GLM I
An Introduction to Generalized Linear Models

CAS Ratemaking and Product Management Seminar
March 2012

Presented by: Tanya D. Havlicek, ACAS, MAAA

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.

Outline

- Overview of Statistical Modeling
- Linear Models
  - ANOVA
  - Simple Linear Regression
  - Multiple Linear Regression
  - Categorical Variables
  - Transformations
- Generalized Linear Models
  - Why GLM?
  - From Linear to GLM
  - Basic Components of GLM's
  - Common GLM structures
- References

Generic Modeling Schematic

Predictor Vars

Driver Age

Region

Relative Equity

Credit Score

Weights

Claims

Exposures

Premium

Response Vars

Losses

Default

Persistency

Statistical Model

Model Results

Parameters

Validation Statistics

Basic Linear Model Structures - Overview

- Simple ANOVA :
  - $Y_{ij} = \mu + e_{ij}$ or more generally $Y_{ij} = \mu + \psi_i + e_{ij}$
  - In Words: Y is equal to the mean for the group with random variation and possibly fixed variation
  - Traditional Classification Rating – Group Means
  - Assumptions: errors independent & follow $N(0,\sigma_e^2)$
  - $\sum \psi_i = 0 \quad i = 1,\ldots,k$ (fixed effects model)
  - $\psi_i \sim N(0,\sigma_\psi^2)$ (random effects model)

Basic Linear Model Structures - Overview

- Simple Linear Regression : $y_i = b_o + b_1 x_i + e_i$
  - Assumptions:
    - linear relationship
    - errors independent and follow $N(0,\sigma_e^2)$
- Multiple Regression : $y_i = b_o + b_1 x_{1i} + \ldots . + b_n x_{ni} + e_i$
  - Assumptions: same, but with n independent random variables (RV's)
- Transformed Regression : transform x, y, or both; maintain errors are $N(0,\sigma_e^2)$

$$y_i = \exp(x_i) \Rightarrow \log(y_i) = x_i$$

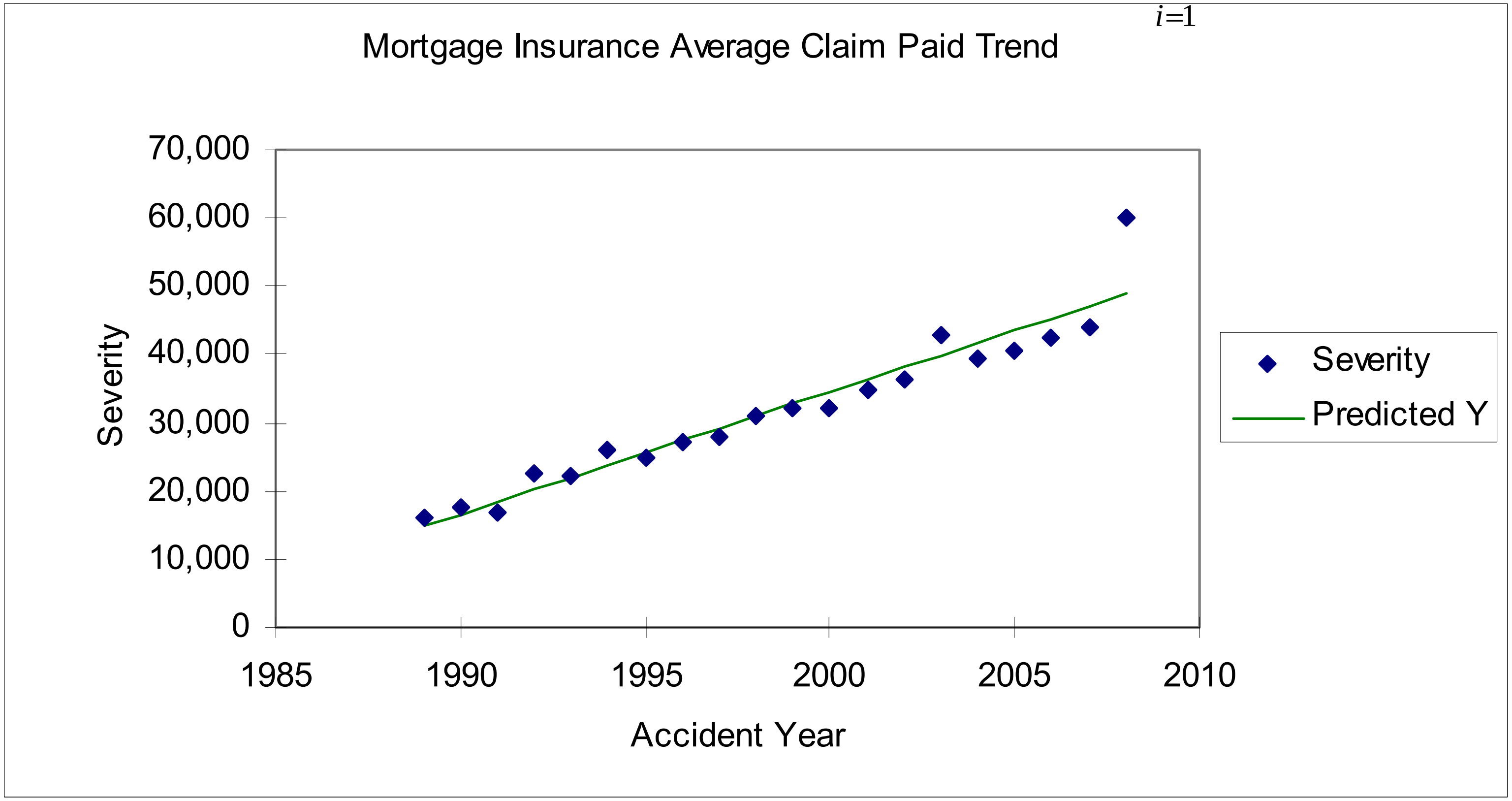
Simple Regression *(special case of multiple regression)*

- Model: $Y_i = b_o + b_1 X_i + e_i$
  - Y is the dependent variable explained by X, the independent variable
  - Y could be Pure Premium, Default Frequency, etc
  - Want to estimate relationship of how Y depends on X using observed data
  - Prediction: $Y= b_o + b_1 x^*$ for some new x^* (usually with some confidence interval)

