 | 0.08171 | 23.0 | | 525 | 10.6 | 1.4270 |
| 8 | agecat (2) | 0.0834 | 0.07938 | 95.1 | | 1,000 | 20.3 | 1.0870 |
| - | agecat (3) | | | | | 1,189 | 24.1 | |
| 9 | agecat (4) | 0.0336 | 0.07913 | 235.4 | | 1,185 | 24.0 | 1.0342 |
| 10 | agecat (5) | -0.0613 | 0.09974 | 162.6 | | 648 | 13.1 | 0.9405 |
| 11 | agecat (6) | 0.0259 | 0.11462 | 442.5 | | 390 | 7.9 | 1.0262 |

For Area=C, Agecat=3

$E[Y] = \exp(\beta_1) \cdot \exp(\beta_2 \cdot 0) \cdot \exp(\beta_3 \cdot 0) \cdot \ldots \cdot \exp(\beta_{11} \cdot 0)$

$= \exp(7.4885) = 1{,}787.46$

# Model output: normal error / log link

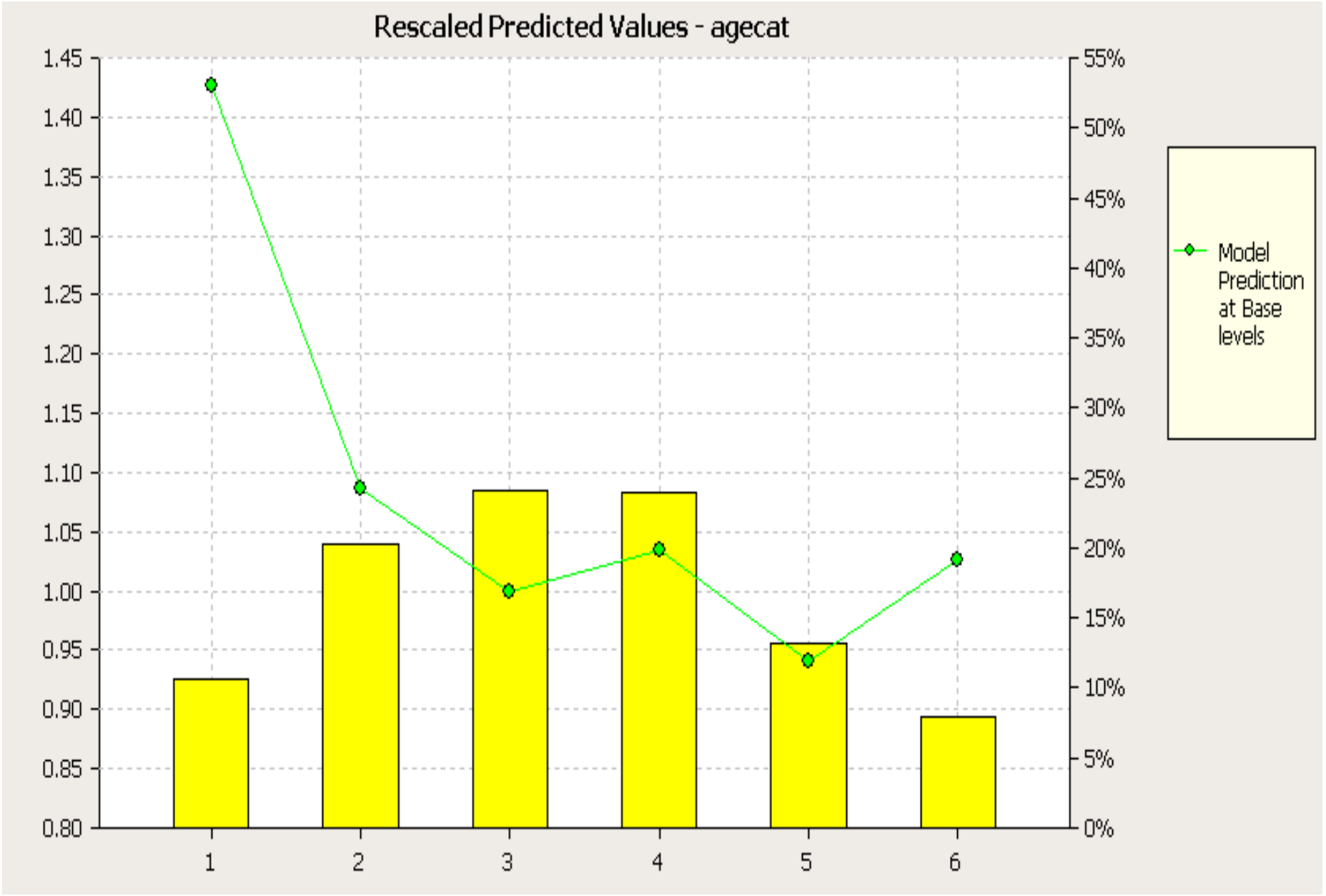
| Parameter Number | Name | Value | Standard Error | Standard Error (%) | Alias Indicator (%) | Weight | Weight (%) | Exp(Value) |
|---|---|---|---|---|---|---|---|---|
| 1 | Mean | 7.4885 | 0.06907 | 0.9 | | 4,937 | 100.0 | 1,787.4562 |
| 2 | area (A) | -0.0841 | 0.07396 | 88.0 | | 1,181 | 23.9 | 0.9194 |
| 3 | area (B) | -0.0840 | 0.07759 | 92.3 | | 1,021 | 20.7 | 0.9194 |
| - | area (C) | | | | | 1,493 | 30.2 | |
| 4 | area (D) | -0.0807 | 0.09882 | 122.5 | | 524 | 10.6 | 0.9225 |
| 5 | area (E) | 0.1103 | 0.09273 | 84.1 | | 413 | 8.4 | 1.1166 |
| 6 | area (F) | 0.3267 | 0.08824 | 27.0 | | 305 | 6.2 | 1.3864 |
| 7 | agecat (1) | 0.3556 | 0.08171 | 23.0 | | 525 | 10.6 | 1.4270 |
| 8 | agecat (2) | 0.0834 | 0.07938 | 95.1 | | 1,000 | 20.3 | 1.0870 |
| - | agecat (3) | | | | | 1,189 | 24.1 | |
| 9 | agecat (4) | 0.0336 | 0.07913 | 235.4 | | 1,185 | 24.0 | 1.0342 |
| 10 | agecat (5) | -0.0613 | 0.09974 | 162.6 | | 648 | 13.1 | 0.9405 |
| 11 | agecat (6) | 0.0259 | 0.11462 | 442.5 | | 390 | 7.9 | 1.0262 |

For Area=A, Agecat=6

$E[Y] = \exp(\beta_1) \cdot \exp(\beta_2 \cdot 1) \cdot \exp(\beta_3 \cdot 0) \cdot \ldots \cdot \exp(\beta_{11} \cdot 1)$

$= \exp(7.4885) \cdot \exp(-0.0841) \cdot \exp(0.0259)$

$= 1{,}787.46 \cdot 0.9194 \cdot 1.0262 = 1{,}686.31$


Rescaled Predicted Values - agecat
1.45
1.40
1.35
1.30
1.25
1.20
1.15
1.10
1.05
1.00
0.95
0.90
0.85
0.80
55%
50%
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
1
2
3
4
5
6
Model Prediction at Base levels

