GLM I
Introduction to Linear & Generalized Linear Models

Casualty Actuarial Society
Ratemaking and Product Management Seminar
March 21, 2011
New Orleans, LA

Ashley Lambeth, FCAS
Safeco Insurance, Liberty Mutual Group

1

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.**

2

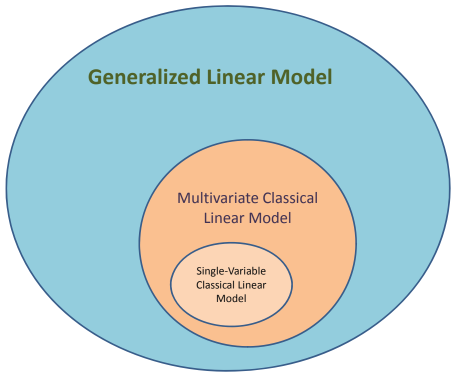
Outline

I. Introduce our Data

II. Classical Linear Modeling

III. Generalized Linear Modeling

3



4

Outline

1. **Introduce our Data**
2. Classical Linear Modeling
3. Generalized Linear Modeling

5

Cracking open the data

veh_value	exposure	clm	numclaims	claimcst0	veh_body	veh_age	gender	area	agecat
1.06	0.303901437	0	0	0	HBACK	3	F	C	2
1.03	0.648870637	0	0	0	HBACK	2	F	A	4
3.26	0.569472964	0	0	0	UTE	2	F	E	2
4.14	0.317590691	0	0	0	STNWG	2	F	D	2
0.72	0.648870637	0	0	0	HBACK	4	F	C	2
2.01	0.854209446	0	0	0	HDTOP	3	M	C	4
1.6	0.854209446	0	0	0	PANVN	3	M	A	4
1.47	0.55578371	0	0	0	HBACK	2	M	B	6
0.52	0.361396304	0	0	0	HBACK	4	F	A	3
0.38	0.52019165	0	0	0	HBACK	4	F	B	4
1.38	0.854209446	0	0	0	HBACK	2	M	A	2
1.22	0.854209446	0	0	0	HBACK	3	M	C	4
1	0.492813142	0	0	0	HBACK	2	F	C	4
7.04	0.314852841	0	0	0	STNWG	1	M	A	5
1.66	0.484599589	1	1	669.5099993	SEDAN	3	M	B	6
2.35	0.391512663	0	0	0	SEDAN	2	M	C	4
1.51	0.993839836	1	1	806.6099987	SEDAN	3	F	F	4
0.76	0.539356605	1	1	401.8054514	HBACK	3	M	C	4
0.27	0.45174538	0	0	0	HBACK	4	F	D	2
0.89	0.594113621	0	0	0	HBACK	3	F	C	3
1.95	0.594113621	0	0	0	HBACK	1	M	A	1
0.39	0.536618754	0	0	0	SEDAN	4	M	C	5
3.86	0.594113621	0	0	0	STNWG	2	F	B	2
1.37	0.59137577	0	0	0	HBACK	1	F	B	1
1.3	0.999315537	0	0	0	HBACK	2	F	A	2
1.44	0.030116359	0	0	0	HBACK	2	F	C	1
1.349	0.462696783	0	0	0	HBACK	1	F	C	5
1.39	0.317590691	0	0	0	HBACK	3	F	C	2
1	0.287474333	0	0	0	STNWG	4	M	C	3
1.51	0.06844627	0	0	0	HBACK	2	F	C	2
4.45	0.594113621	0	0	0	STNWG	1	F	C	3
2.17	0.536618754	0	0	0	SEDAN	2	F	A	4
0.87	0.854209446	0	0	0	HBACK	3	M	D	3
4.09	0.848733744	0	0	0	UTE	1	M	A	2
1.31	0.405201917	0	0	0	HBACK	2	M	C	1

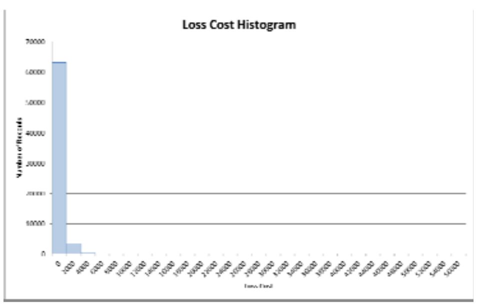
6

What is our target?