Simple Regression

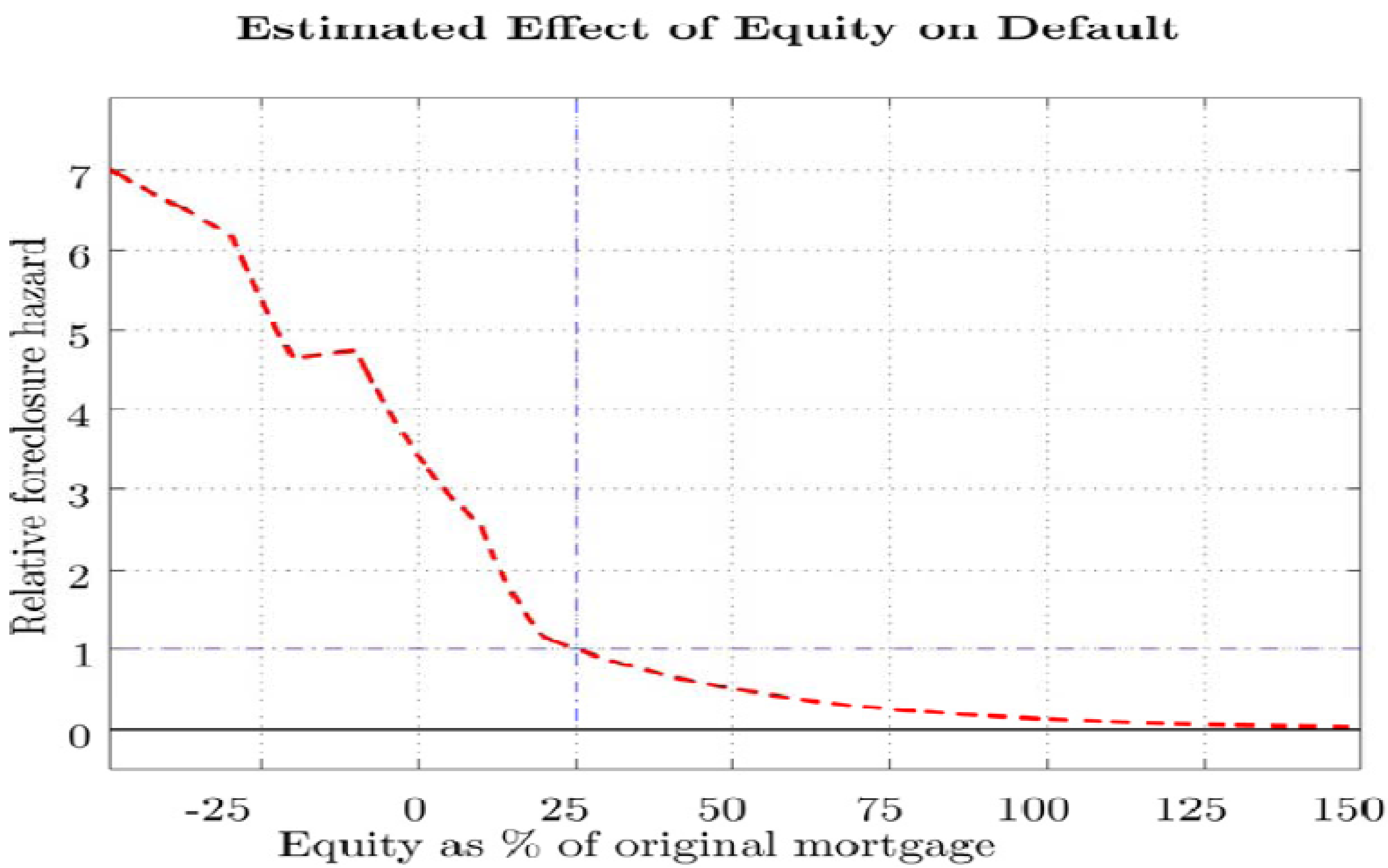
- A formalization of best fitting a line through data with a ruler and a pencil
- Correlative relationship
- Simple e.g. determine a trend to apply

$$\beta = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{N}(X_i - \bar{X})^2}, \quad a = \bar{Y} - \beta\bar{X}$$



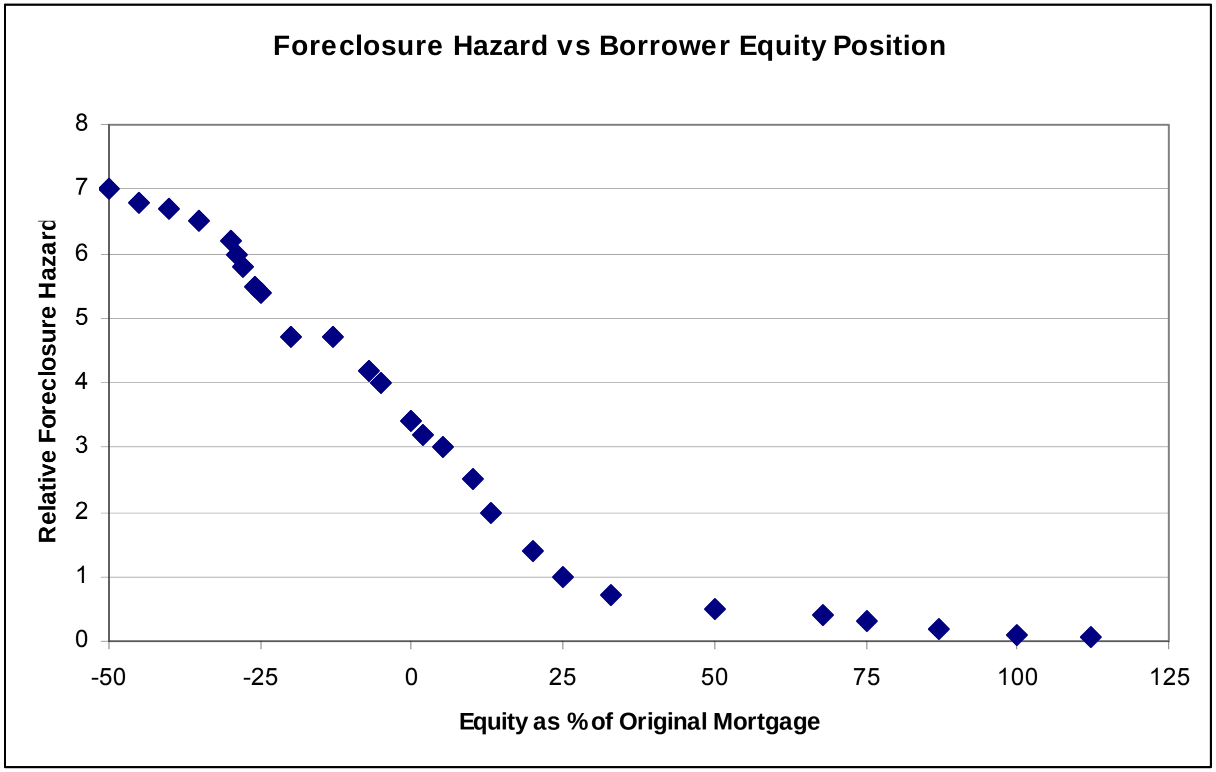
Note: All data in this presentation are for illustrative purposes only