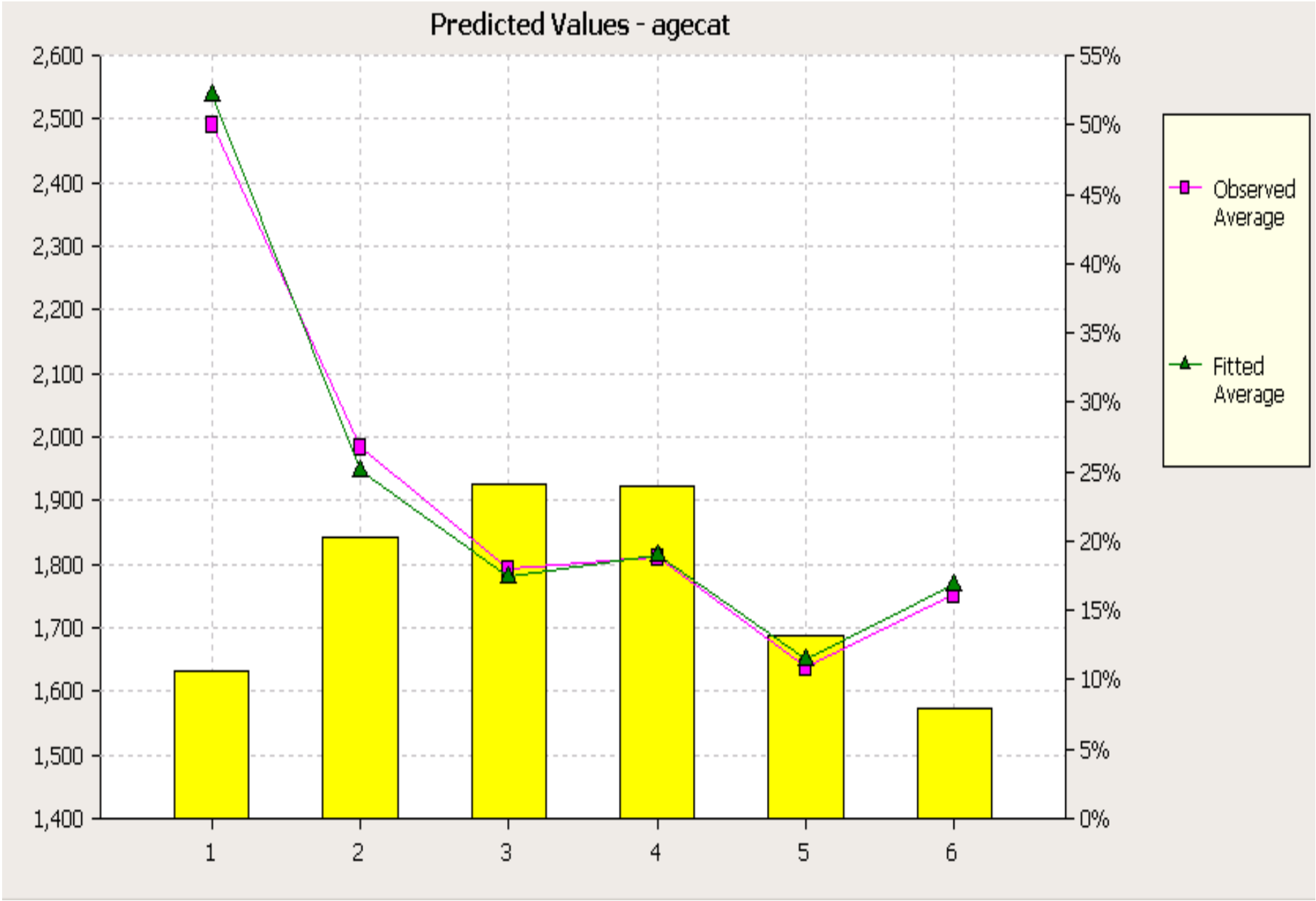
# Do estimates match the data?

**Average loss by driver's age and area**

| | driver's age | | | | | |
|---|---|---|---|---|---|---|
| area | 1 | 2 | 3 | 4 | 5 | 6 |
| A | 2,536 | 2,052 | 1,583 | 2,031 | 1,794 | 1,490 |
| B | 2,181 | 2,066 | 2,081 | 1,652 | 1,580 | 1,367 |
| C | 2,363 | 2,296 | 1,808 | 1,858 | 1,821 | 2,532 |
| D | 2,064 | 2,258 | 2,122 | 1,678 | 1,090 | 1,518 |
| E | 3,415 | 1,766 | 2,012 | 2,708 | 1,719 | 2,351 |
| F | 5,313 | 2,211 | 2,426 | 2,789 | 3,450 | 2,497 |

**Model fitted values**

| | driver' age | | | | | |
|---|---|---|---|---|---|---|
| area | 1 | 2 | 3 | 4 | 5 | 6 |
| A | 2,517 | 1,997 | 1,794 | 1,850 | 1,641 | 1,817 |
| B | 2,474 | 1,954 | 1,751 | 1,807 | 1,598 | 1,774 |
| C | 2,639 | 2,119 | 1,916 | 1,972 | 1,763 | 1,939 |
| D | 2,467 | 1,947 | 1,744 | 1,800 | 1,591 | 1,767 |
| E | 2,855 | 2,335 | 2,132 | 2,188 | 1,980 | 2,155 |
| F | 3,442 | 2,922 | 2,719 | 2,775 | 2,566 | 2,742 |


Predicted Values - agecat
2,600
2,500
2,400
2,300
2,200
2,100
2,000
1,900
1,800
1,700
1,600
1,500
1,400
55%
50%
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
1
2
3
4
5
6
Observed Average
Fitted Average

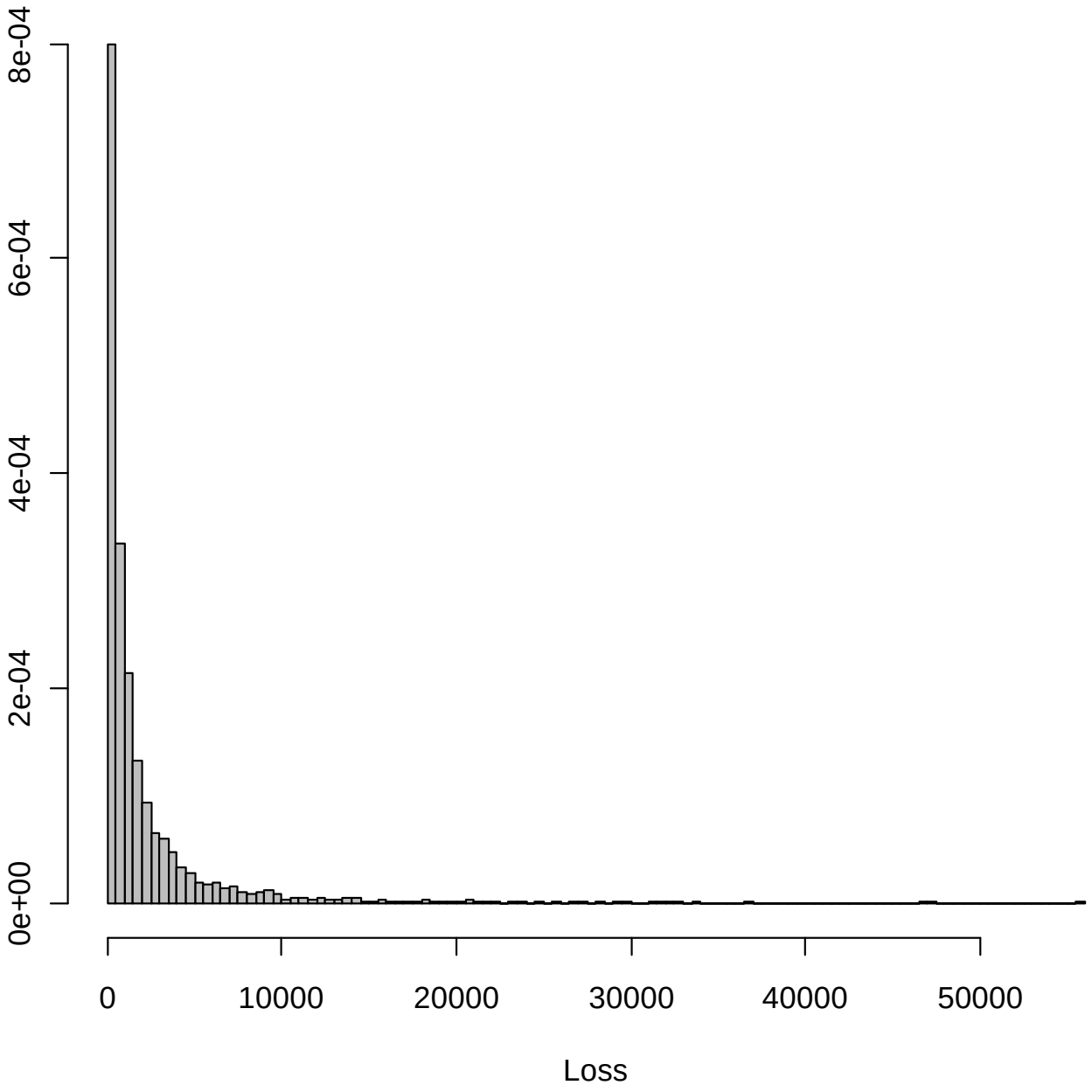
# Linear vs. Generalized Linear Model

| Assumption | Linear Regression Model | Generalized Linear Model |
|---|---|---|
| Relationship between X and Y | Y is a linear combination of X | Y is a function of a linear combination of X |
| Distribution of Y | Normal | Any distribution from the Exponential family |
| Variance of Y | Constant | Function of the mean |