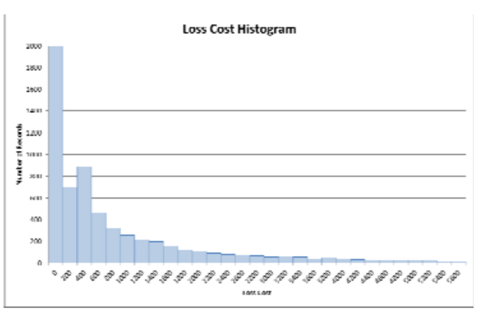
veh_value	exposure	clm	numclaims	claimcst0	veh_body	veh_age	gender	area	agecat
1.06	0.303901437	0	0	0	HBACK	3	F	C	2
1.03	0.648870637	0	0	0	HBACK	2	F	A	4
3.26	0.569472964	0	0	0	UTE	2	F	E	2
4.14	0.317590691	0	0	0	STNWG	2	F	D	2
0.72	0.648870637	0	0	0	HBACK	4	F	C	2
2.01	0.854209446	0	0	0	HDTOP	3	M	C	4
1.6	0.854209446	0	0	0	PANVN	3	M	A	4
1.47	0.55578371	0	0	0	HBACK	2	M	B	6
0.52	0.361396304	0	0	0	HBACK	4	F	A	3
0.38	0.52019165	0	0	0	HBACK	4	F	B	4
1.38	0.854209446	0	0	0	HBACK	2	M	A	2
1.22	0.854209446	0	0	0	HBACK	3	M	C	4
1	0.492813142	0	0	0	HBACK	2	F	C	4
7.04	0.314852841	0	0	0	STNWG	1	M	A	5
1.66	0.484599589	1	1	669.5099993	SEDAN	3	M	B	6
2.35	0.391512663	0	0	0	SEDAN	2	M	C	4
1.51	0.993839836	1	1	806.6099987	SEDAN	3	F	F	4
0.76	0.539356605	1	1	401.8054514	HBACK	3	M	C	4
0.27	0.45174538	0	0	0	HBACK	4	F	D	2
0.89	0.594113621	0	0	0	HBACK	3	F	C	3
1.95	0.594113621	0	0	0	HBACK	1	M	A	1
0.39	0.536618754	0	0	0	SEDAN	4	M	C	5
3.86	0.594113621	0	0	0	STNWG	2	F	B	2
1.37	0.59137577	0	0	0	HBACK	1	F	B	1
1.3	0.999315537	0	0	0	HBACK	2	F	A	2
1.44	0.030116359	0	0	0	HBACK	2	F	C	1
1.349	0.462696783	0	0	0	HBACK	1	F	C	5
1.39	0.317590691	0	0	0	HBACK	3	F	C	2
1	0.287474333	0	0	0	STNWG	4	M	C	3
1.51	0.06844627	0	0	0	HBACK	2	F	C	2
4.45	0.594113621	0	0	0	STNWG	1	F	C	3
2.17	0.536618754	0	0	0	SEDAN	2	F	A	4
0.87	0.854209446	0	0	0	HBACK	3	M	D	3
4.09	0.848733744	0	0	0	UTE	1	M	A	2

Number of Records =	67,856
Number of Claims =	4,937
Average Frequency =	0.0728
Average Claim Size =	$ 1,886.69
Max Claim Size =	$ 55,922.13
Min Claim Size =	$ 200.00

Exploring the Data



Exploring the Data



Frequency & Severity Component Modeling

veh_value	exposure	clm	numclaims	claimcst0	veh_body	veh_age	gender	area	agecat
1.06	0.303901437	0	0	0	HBACK	3	F	C	2
1.03	0.648870637	0	0	0	HBACK	2	F	A	4
3.26	0.569472964	0	0	0	UTE	2	F	E	2
4.14	0.317590691	0	0	0	STNWG	2	F	D	2
0.72	0.648870637	0	0	0	HBACK	4	F	C	2
2.01	0.854209446	0	0	0	HDTOP	3	M	C	4
1.6	0.854209446	0	0	0	PANVN	3	M	A	4
1.47	0.55578371	0	0	0	HBACK	2	M	B	6
0.52	0.361396304	0	0	0	HBACK	4	F	A	3
0.38	0.52019165	0	0	0	HBACK	4	F	B	4
1.38	0.854209446	0	0	0	HBACK	2	M	A	2
1.22	0.854209446	0	0	0	HBACK	3	M	C	4
1	0.492813142	0	0	0	HBACK	2	F	C	4
7.04	0.314852841	0	0	0	STNWG	1	M	A	5
1.66	0.484599589	1	1	669.5099993	SEDAN	3	M	B	6
2.35	0.391512663	0	0	0	SEDAN	2	M	C	4
1.51	0.993839836	1	1	806.6099987	SEDAN	3	F	F	4
0.76	0.539356605	1	1	401.8054514	HBACK	3	M	C	4
0.27	0.45174538	0	0	0	HBACK	4	F	D	2
0.89	0.594113621	0	0	0	HBACK	3	F	C	3
1.95	0.594113621	0	0	0	HBACK	1	M	A	1
0.39	0.536618754	0	0	0	SEDAN	4	M	C	5
3.86	0.594113621	0	0	0	STNWG	2	F	B	2
1.37	0.59137577	0	0	0	HBACK	1	F	B	1
1.3	0.999315537	0	0	0	HBACK	2	F	A	2
1.44	0.030116359	0	0	0	HBACK	2	F	C	1
1.349	0.462696783	0	0	0	HBACK	1	F	C	5
1.39	0.317590691	0	0	0	HBACK	3	F	C	2
1	0.287474333	0	0	0	STNWG	4	M	C	3
1.51	0.06844627	0	0	0	HBACK	2	F	C	2
4.45	0.594113621	0	0	0	STNWG	1	F	C	3
2.17	0.536618754	0	0	0	SEDAN	2	F	A	4
0.87	0.854209446	0	0	0	HBACK	3	M	D	3
4.09	0.848733744	0	0	0	UTE	1	M	A	2
1.31	0.405201917	0	0	0	HBACK	2	M	C	1

<u>Frequency Model</u>: Predicts Likelihood to file a claim.