Regression – Observe Data

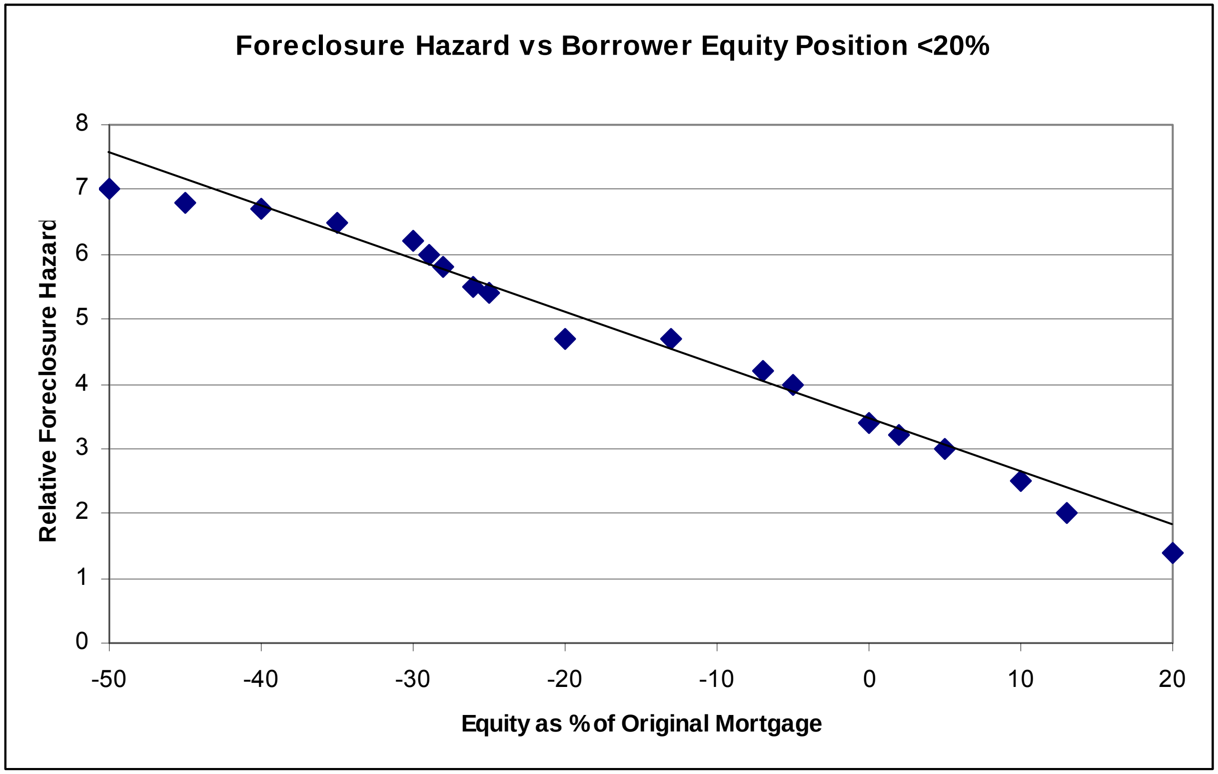


Source: Foote et al., "Negative Equity and Foreclosure: Theory and Evidence."[32]

Regression – Observe Data



Regression – Observe Data



Simple Regression

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	52.7482	52.7482	848.2740	<0.0001
Residual	17	1.0571	0.0622		
Total	18	53.8053			

- How much of the sum of squares is explained by the regression?

  SS = Sum Squared Errors

  SSTotal = SSRegression + SSResidual (Residual also called Error)

  SSTotal = $\sum (y_i - \bar{y})^2 = 53.8053$

  SSRegression = $b_{1\ est} * [\sum x_i y_i - 1/n(\sum x_i)(\sum y_i)] = 52.7482$

  SSResidual = $\sum (y_i - y_{i\ est})^2$

  = SSTotal – SSRegression

  1.0571 = 53.8053 – 52.742

Simple Regression

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	52.7482	52.7482	848.2740	<0.0001
Residual	17	1.0571	0.0622		
Total	18	53.8053			

Regression Statistics	
Multiple R	0.9901
R Square	0.9804
Adjusted R Square	0.9792

- MS = SS divided by df
- R^2: (SS Regression/SS Total)

  0.9804 = 52.7482 / 53.8053

  - percent of variance explained
- F statistic: (MS Regression/MS Residual)
- significance of regression:
  - F tests H_o: b_1=0 v. H_A: $b_1 \neq 0$