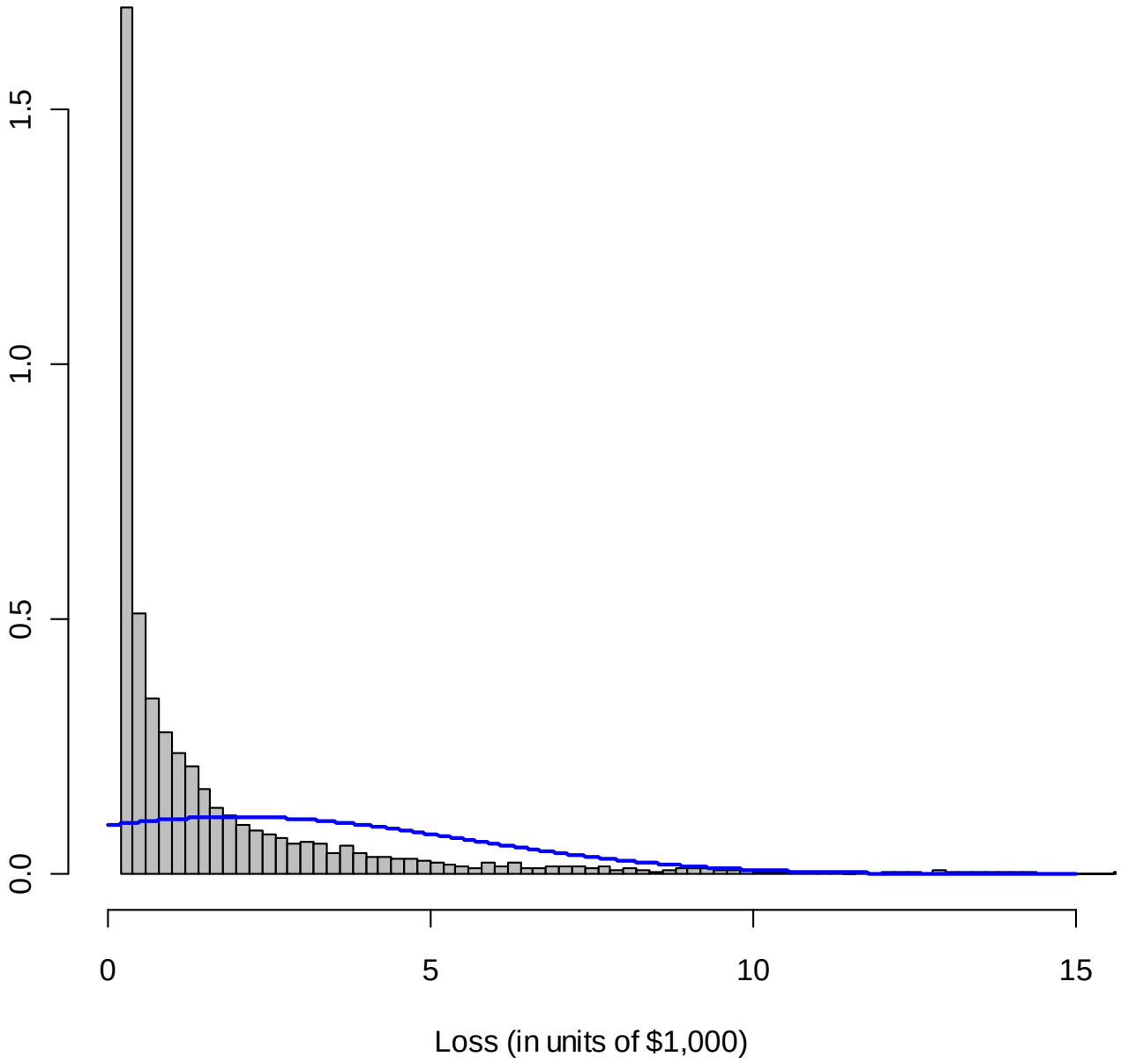
# Flexibility of Distribution of Y

- Least-squares estimation implicitly assumes observations come from normal distribution