<u>Severity Model</u>: Predicts size of claim given a claim is filed

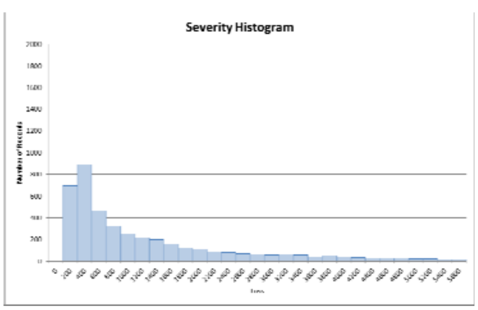
10

Our New Severity Dataset:

veh_value	exposure	clm	numclaims	claimcst0	veh_body	veh_age	gender	area	agecat
1.66	0.484599589	1	1	669.5099993	SEDAN	3	M	B	6
1.51	0.993839836	1	1	806.6099987	SEDAN	3	F	F	4
0.76	0.539356605	1	1	401.8054514	HBACK	3	M	C	4
1.89	0.654346338	1	2	1811.709997	STNWG	3	M	F	2
4.06	0.851471595	1	1	5434.439987	STNWG	2	M	F	3
1.39	0.317590691	1	1	865.789999	HBACK	3	F	A	4
2.66	0.317590691	1	1	1105.769999	STNWG	1	F	F	5
0.5	0.071184121	1	1	200	HBACK	4	F	A	5
1.16	0.596851472	1	2	739.2299995	STNWG	4	F	B	2
3.56	0.906228611	1	1	3230.599999	MCARA	3	M	F	4
2.15	0.654346338	1	1	200	SEDAN	3	F	A	5
3.03	0.90349076	1	1	200	TRUCK	1	M	C	1
2.41	0.339493498	1	1	407.8399997	STNWG	3	M	E	1
1.72	0.616016427	1	1	1619.829998	STNWG	2	F	A	6
1.92	0.616016427	1	1	1010.919998	STNWG	4	M	A	5
2.649	0.257357974	1	1	9424.349976	TRUCK	1	M	A	1
1.86	0.616016427	1	1	238.6400003	SEDAN	2	F	C	3
1.54	0.616016427	1	1	200	STNWG	2	F	B	3
1.88	0.750171116	1	1	369.1799998	PANVN	3	M	D	3
0.89	0.684462697	1	1	353.77	SEDAN	3	F	D	3
4.46	0.191649555	1	1	989.9199982	COUPE	2	M	B	2
4.32	0.646132786	1	1	736.8499985	STNWG	2	M	A	5
5.59	0.851471595	1	1	369.1499996	STNWG	1	M	C	3
1.44	0.665297741	1	1	500	UTE	4	M	F	2
2.81	0.851471595	1	1	353.77	STNWG	1	M	C	3
3	0.613278576	1	1	6372.029999	STNWG	3	F	C	2
2.29	0.353182752	1	1	500	STNWG	4	F	F	4
0.39	0.925393566	1	1	1379.039997	SEDAN	4	F	C	6
1.29	0.925393566	1	1	200	SEDAN	3	M	D	3
1.13	0.377823409	1	1	937.5599957	SEDAN	4	F	B	1
0	0.999315537	1	2	2724.339996	STNWG	4	M	A	3
2.41	0.750171116	1	1	12879.65997	HBACK	1	F	B	3
3.47	0.246406571	1	1	937.2799988	SEDAN	1	M	B	3
3.14	0.851471595	1	2	19142.13268	TRUCK	3	M	C	1
1.98	0.911704312	1	1	2659.369995	STNWG	4	M	C	6

11

Exploring our new dataset



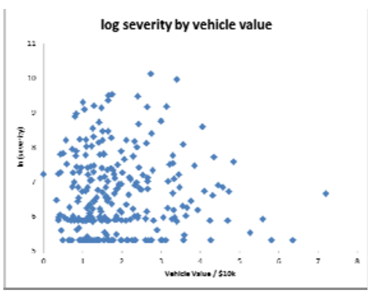
12

Our Severity Dataset reorganized:

Predictor Variables						Target Variable
veh_value	veh_body	veh_age	gender	area	agecat	Severity
1.66	SEDAN	3	M	B	6	669.5099993
1.51	SEDAN	3	F	F	4	806.6099987
0.76	HBACK	3	M	C	4	401.8054514
1.89	STNWG	3	M	F	2	905.8549986
4.06	STNWG	2	M	F	3	5434.439987
1.39	HBACK	3	F	A	4	865.789999
2.66	STNWG	1	F	F	5	1105.769999
0.5	HBACK	4	F	A	5	200
1.16	STNWG	4	F	B	2	369.6149998
3.56	MCARA	3	M	F	4	3230.599999
2.15	SEDAN	3	F	A	5	200
3.03	TRUCK	1	M	C	1	200
2.41	STNWG	3	M	E	1	407.8399997
1.72	STNWG	2	F	A	6	1619.829998
1.92	STNWG	4	M	A	5	1010.919998
2.649	TRUCK	1	M	A	1	9424.349976
1.86	SEDAN	2	F	C	3	238.6400003
1.54	STNWG	2	F	B	3	200
1.88	PANVN	3	M	D	3	369.1799998
0.89	SEDAN	3	F	D	3	353.77
4.46	COUPE	2	M	B	2	989.9199982
4.32	STNWG	2	M	A	5	736.8499985
5.59	STNWG	1	M	C	3	369.1499996
1.44	UTE	4	M	F	2	500
2.81	STNWG	1	M	C	3	353.77
3	STNWG	3	F	C	2	6372.029999
2.29	STNWG	4	F	F	4	500
0.39	SEDAN	4	F	C	6	1379.039997
1.29	SEDAN	3	M	D	3	200
1.13	SEDAN	4	F	B	1	937.5599957
0	STNWG	4	M	A	3	1362.169998