Simple Regression

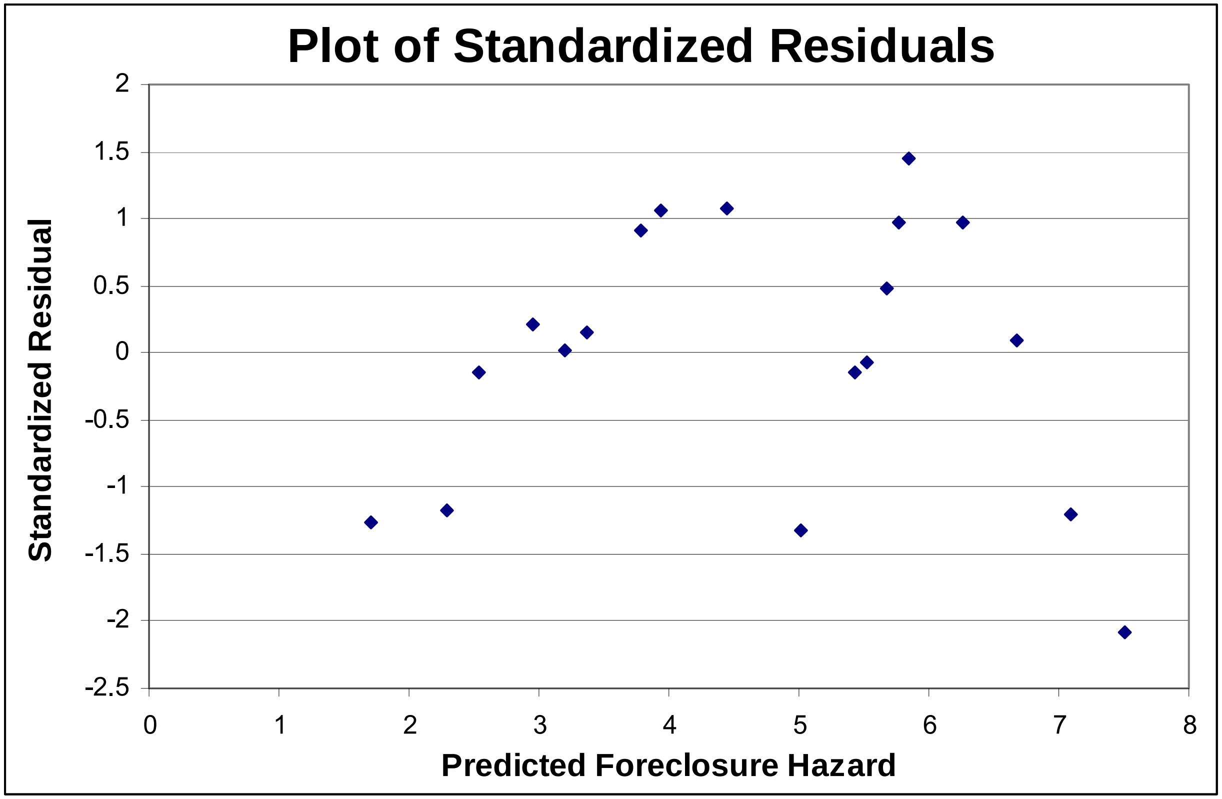
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	3.3630	0.0730	46.0615	0.0000	3.2090	3.5170	3.2090	3.5170
X	-0.0828	0.0028	-29.1251	0.0000	-0.0888	-0.0768	-0.0888	-0.0768

<u>T statistics</u>: $(b_{i\ est} - H_o(b_i))\ /\ s.e.(b_{i\ est})$

- significance of individual coefficients
- $T^2 = F$ for b_1 in simple regression
- $(-29.1251)^2 = 848.2740$
- F in multiple regression tests that at least one coefficient is nonzero. For the simple case, at least one is the same as the entire model. F stat tests the global null model.

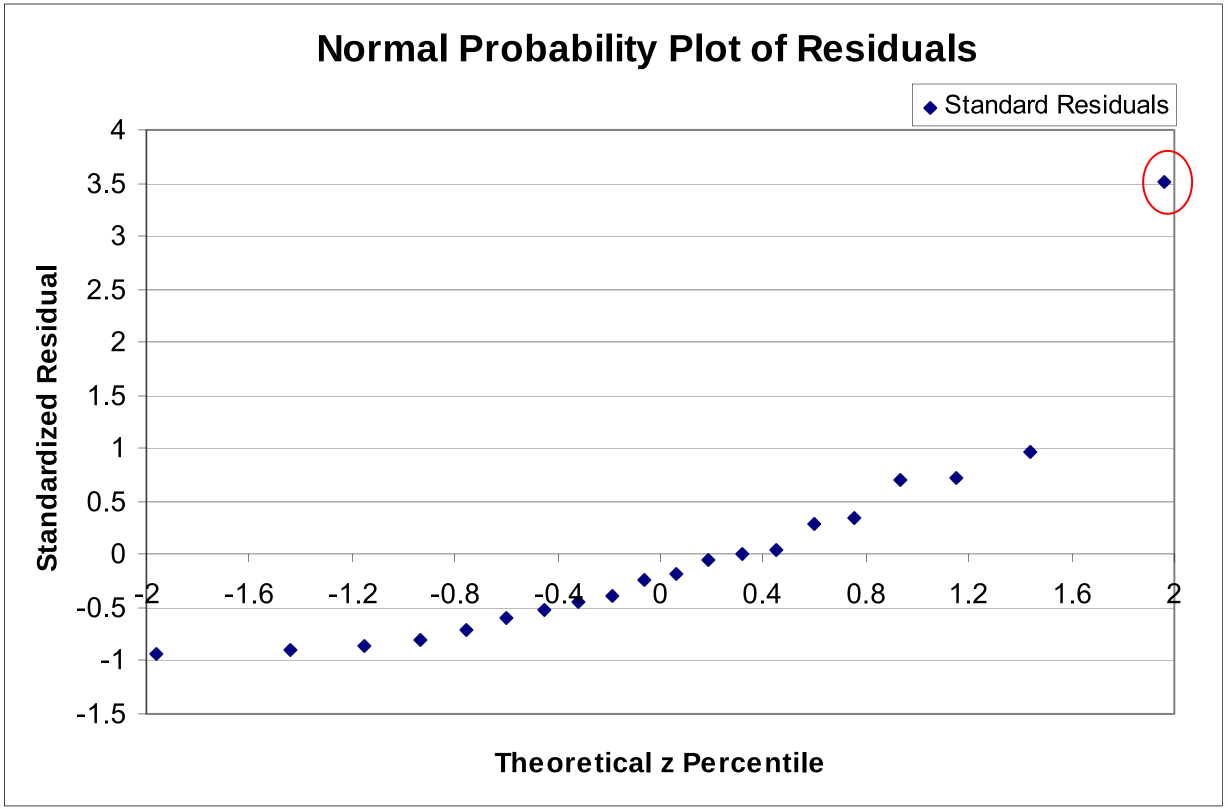
Residuals Plot

- Looks at ($y_{obs} - y_{pred}$) vs. y_{pred}
- Can assess linearity assumption, constant variance of errors, and look for outliers
- Standardized Residuals (raw residual scaled by standard error) should be random scatter around 0, standard residuals should lie between -2 and 2
- With small data sets, it can be difficult to assess assumptions