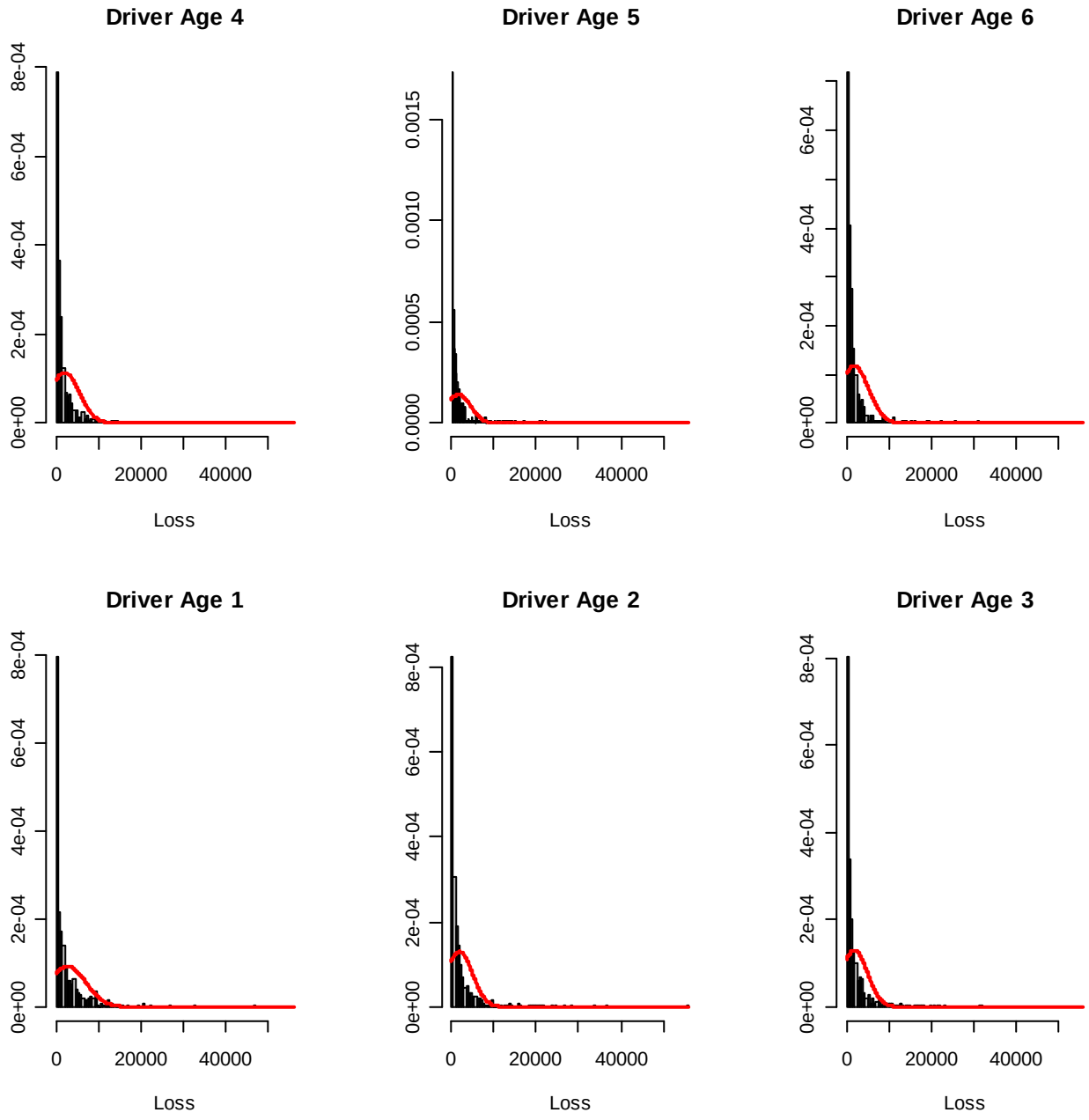
# Histogram of loss



# Histogram of Loss with normal distribution



# Histograms of Loss with normal distribution



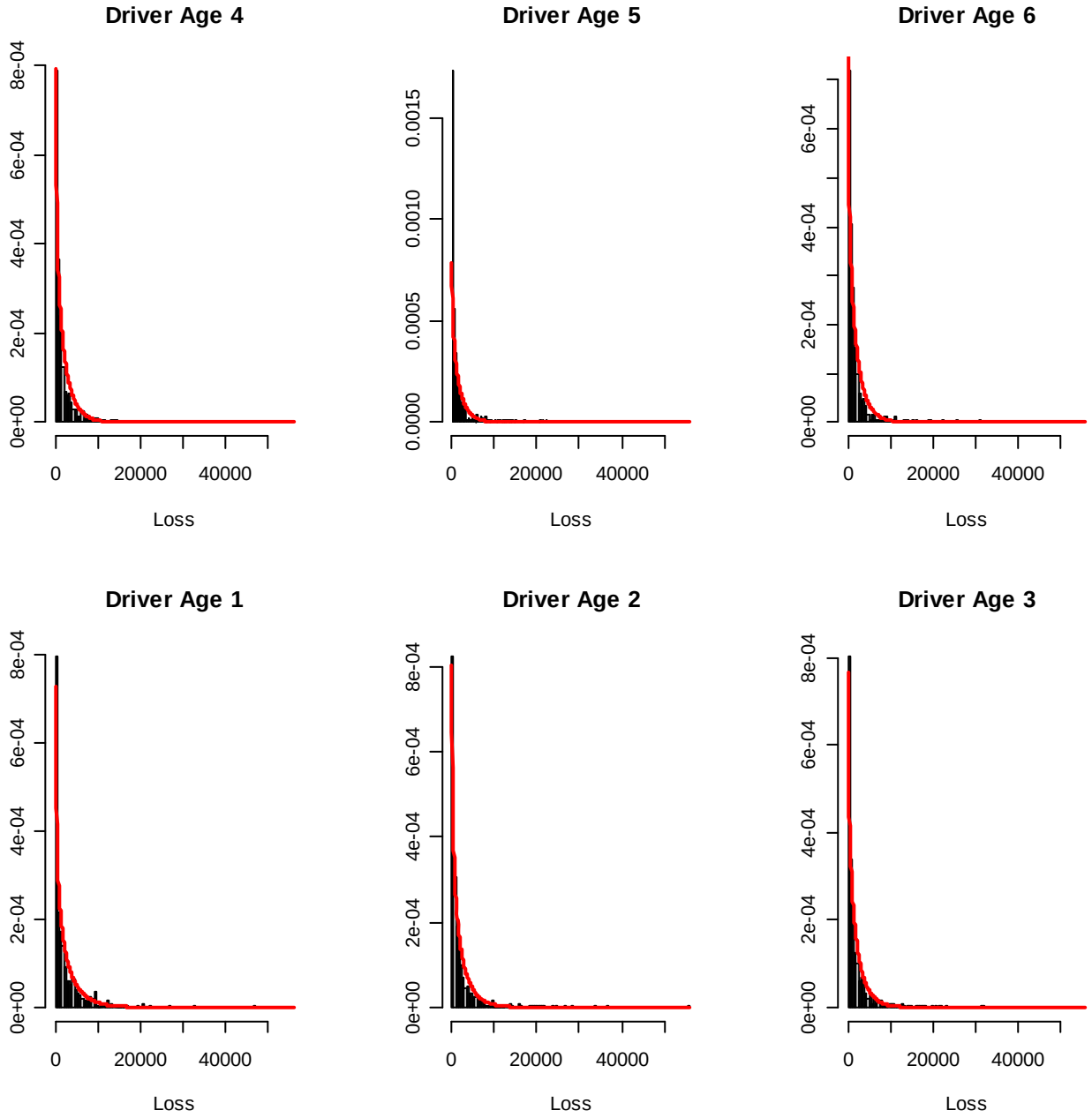
# Flexibility of Distribution of Y

- Least-squares estimation implicitly assumes observations come from normal distribution
- Problems with normal distribution assumption
  - Severity distributions usually skewed to right
  - Higher mean of Y associated with higher variance
  - Values of response may be restricted to positive

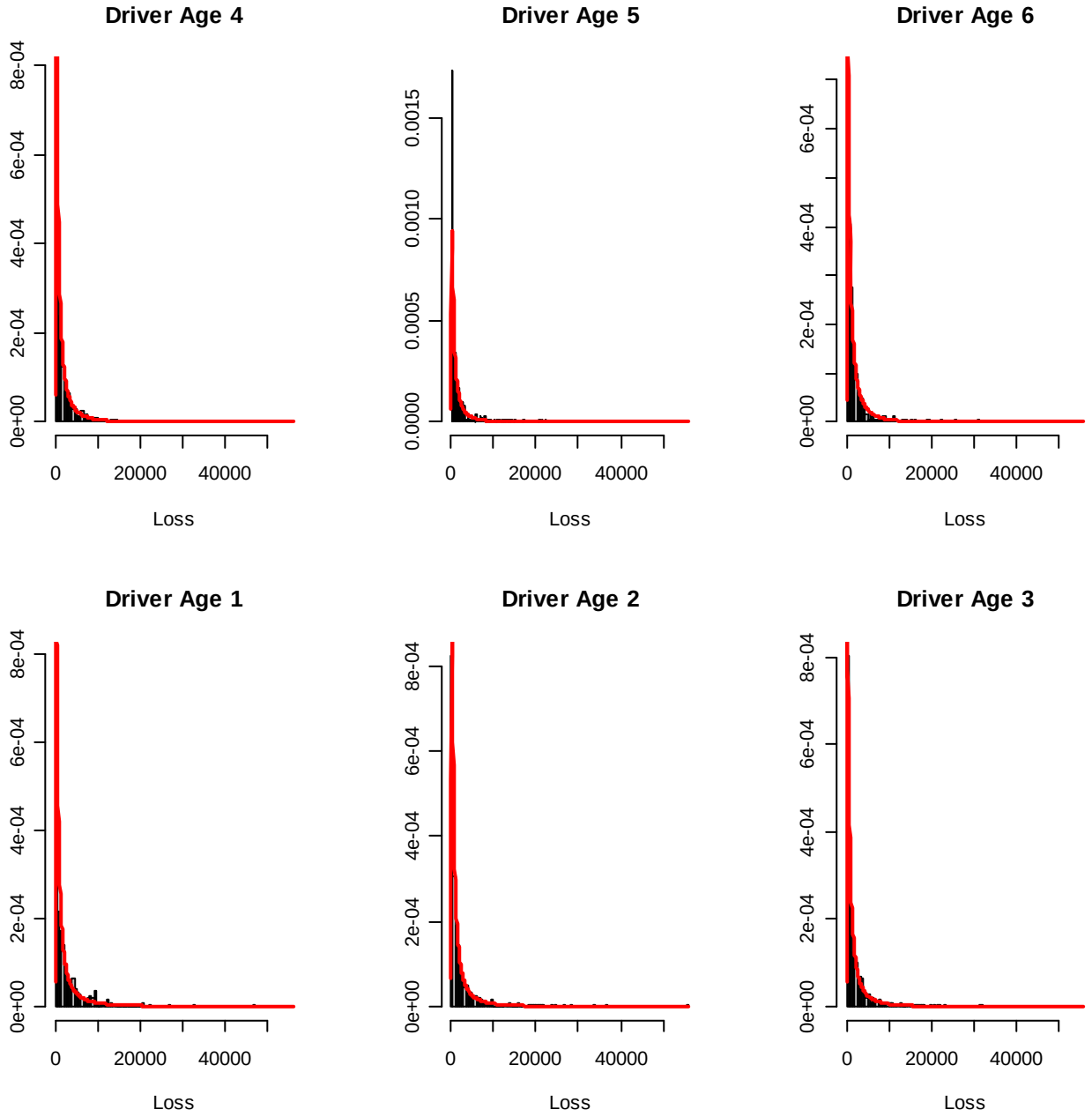
# Exponential Family of Distributions

- In a GLM, $Y_i$ may be distributed according to any member of the Exponential family of distributions
- Two Key Features of the Exponential Family:
  - The distribution is completely specified in terms of its mean and variance
  - The variance of $Y_i$ is a function of the mean
- Familiar Examples: Normal, Poisson, Gamma, Inverse Gaussian

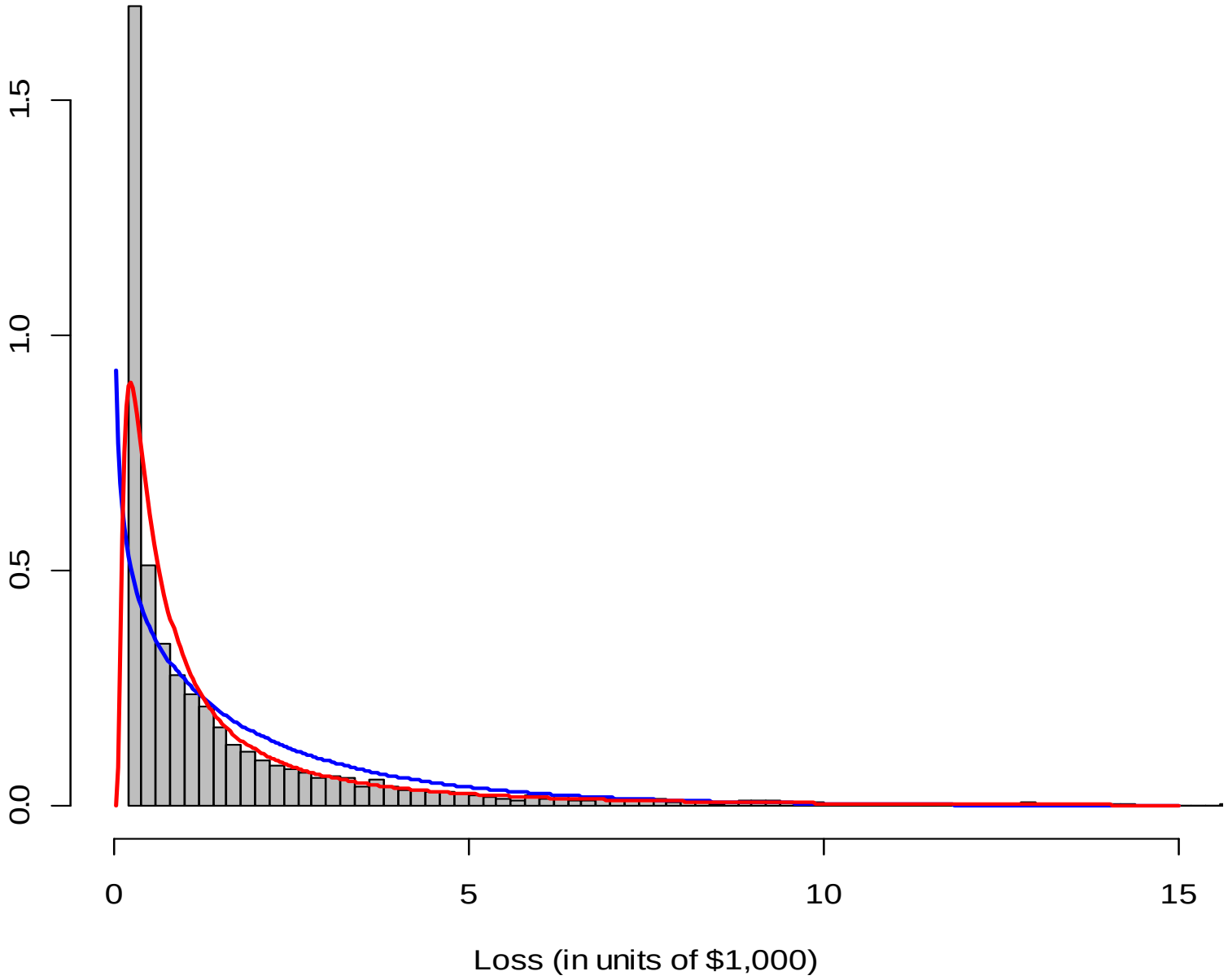
# Histograms of Loss with gamma distribution