13

Variable Exploration (continuous)



14

Variable Exploration (categorical)

Veh_Body	Claims	Av Severity
BUS	10	$ 1,336
CONVT	3	$ 2,296
COUPE	75	$ 2,503
HBACK	1330	$ 1,947
HDTOP	136	$ 2,168
MCARA	15	$ 712
MIBUS	45	$ 2,580
PANVN	68	$ 1,958
RDSTR	3	$ 456
SEDAN	**1598**	**$ 1,678**
STNWG	1248	$ 1,894
TRUCK	130	$ 2,458
UTE	276	$ 2,164

Veh_Age (1 is newest)	Claims	Av Severity
1	876	$ 1,775
2	1354	$ 1,836
3	**1446**	**$ 1,880**
4	1261	$ 2,026

Gender	Claims	Av Severity
M	2105	$ 2,093
F	**2832**	**$ 1,733**

Area	Claims	Av Severity
A	1181	$ 1,754
B	1021	$ 1,758
C	**1493**	**$ 1,919**
D	524	$ 1,739
E	413	$ 2,104
F	305	$ 2,629

Agecat (1 is youngest)	Claims	Av Severity
1	525	$ 2,490
2	1000	$ 1,985
3	**1189**	**$ 1,793**
4	1185	$ 1,810
5	648	$ 1,638
6	390	$ 1,753

15

Outline

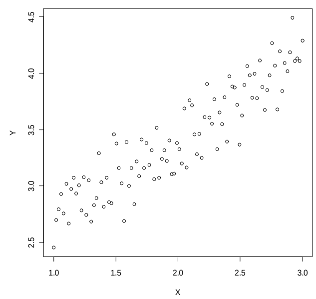
I. Introduce our Data

II. Classical Linear Modeling

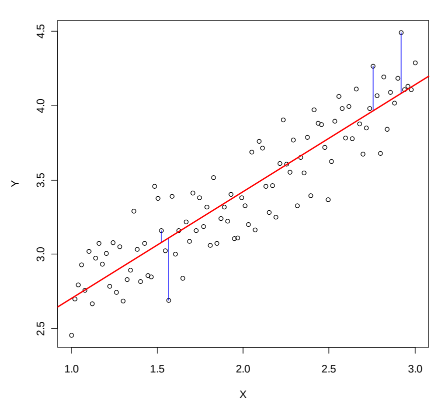
III. Generalized Linear Modeling

16

What is Linear Modeling?



17



18

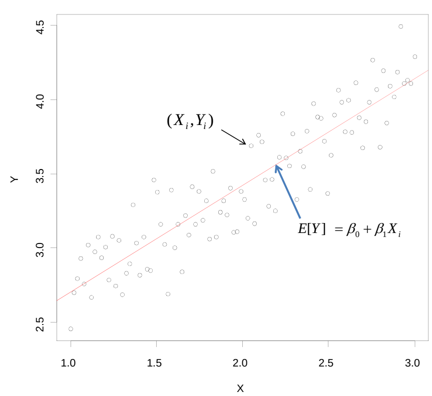
Classical Linear Model; Moving Parts

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

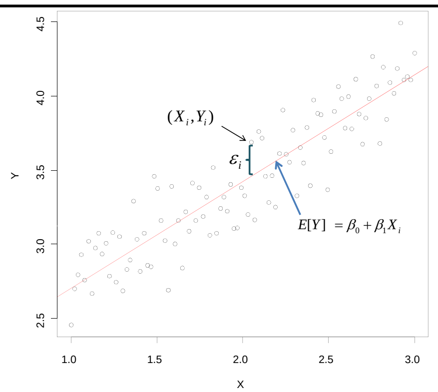
1. Y_i is the (observed) value of response variable in the i^{th} trial
2. β_0 and β_1 are parameters
3. X_i is the (observed) value of the predictor variable in the i^{th} trial
4. ε_i is a random error term with mean 0 and variance σ^2
5. i= 1, 2, . . ., n

19

19



20



$$Y_i = E[Y_i] + \varepsilon_i$$

21

Multivariate Classical Linear Model

$$E[Y_i] = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} \ldots$$

Or, in matrix notation:

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

22

How does this pertain to insurance modeling?

Gender	Weight	Av Severity
M	2105	$ 2,093
F	**2832**	**$ 1,733**

This categorical variable requires a two-parameter model.

23

How does this pertain to insurance modeling?

Gender	Weight	Av Severity
M	2105	$ 2,093
F	**2832**	**$ 1,733**

This categorical variable requires a two-parameter model.

Predicted Severity

Binary Variable:
1 for male, 0 for female

$$E[Y] = \beta_0 + \beta_1 X_1$$

Base (female) Severity

Additional Severity for being male

24

Gender		
	Weight	Av Severity
M	2105	$ 2,093
F	**2832**	**$ 1,733**

This categorical variable requires a two-parameter model.

$$E[Y] = \beta_0 + \beta_1 X_1$$

Parameter Estimates

β_0 = 1733
β_1 = 360

25

Let's add another variable

Gender		
	Weight	Av Severity
M	2105	$ 2,093
F	**2832**	**$ 1,733**

Area		
	Weight	Av Severity
A	1181	$ 1,754
B	1021	$ 1,758
C	**1493**	**$ 1,919**
D	524	$ 1,739
E	413	$ 2,104
F	305	$ 2,629