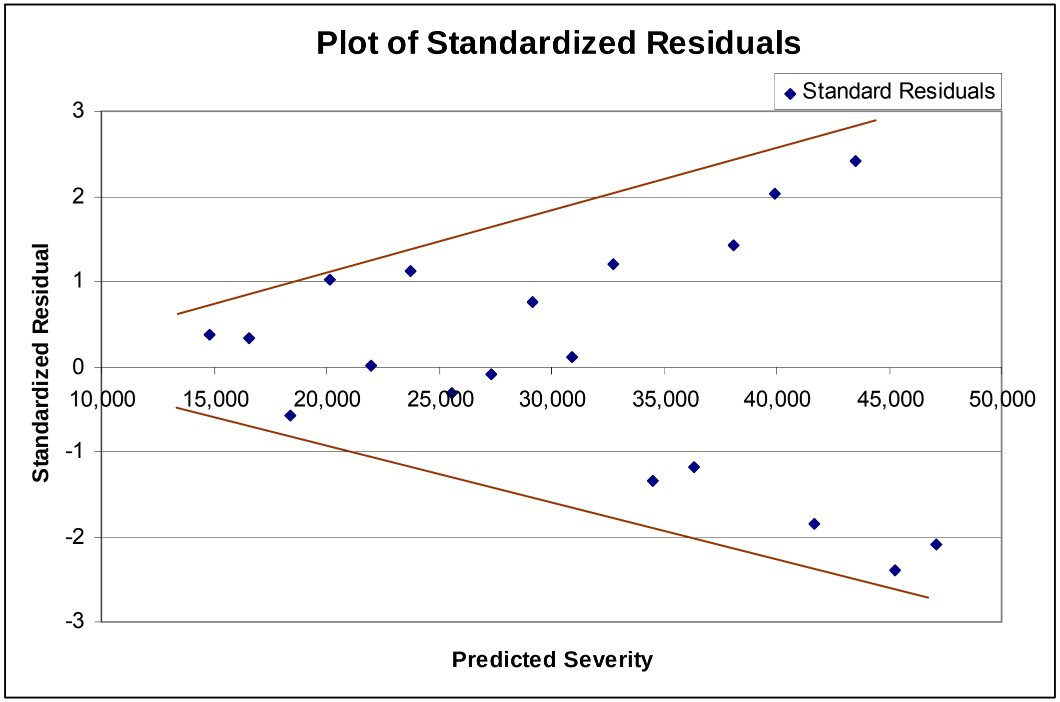
Normal Probability Plot

- Can evaluate assumption $e_i \sim N(0,\sigma_e^2)$
  - Plot should be a straight line with intercept μ and slope σ_e^2
  - Can be difficult to assess with small sample sizes



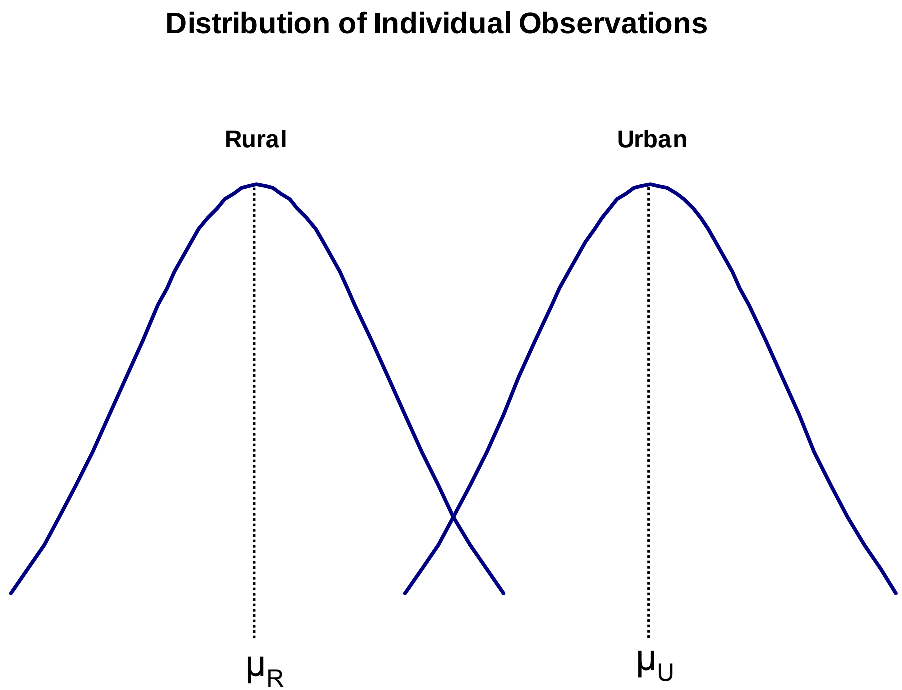
Residuals

- If absolute size of residuals increases as predicted value increases, may indicate nonconstant variance
- May indicate need to transform dependent variable
- May need to use weighted regression
- May indicate a nonlinear relationship