# Histograms of Loss with inv. Gaussian distributions



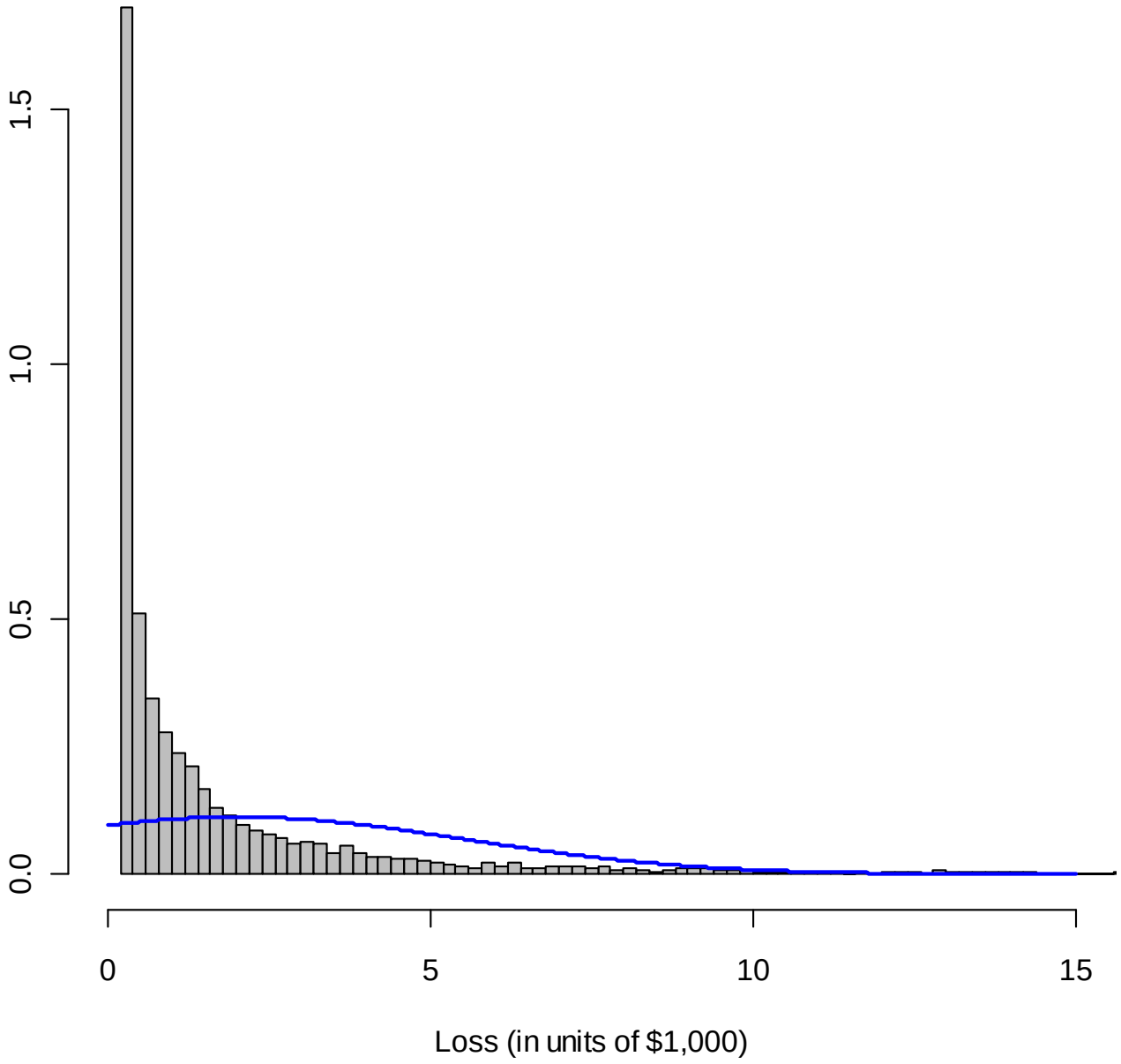
# Histogram of Loss with gamma and inv Gaussian



# Least Squares vs. Maximum Likelihood

- For each observation ($X_i$, $Y_i$), consider the probability of $Y_i$ based on assumed distribution.
- Further, consider the product of the n probabilities.
- The estimators ($\beta$) are those values that maximize the product of the n probabilities.
- (If a normal distribution is assumed, maximum likelihood is equivalent to minimizing sum of squared errors.)

# Histogram of Loss with normal distribution



# Flexibility of Variance of Y

- The variance of $Y_i$ is allowed to vary with the expected value of $Y_i$ ($\mu$)
- Variance functions link the variability of $Y_i$ to the expected value of $Y_i$ ($\mu$)

| Distribution of Y | Variance Function |
|---|---|
| Normal | 1 (variance is constant across cells) |
| Poisson | $\mu$ (variance is proportional to mean) |
| Gamma | $\mu^2$ (CV is constant across cells) |
| Inverse Gaussian (Normal) | $\mu^3$ |
| Binomial | $\mu(1-\mu)$ |
| More General Case | $\mu^p$ (Tweedie if $p < 0$, $1 < p < 2$, $p > 2$) |