26

Let's add another variable

Gender		
	Weight	Av Severity
M	2105	$ 2,093
F	**2832**	**$ 1,733**

Area		
	Weight	Av Severity
A	1181	$ 1,754
B	1021	$ 1,758
C	**1493**	**$ 1,919**
D	524	$ 1,739
E	413	$ 2,104
F	305	$ 2,629

$$E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6$$

X1: Binary variable (Male = 1, Female = 0)
X2: Binary variable (Area A = 1, else = 0)
X3: Binary variable (Area B = 1, else = 0)
X4: Binary variable (Area D = 1, else = 0)
X5: Binary variable (Area E = 1, else = 0)
X6: Binary variable (Area F = 1, else = 0)

β0: Base severity (female from Area C)
β1: Add'l severity from being male
β2: Add'l severity from Area A
β3: Add'l severity from Area B
β4: Add'l severity from Area D
β5: Add'l severity from Area E
β6: Add'l severity from Area F

Example: Female from Area F $E[Y] = \beta_0 + \beta_6$

27

Modeling Software Output

[GLM fit: Identity Link Function, Normal Error Structure]

Parameter Number	Name	Value	Standard Error	Standard Error (%)	Weight	Weight (%)
1	Mean	1,769.64	99.14477	5.6	4,937	100
-	gender (F)				2,832	57.4
2	gender (M)	361.864	100.18706	27.7	2,105	42.6
3	area (A)	-174.419	135.53147	77.7	1,181	23.9
4	area (B)	-170.408	141.33498	82.9	1,021	20.7
-	area (C)				1,493	30.2
5	area (D)	-174.622	176.69103	101.2	524	10.6
6	area (E)	173.704	193.48309	111.4	413	8.4
7	area (F)	707.8558	218.65201	30.9	305	6.2

28

How good is our fit?

Claims

Area	Gender M	Gender F
A	519	662
B	449	572
C	618	875
D	208	316
E	183	230
F	128	177

Actual Severity

Area	Gender M	Gender F
A	$ 1,899	$ 1,641
B	$ 1,939	$ 1,616
C	$ 2,100	**$ 1,792**
D	$ 1,666	$ 1,787
E	$ 2,579	$ 1,726
F	$ 3,386	$ 2,082

Predicted Severity

Area	Gender M	Gender F
A	$ 1,957	$ 1,595
B	$ 1,961	$ 1,599
C	$ 2,132	**$ 1,770**
D	$ 1,957	$ 1,595
E	$ 2,305	$ 1,943
F	$ 2,839	$ 2,477

29

Classical Linear Model; Assumptions

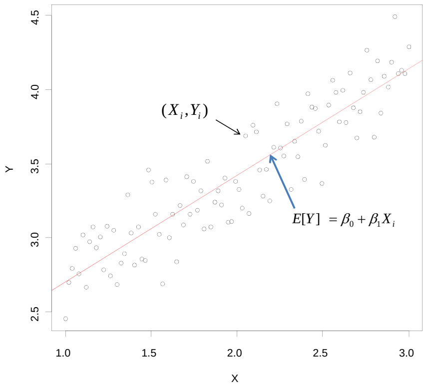
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$Y_i = \mu_{Y_i} + \varepsilon_i$$

1. Y_i is the sum of a constant term and a random term
2. $E\{Y_i\} = \beta_0 + \beta_1 X_i$ = "Linear Predictor"
   1. This implies a linear relationship between X_i and Y_i
3. The error terms ε_i are random variables which;
   1. Are independent
   2. Are normally distributed
   3. Have constant variance, σ^2.
4. Therefore, the responses, Y_i are also independent normally distributed random variables with constant variance, σ^2.

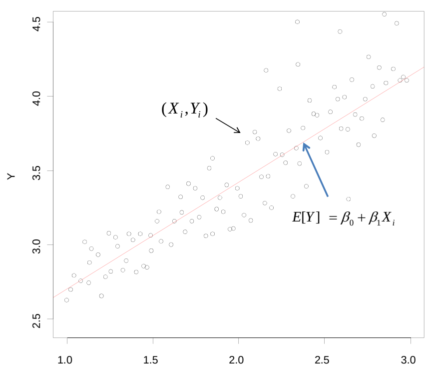
30

30



31

A slightly different scatterplot



32

Outline

I. Introduce our Data

II. Classical Linear Modeling

III. Generalized Linear Modeling

33

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

34

Flexibility of Relationship between X & Y

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

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

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

35

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:

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

36

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)

37

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

38

Why choose log link?

- Convenience: match structure of rating plan/scorecard
- Intuition: Do rating variables have additive or multiplicative effects on severity?
- Evaluation: We can test the appropriateness of the link function

39

Model output: normal error / log link

[GLM fit: Log Link Function, Normal Error Structure]

Parameter Number	Name	Value	Standard Error	Standard Error (%)	Weight	Weight (%)	Exp(Value)
1	Mean	7.47	0.0539	0.7	4,937	100	1,749.90
-	gender (F)				2,832	57.4	
2	gender (M)	0.2051	0.05196	25.3	2,105	42.6	1.2277
3	area (A)	-0.0959	0.07417	77.3	1,181	23.9	0.9085
4	area (B)	-0.0918	0.0774	84.3	1,021	20.7	0.9123
-	area (C)				1,493	30.2	
5	area (D)	-0.1069	0.09972	93.3	524	10.6	0.8986
6	area (E)	0.098	0.09284	94.7	413	8.4	1.103
7	area (F)	0.3303	0.08776	26.6	305	6.2	1.3914