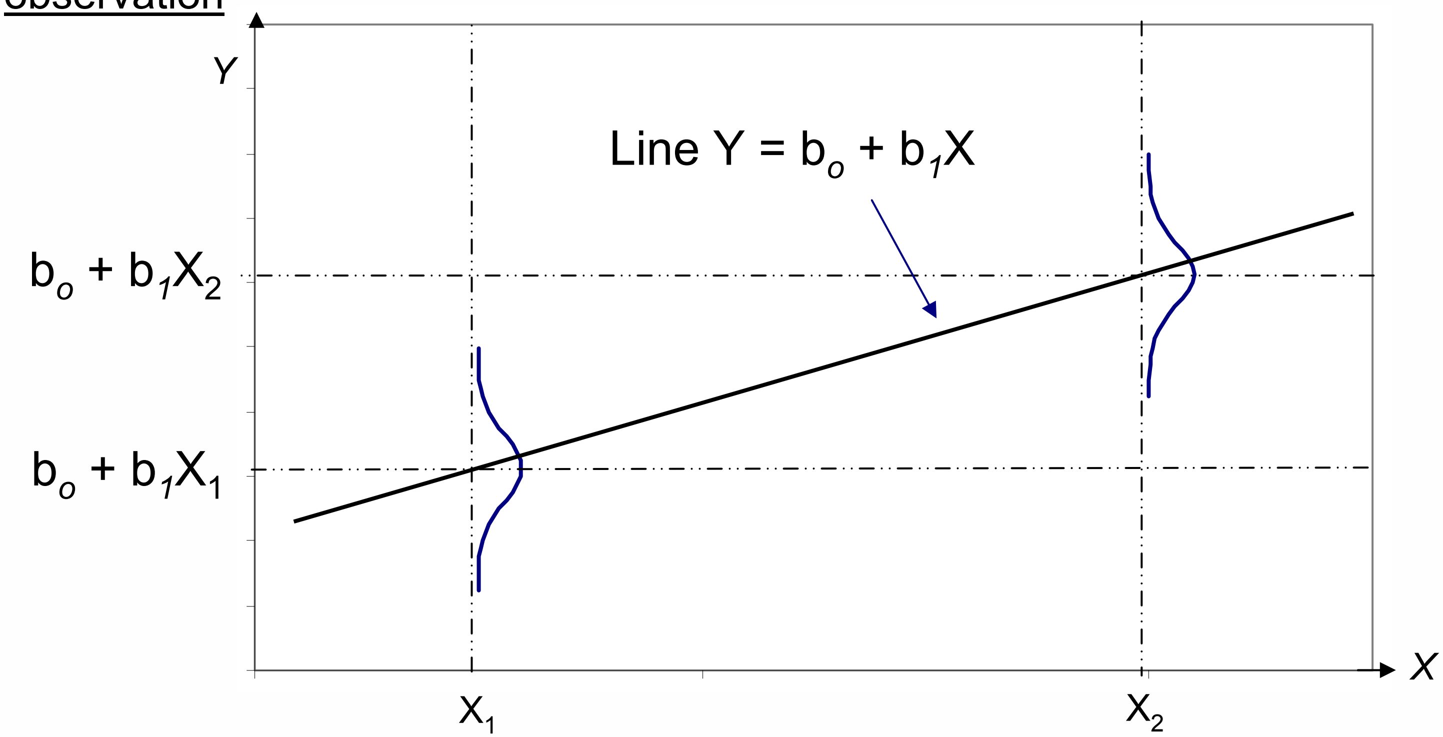
Distribution of Observations

- Average claim amounts for Rural drivers are normally distributed as are average claim amounts for Urban drivers
- Mean for Urban drivers is twice that of Rural drivers
- The variance of the observations is equal for Rural and Urban
- The total distribution of average claim amounts across Rural and Urban is not Normal
  - here it is bimodal



Distribution of Observations

- The basic form of the regression model is $Y = b_o + b_1X + e$
- $\mu_i = E[Y_i] = E[b_o + b_1X_i + e_i] = b_o + b_1X_i + E[e_i] = b_o + b_1X_i$
- The mean value of Y, rather than Y itself, is a linear function of X
- The observations Y_i are normally distributed about their mean μ_i $Y_i \sim N(\mu_i , \sigma_e^2)$
- Each Y_i can have a different mean μ_i but the variance σ_e^2 <u>is the same for each observation</u>



Multiple Regression *(special case of a GLM)*

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon$
- $E[\underline{Y}] = \underline{\beta}\,\mathbf{X}$

  $\underline{\beta}$ is a vector of the parameter coefficients

  $\underline{Y}$ is a vector of the dependent variable

  X is a matrix of the independent variables

  - Each column is a variable
  - Each row is an observation

- Same assumptions as simple regression
  1) model is correct (there exists a linear relationship)
  2) errors are independent
  3) variance of e_i constant
  4) $e_i \sim N(0,\sigma_e^2)$
- Added assumption the n variables are independent

Multiple Regression

- Uses more than one variable in regression model
  - R-sq always goes up as add variables
  - Adjusted R-Square puts models on more equal footing
  - Many variables may be insignificant
- Approaches to model building
  - Forward Selection - Add in variables, keep if “significant”
  - Backward Elimination - Start with all variables, remove if not “significant”
  - Fully Stepwise Procedures – Combination of Forward and Backward

Multiple Regression

- **Goal** : Find a simple model that explains things well with assumptions reasonably satisfied
- **Cautions**:
  - All predictor variables assumed independent
    - as add more, they may not be
    - multicollinearity— linear relationships among the X's
  - Tradeoff:
    - Increase # of parameters (1 for each variable in regression) ➔ lose degrees of freedom (df)
    - keep df as high as possible for general predictive power ➔ problem of over-fitting

Multiple Regression

- ***Model***: Claim Rate = *ƒ* (Loan-to-Value (LTV), Delinquency Status, Home Price Appreciation (HPA))
- Degrees of freedom ~ # observations - # parameters
- Any parameter with a t-stat with absolute value less than 2 is not significant

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.97
R Square	0.94
Adjusted R Square	0.94
Standard Error	0.05
Observations	586