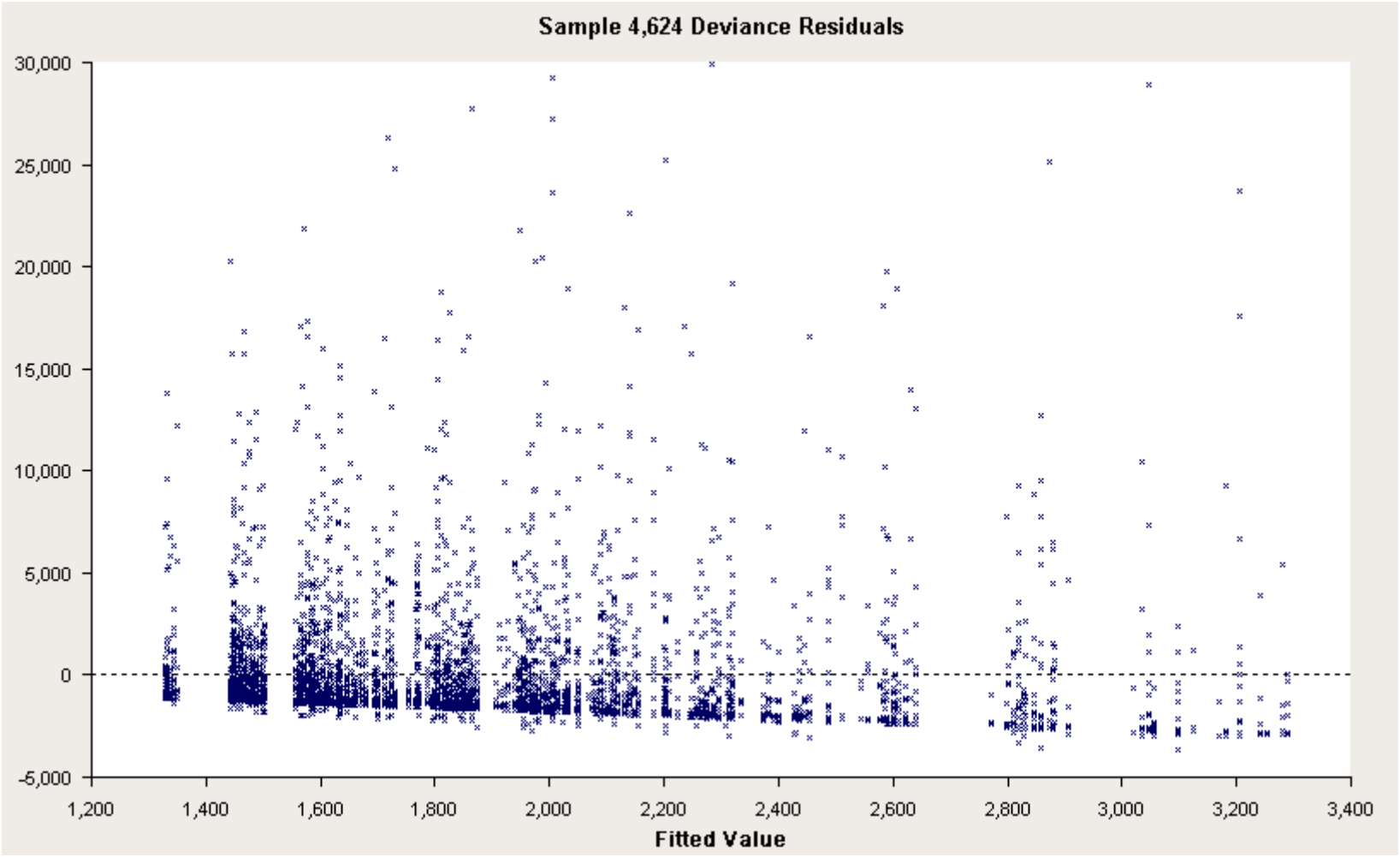
# Error structure Diagnostics

Deviance residuals against fitted value

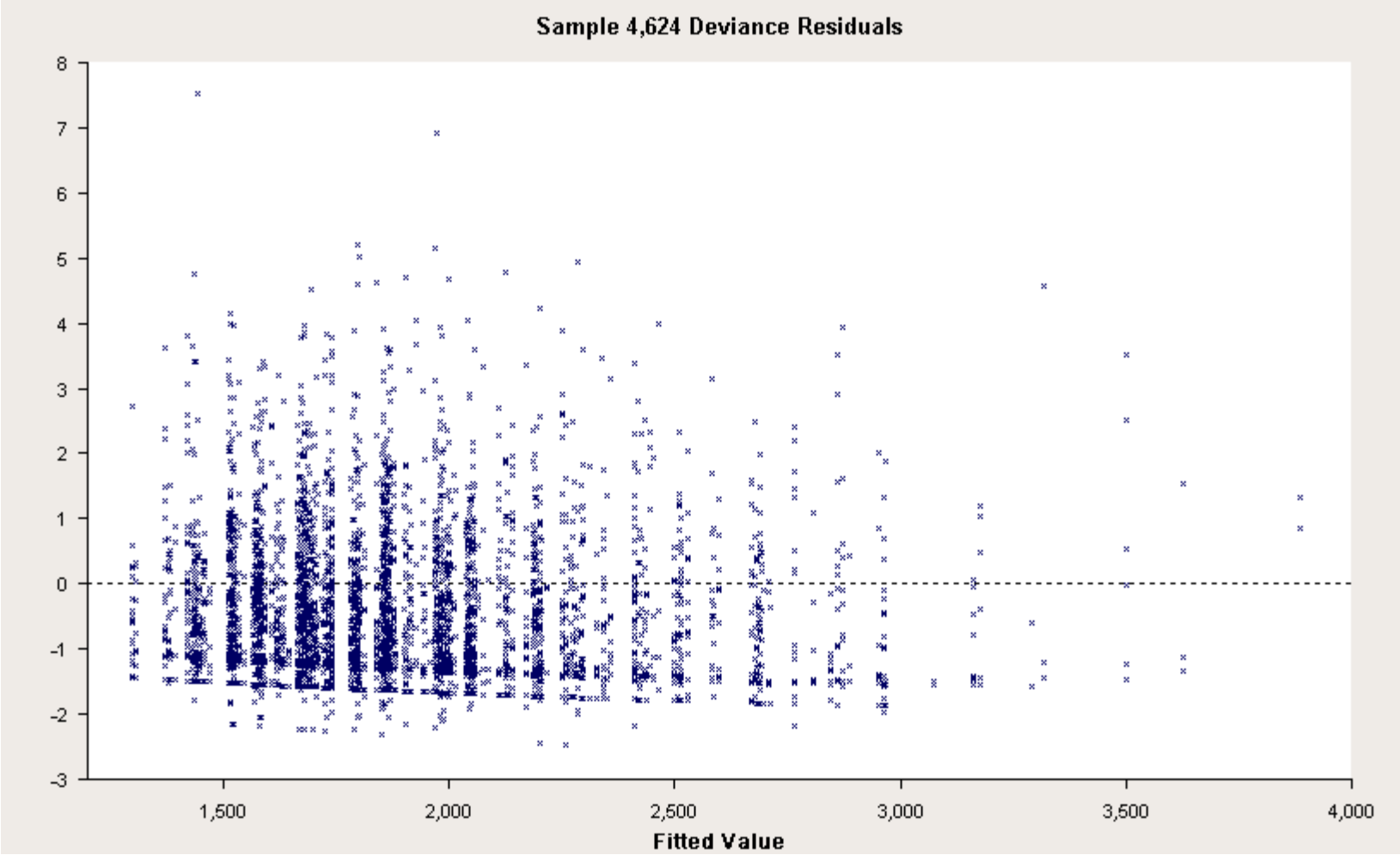
- *Deviance:* in a GLM, more weight given to differences in fitted vs. actual when variance function is small
- *Deviance residual*: square root of an observation's contribution to total deviance
- Plotting *deviance residual* against fitted value can highlight problems with error structure assumption