$E[Y] = \exp(\beta_1) \cdot \exp(\beta_2 X_2) \cdot \exp(\beta_3 X_3) \cdot \ldots \cdot \exp(\beta_7 X_7)$

40

Model output: normal error / log link

[GLM fit: Log Link Function, Normal Error Structure]

Parameter Number	Name	Value	Standard Error	Standard Error (%)	Weight	Weight (%)	Exp(Value)
1	Mean	7.47	0.0539	0.7	4,937	100	1,749.90
-	gender (F)				2,832	57.4	
2	gender (M)	0.2051	0.05196	25.3	2,105	42.6	1.2277
3	area (A)	-0.0959	0.07417	77.3	1,181	23.9	0.9085
4	area (B)	-0.0918	0.0774	84.3	1,021	20.7	0.9123
-	area (C)				1,493	30.2	
5	area (D)	-0.1069	0.09972	93.3	524	10.6	0.8986
6	area (E)	0.098	0.09284	94.7	413	8.4	1.103
7	area (F)	0.3303	0.08776	26.6	305	6.2	1.3914

<u>For Gender=F, Area=C</u>

$E[Y] = \exp(\beta_1) \cdot \exp(\beta_2 \cdot 0) \cdot \exp(\beta_3 \cdot 0) \cdot \ldots \cdot \exp(\beta_7 \cdot 0)$

$= \exp(7.47) = 1{,}749.90$

41

Model output: normal error / log link

[GLM fit: Log Link Function, Normal Error Structure]

Parameter Number	Name	Value	Standard Error	Standard Error (%)	Weight	Weight (%)	Exp(Value)
1	Mean	7.47	0.0539	0.7	4,937	100	1,749.90
-	~~gender (F)~~				2,832	57.4	
2	gender (M)	0.2051	0.05196	25.3	2,105	42.6	1.2277
3	area (A)	-0.0959	0.07417	77.3	1,181	23.9	0.9085
4	area (B)	-0.0918	0.0774	84.3	1,021	20.7	0.9123
-	area (C)				1,493	30.2	
5	area (D)	-0.1069	0.09972	93.3	524	10.6	0.8986
6	area (E)	0.098	0.09284	94.7	413	8.4	1.103
7	area (F)	0.3303	0.08776	26.6	305	6.2	1.3914

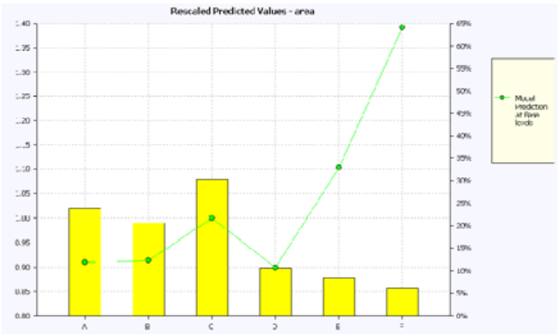
<u>For Gender=M, Area=A</u>

$E[Y] = \exp(\beta_1) \cdot \exp(\beta_2 \cdot 1) \cdot \exp(\beta_3 \cdot 1) \cdot \ldots \cdot \exp(\beta_7 \cdot 0)$

$= \exp(7.47) \cdot \exp(0.2051) \cdot \exp(-0.0959)$

$= 1{,}749.90 \cdot 1.2277 \cdot .9085 = 1{,}951.78$

42



43

Do estimates match the data?

Actual Severity

Area	Gender M	Gender F
A	$ 1,899	$ 1,641
B	$ 1,939	$ 1,616
C	$ 2,100	$ 1,792
D	$ 1,666	$ 1,787
E	$ 2,579	$ 1,726
F	$ 3,386	$ 2,082

Predicted Severity (additive)

Area	Gender M	Gender F
A	$ 1,957	$ 1,595
B	$ 1,961	$ 1,599
C	$ 2,132	$ 1,770
D	$ 1,957	$ 1,595
E	$ 2,305	$ 1,943
F	$ 2,839	$ 2,477

Predicted Severity (multiplicative)

Area	Gender M	Gender F
A	$ 1,952	$ 1,590
B	$ 1,960	$ 1,596
C	$ 2,148	$ 1,750
D	$ 1,931	$ 1,573
E	$ 2,370	$ 1,930
F	$ 2,989	$ 2,435

44

Linear vs. Generalized Linear Model

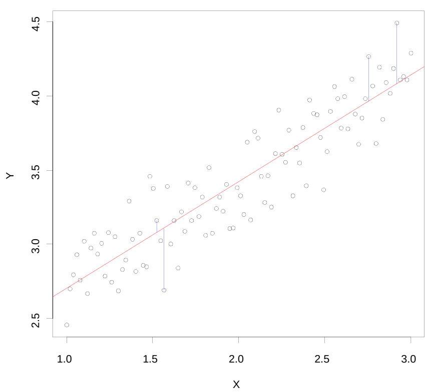
Assumption	Linear Regression Model	Generalized Linear Model
Relationship between X and Y	Y is a linear combination of X	Y is a function of a linear combination of X
Distribution of Y	Normal	Any distribution from the Exponential family
Variance of Y	Constant	Function of the mean