ANOVA

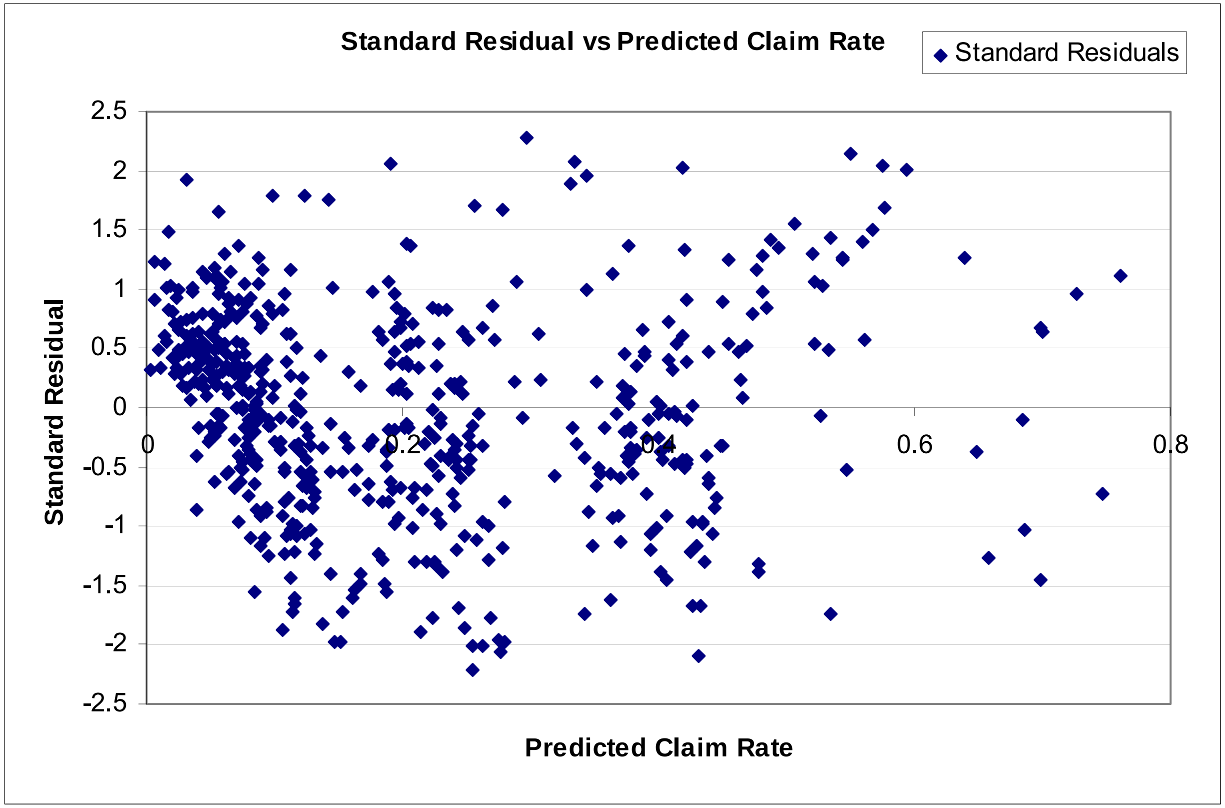
	df	*SS*	*MS*	*F*	*Significance F*
Regression	10	17.716	1.772	849.031	< 0.00001
Residual	575	1.200	0.002		
Total	585	18.916			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	1.30	0.03	41.4	0.00	1.24	1.36
ltv85	-0.10	0.01	-12.9	0.00	-0.11	-0.09
ltv90	-0.07	0.01	-9.1	0.00	-0.08	-0.06
ltv95	-0.04	0.01	-9.1	0.00	-0.05	-0.03
ltv97	-0.02	0.01	-6.0	0.00	-0.03	-0.01
ss30	-0.75	0.01	-55.3	0.00	-0.77	-0.73
ss60	-0.61	0.01	-56.0	0.00	-0.63	-0.59
ss90	-0.45	0.01	-53.5	0.00	-0.47	-0.43
ss120	-0.35	0.01	-40.1	0.00	-0.37	-0.33
ssFCL	-0.24	0.01	-22.8	0.00	-0.26	-0.22
HPA	-0.48	0.03	-18.0	0.00	-0.53	-0.43

- *T-stats are also used for evaluating significance of coefficients in GLM's*

Multiple Regression

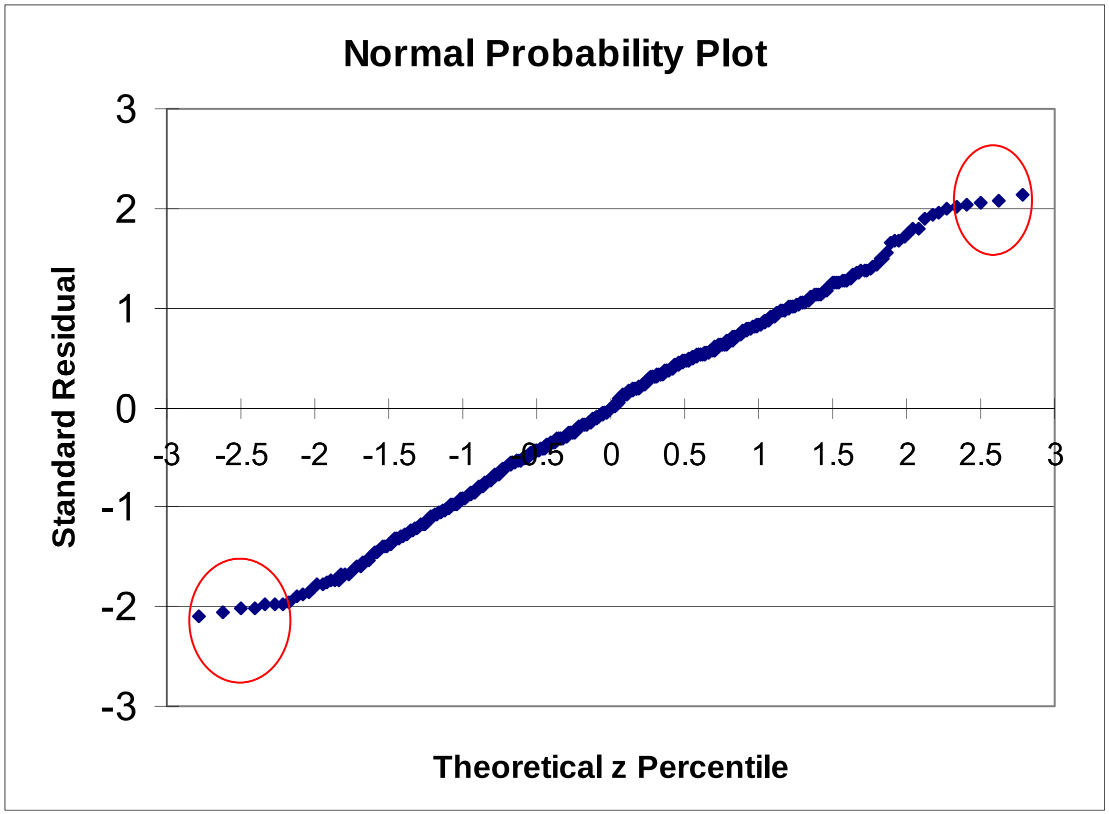
- Residuals Plot



- *Residual Plots are also used to evaluate fits of GLM's*

Multiple Regression

- Normal Probability Plot



- *Percentile or Quantile Plots are also used to evaluate fits of GLM's*

Categorical Variables *(used in LM’s and GLM’s)*

- Explanatory variables can be discrete or continuous
- Discrete variables generally referred to as “factors”
- Values each factor takes on referred to as “levels”
- Discrete variables also called Categorical variables
- In the multiple regression example given, all variables were categorical except HPA

Categorical Variables

- Assign each level a “Dummy” variable
  - A binary valued variable
  - X=1 means member of category and 0 otherwise
  - Always a reference category
    - defined by being 0 for all other levels
  - If only one factor in model, then reference level will be intercept of regression
  - If a category is not omitted, there will be linear dependency
    - “Intrinsic Aliasing”