Histogram of deviance residuals

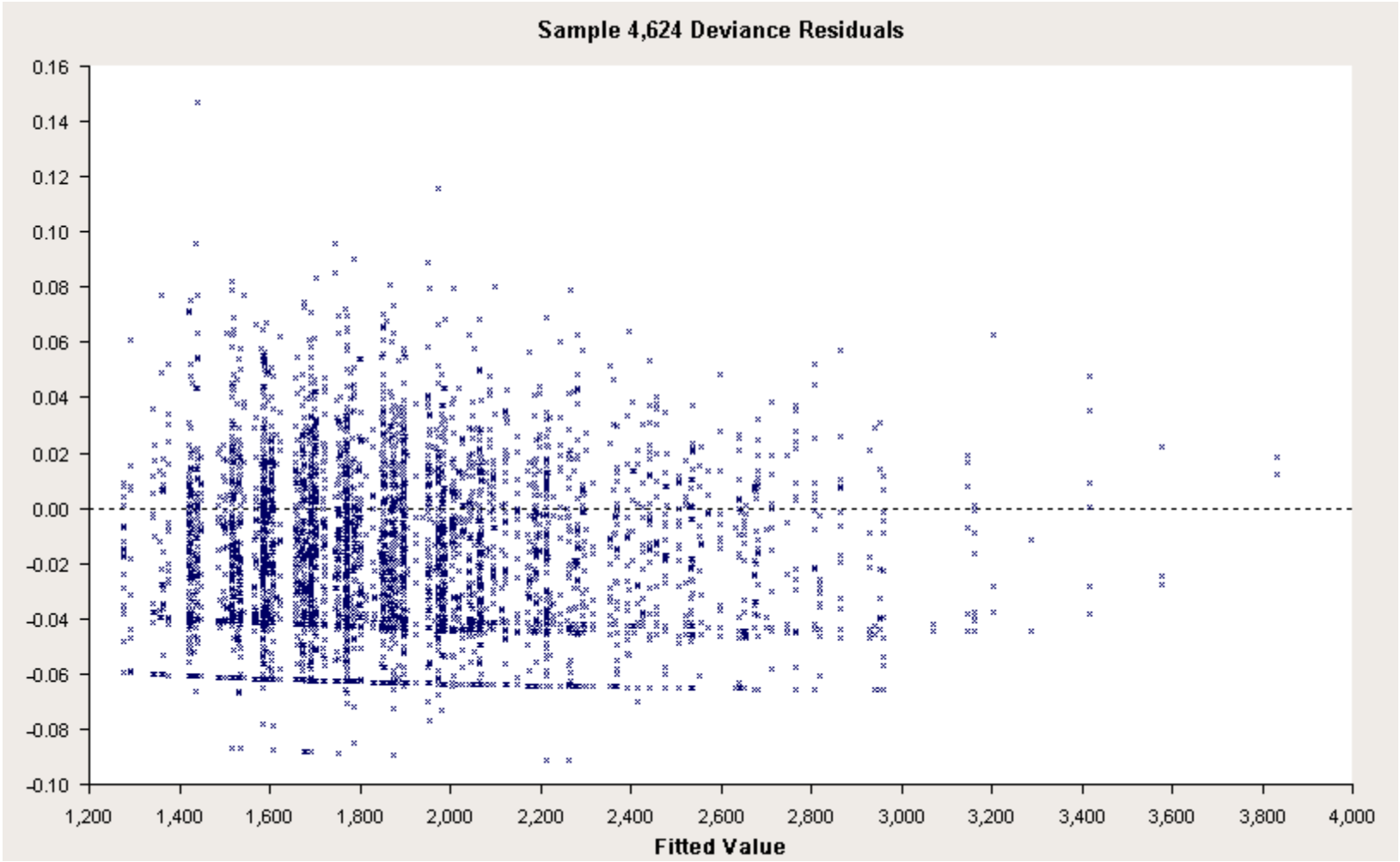
# Error structure diagnostics: Normal



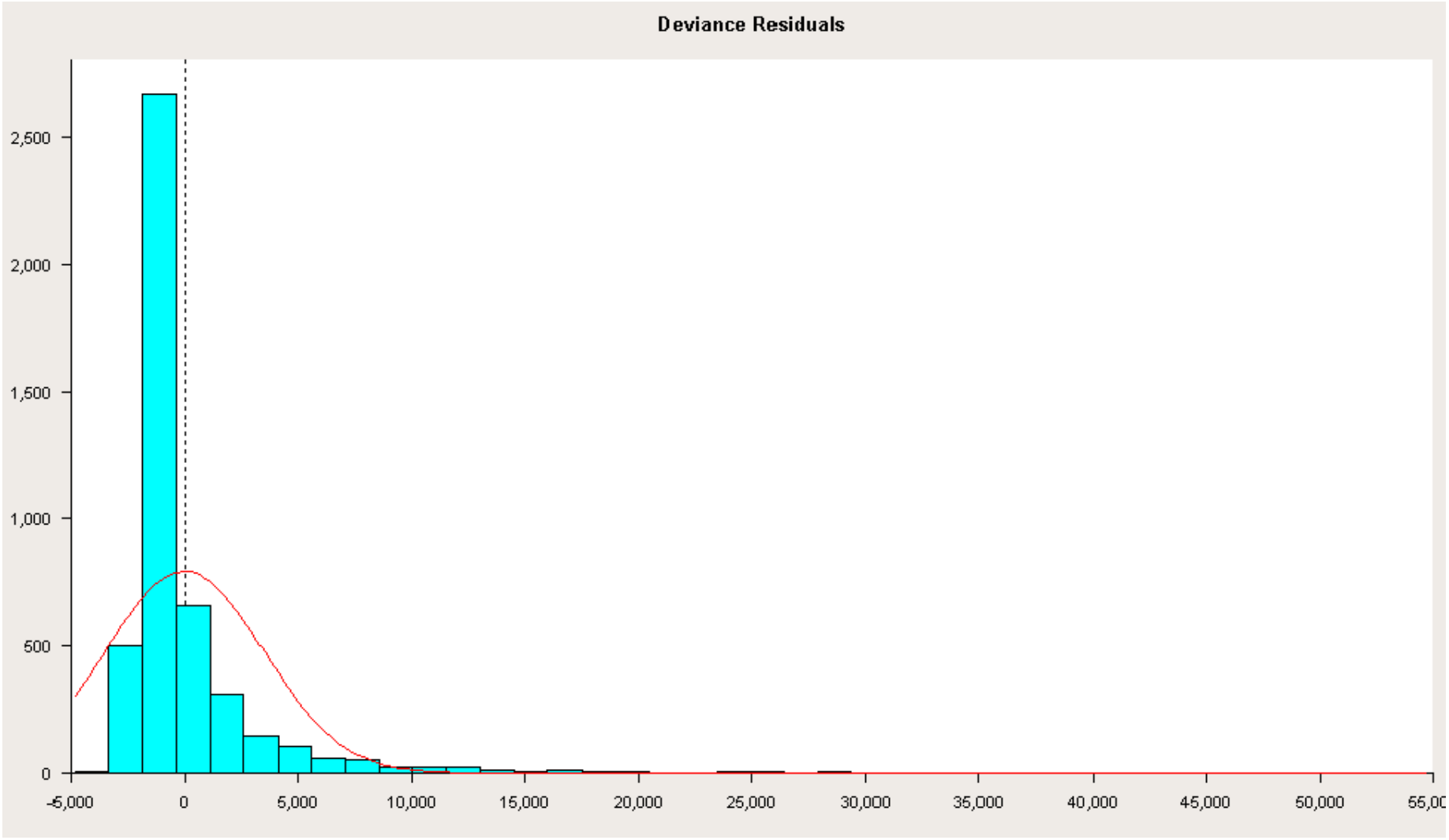
# Error structure diagnostics: Gamma



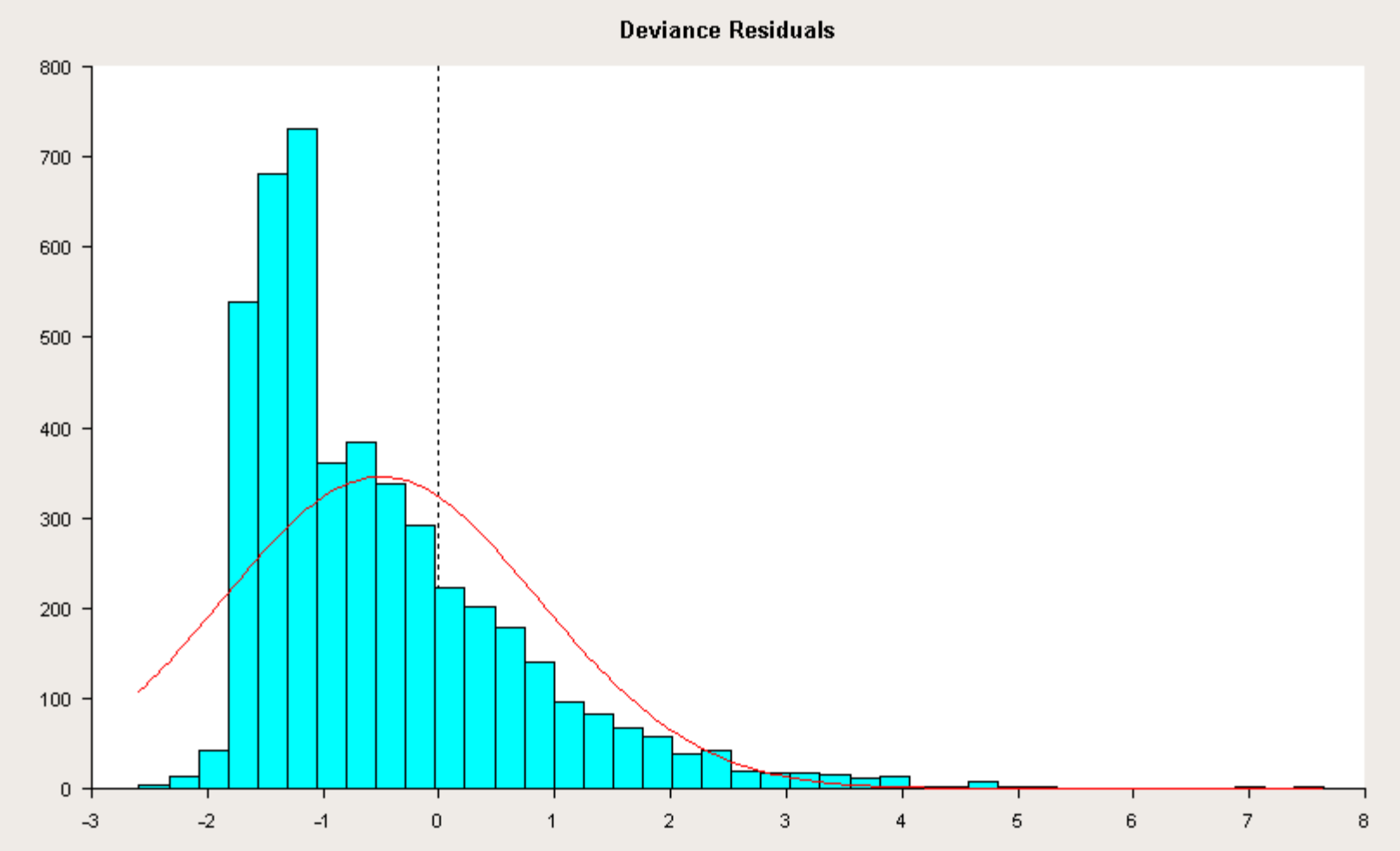
# Error structure diagnostics: Inv. Gaussian