45

Flexibility of Distribution of Y

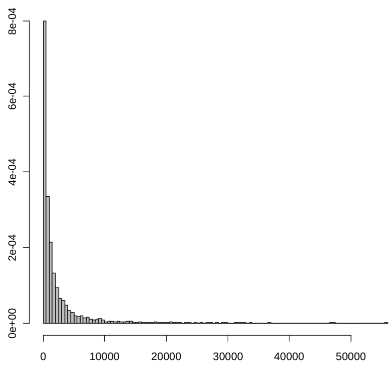
- Least-squares estimation implicitly assumes observations come from normal distribution

46



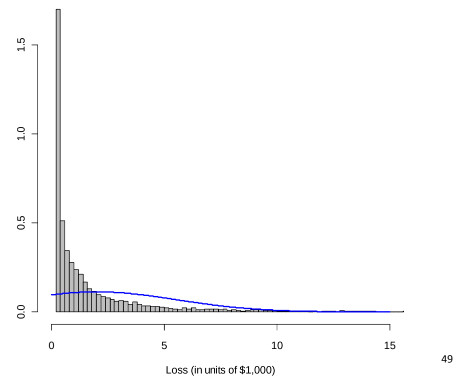
47

Histogram of loss



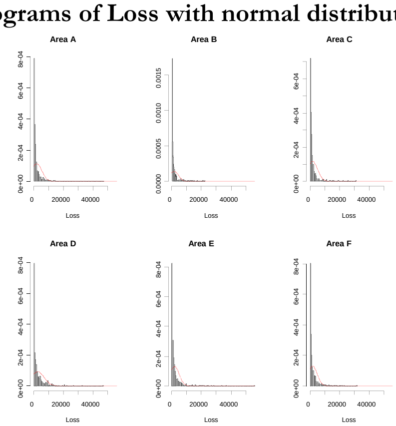
48

Histogram of Loss with normal distribution



49

Histograms of Loss with normal distribution



50

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

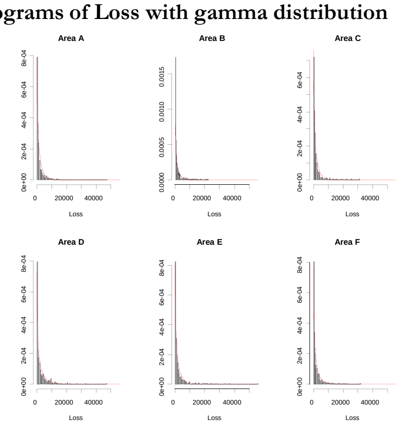
51

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

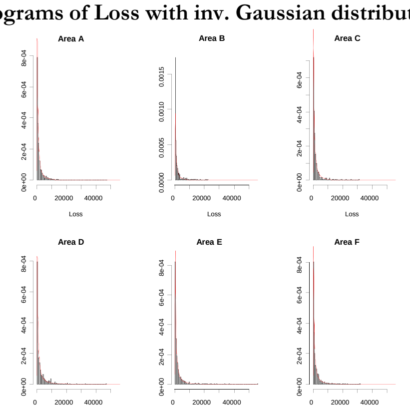
52

Histograms of Loss with gamma distribution



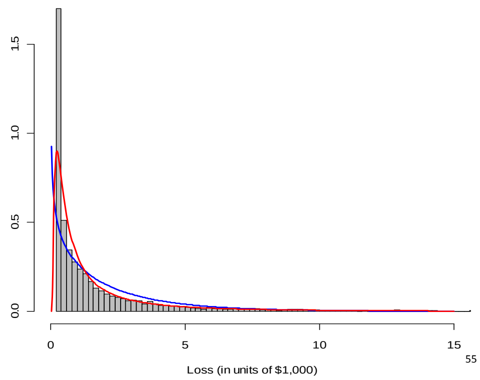
53

Histograms of Loss with inv. Gaussian distributions



54

Histogram of Loss with gamma and inv Gaussian



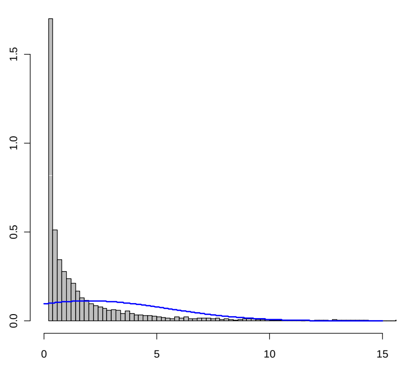
55

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 (β) 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.)

56

Histogram of Loss with normal distribution

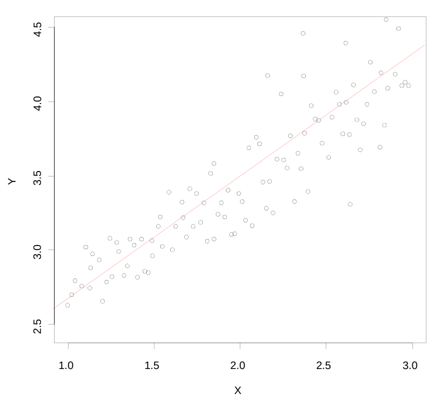


57

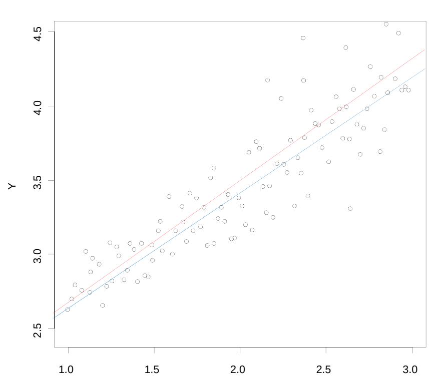
Linear vs. Generalized Linear Model

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

58



59



60

Flexibility of Variance of Y

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

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

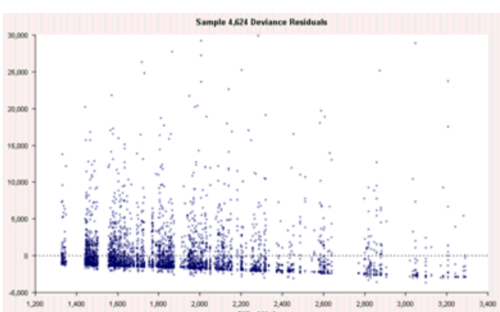
61

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

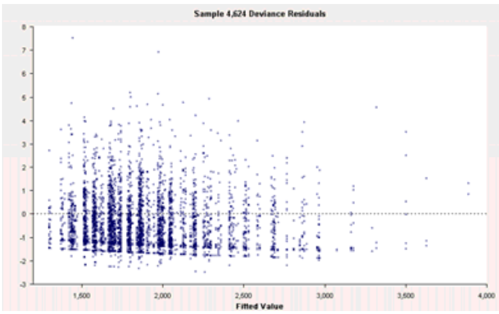
62

Error structure diagnostics: Normal



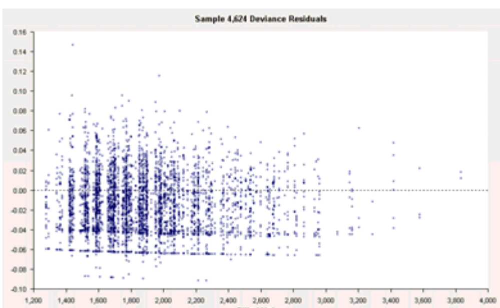
63

Error structure diagnostics: Gamma



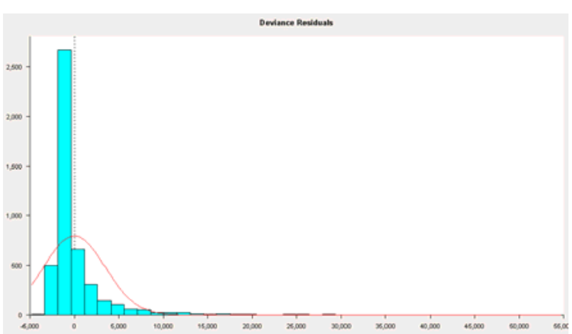
64

Error structure diagnostics: Inv. Gaussian



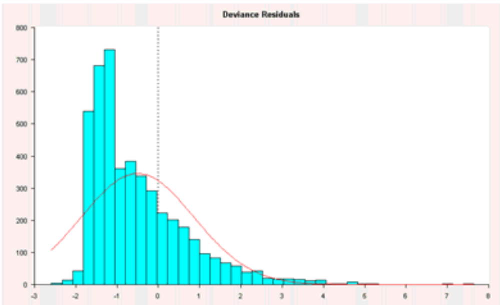
65

Error structure diagnostics: Normal



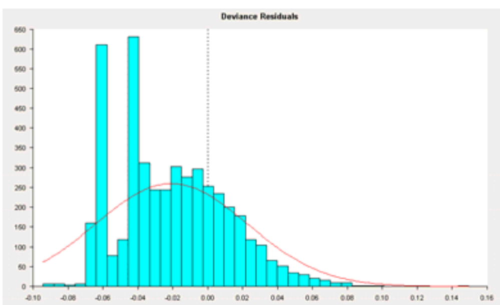
66

Error structure diagnostics: Gamma



67

Error structure diagnostics: Inv. Gaussian



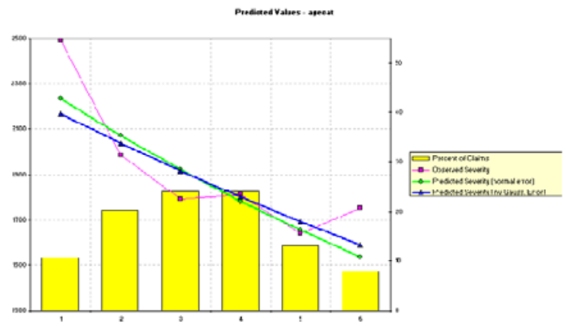
68

Model comparison: normal vs. inverse Gaussian

		[Log Link, Normal Error Structure]			[Log Link, Inv Gaussian Error Structure]		
Parameter	Name	Value	Standard Error	Exp(Value)	Value	Standard Error	Exp(Value)
1	Mean	7.47	0.0539	1,749.90	7.49	0.0507	1,783.61
-	gender (F)						
2	gender (M)	0.2051	0.05196	1.2277	0.1711	0.05212	1.1867
3	area (A)	-0.0959	0.07417	0.9085	-0.0934	0.06874	0.9108
4	area (B)	-0.0918	0.0774	0.9123	-0.0941	0.07155	0.9102
-	area (C)						
5	area (D)	-0.1069	0.09972	0.8986	-0.0783	0.08922	0.9247
6	area (E)	0.098	0.09284	1.103	0.0664	0.10333	1.0687
7	area (F)	0.3303	0.08776	1.3914	0.2866	0.12842	1.3319

69

Parameters: Normal and Inv. Gaussian Error



70

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

71

Further modeling

- Explore significance of other variables
- Group levels on our chosen variables
- Add interactions
- (see GLM II)

72

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

73