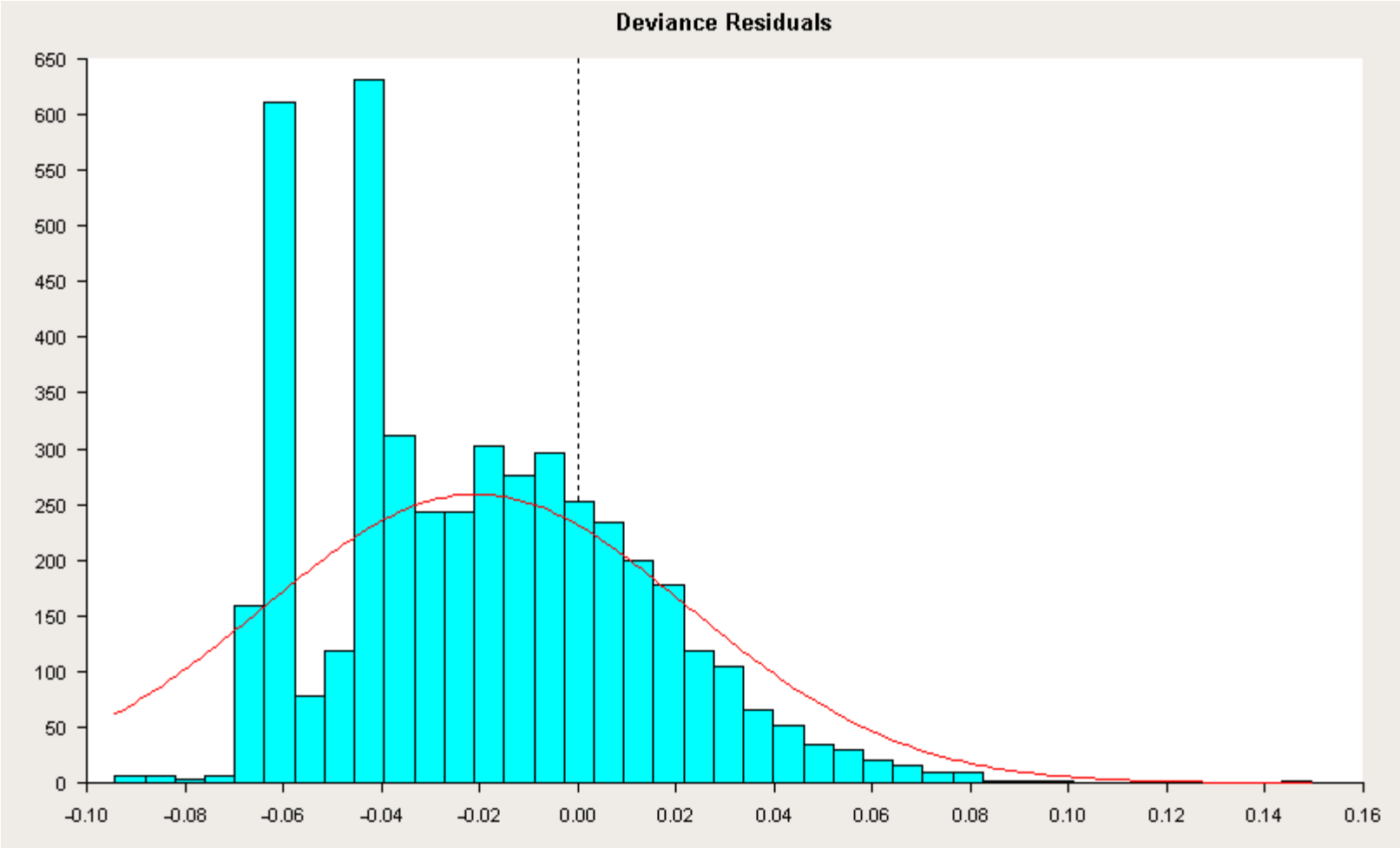
# Error structure diagnostics: Normal



# Error structure diagnostics: Gamma



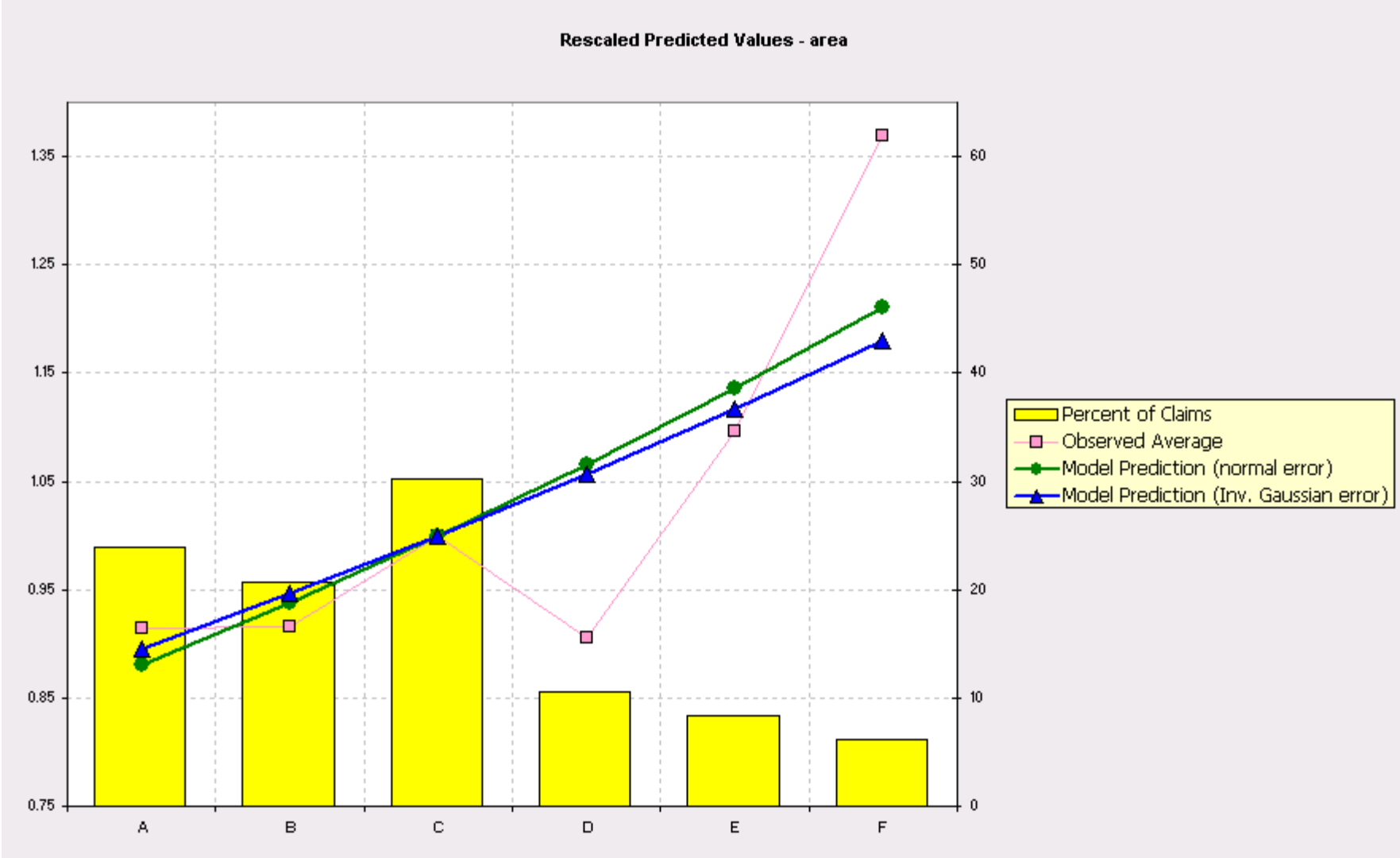
# Error structure diagnostics: Inv. Gaussian



# Model comparison: normal vs. inverse Gaussian

| | | Normal Error Structure | | | Inverse Gaussian Error Structure | | |
|---|---|---|---|---|---|---|---|
| Parameter | Name | Value | Standard Error | Exp(Value) | Value | Standard Error | Exp(Value) |
| 1 | Mean | 7.4885 | 0.06907 | 1,787.46 | 7.5198 | 0.06576 | 1,844.20 |
| 2 | area (A) | -0.0841 | 0.07396 | 0.9194 | -0.0997 | 0.06974 | 0.9052 |
| 3 | area (B) | -0.084 | 0.07759 | 0.9194 | -0.09 | 0.07287 | 0.9139 |
| - | area (C) | | | | | | |
| 4 | area (D) | -0.0807 | 0.09882 | 0.9225 | -0.1069 | 0.08989 | 0.8986 |
| 5 | area (E) | 0.1103 | 0.09273 | 1.1166 | 0.0747 | 0.10518 | 1.0776 |
| 6 | area (F) | 0.3267 | 0.08824 | 1.3864 | 0.2781 | 0.13145 | 1.3206 |
| 7 | agecat (1) | 0.3556 | 0.08171 | 1.427 | 0.2958 | 0.10378 | 1.3442 |
| 8 | agecat (2) | 0.0834 | 0.07938 | 1.087 | 0.11 | 0.07894 | 1.1162 |
| - | agecat (3) | | | | | | |
| 9 | agecat (4) | 0.0336 | 0.07913 | 1.0342 | 0.0087 | 0.07321 | 1.0088 |
| 10 | agecat (5) | -0.0613 | 0.09974 | 0.9405 | -0.0958 | 0.0841 | 0.9086 |
| 11 | agecat (6) | 0.0259 | 0.11462 | 1.0262 | -0.0235 | 0.10238 | 0.9767 |

# Parameters: Normal and Inv. Gaussian Error



# Common choices for some model types

| Target | Link Function | Error |
|---|---|---|
| Claim Frequency | log | Poisson |
| Claim Severity | log | gamma |
| Loss Costs | log | Tweedie |
| Probability of Renewal | logit | binomial |

# Further modeling

Explore significance of other variables

Group levels on our chosen variables

Better handling of $200 and $300 claims

Add interactions

(see GLM II)

# Main Ideas in Generalized Linear Modeling

Link function

Deviance

Distribution of Error

# References/Resources


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

Anderson, D., Feldblum, S., Modlin, C., Schirmacher, D., Schirmacher, E., Thandi, N. 2007. *A Practitioner's Guide to Generalized Linear Models.* CAS Discussion Paper Program

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