Categorical Variables

- Example: Loan – To – Value (LTV)
  - Grouped for premium – 5 Levels
    - <=85%, LTV85
    - 85.01% - 90%, LTV90
    - 90.01% - 95%, LTV95
    - 95.01% - 97%, LTV97
    - >97% Reference
  - Generally positively correlated with claim frequency
  - Allowing each level it's own dummy variable allows for the possibility of non-monotonic relationship
  - Each modeled coefficient will be relative to reference level

		X1	X2	X3	X4
Loan #	LTV	LTV85	LTV90	LTV95	LTV97
1	97	0	0	0	1
2	93	0	0	1	0
3	95	0	0	1	0
4	85	1	0	0	0
5	100	0	0	0	0

Design Matrix

Transformations

- A possible solution to nonlinear relationship or unequal variance of errors
- Transform predictor variables, response variable, or both
- Examples:
  - $Y' = \log(Y)$
  - $X' = \log(X)$
  - $X' = 1/X$
  - $Y' = \sqrt{Y}$
- Substitute transformed variable into regression equation
- Maintain assumption that errors are $N(0,\sigma_e^2)$

Why GLM?

- What if the variance of the errors increases with predicted values?
  - More variability associated with larger claim sizes
- What if the values for the response variable are strictly positive?
  - assumption of normality violates this restriction
- If the response variable is strictly non-negative, intuitively the variance of Y tends to zero as the mean of X tends to zero
  - Variance is a function of the mean (poisson, gamma)
- What if predictor variables do not enter additively?
  - Many insurance risks tend to vary multiplicatively with rating factors

Classic Linear Model to Generalized Linear Model

- LM:
  - **X** is a matrix of the independent variables
    - Each column is a variable
    - Each row is an observation
  - $\underline{\beta}$ is a vector of parameter coefficients
  - $\underline{\varepsilon}$ is a vector of residuals

LM

$\underline{Y} = \underline{\beta}\,\mathbf{X} + \underline{\varepsilon}$

$E[\underline{Y}] = \underline{\beta}\,\mathbf{X}$

$E[\underline{Y}] = \underline{\mu} = \underline{\eta}$

$\varepsilon \sim N(0, \sigma_e^2)$

- GLM:
  - **X**, $\underline{\beta}$ same as in LM
  - $\underline{\varepsilon}$ is still vector of residuals
  - g is called the "link function"

GLM

$g(\underline{\mu}) = \underline{\eta} = \underline{\beta}\,\mathbf{X}$

$E[\underline{Y}] = \underline{\mu} = g^{-1}(\underline{\eta})$

$\underline{Y} = g^{-1}(\underline{\eta}) + \underline{\varepsilon}$

$\varepsilon \sim$ exponential family

Classic Linear Model to Generalized Linear Model

- <u>LM</u>:
  1) *Random Component* : Each component of $\underline{Y}$ is independent and normally distributed. The mean μ_i allowed to differ, but all Y_i have common variance σ_e^2
  2) *Systematic Component* : The n covariates combine to give the "linear predictor"

$$\underline{\eta} = \underline{\beta}\,\mathbf{X}$$

  3) *Link Function* : The relationship between the random and systematic components is specified via a link function. In linear model, link function is identity fnc.

$$E[\underline{Y}] = \underline{\mu} = \underline{\eta}$$

- <u>GLM</u>:
  1) *Random Component* : Each component of $\underline{Y}$ is independent and from one of the exponential family of distributions
  2) *Systematic Component* : The n covariates are combined to give the "linear predictor"

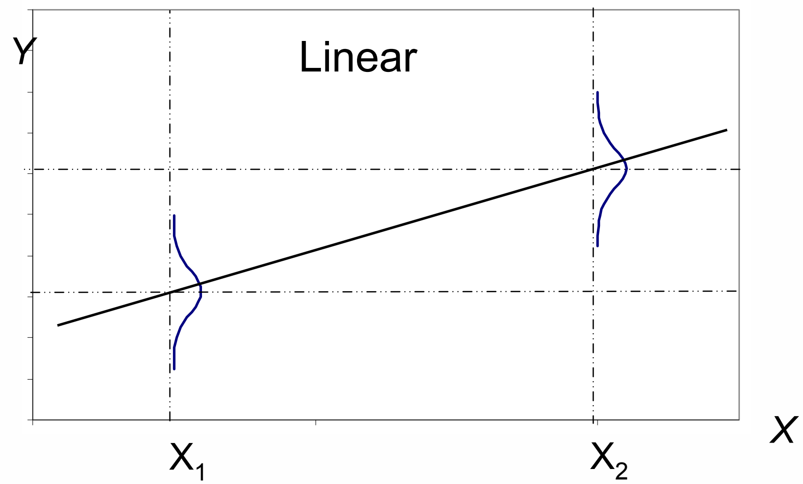
$$\underline{\eta} = \underline{\beta}\,\mathbf{X}$$

  3) *Link Function* : The relationship between the random and systematic components is specified via a link function g, that is differentiable and monotonic

$$E[\underline{Y}] = \underline{\mu} = g^{-1}(\underline{\eta})$$

Linear Transformation versus a GLM

- Linear transformation uses transformed variables
  - GLM transforms the mean
  - GLM not trying to transform Y in a way that approximates uniform variability



- The error structure
  - Linear transformation retains assumption $Y_i \sim N(\mu_i, \sigma_e^2)$
  - GLM relaxes normality
  - GLM allows for non-uniform variance
  - Variance of each observation Y_i is a function of the mean $E[Y_i] = \mu_i$

The Link Function

- Example: the log link function $g(x) = \mathit{ln}\ (x)$; $g^{-1}(x) = e^x$
- Suppose Premium (Y) is a multiplicative function of Policyholder Age (X_1) and Rating Area (X_2) with estimated parameters β_1 , β_2
  - $\eta_i = \beta_1 X_1 + \beta_2 X_2$
  - $g(\mu_i) = \eta_i$
  - $E[Y_i] = \mu_i = g^{-1}(\eta_i)$
  - $E[Y_i] = \exp(\beta_1 X_1 + \beta_2 X_2)$
  - $E[\underline{Y}] = g^{-1}(\underline{\beta}\ \mathbf{X})$

  - $E[Y_i] = \exp(\beta_1 X_1) \bullet \exp(\beta_2 X_2) = \mu_i$
  - $g(\mu_i) = \ln[\exp(\beta_1 X_1) \bullet \exp(\beta_2 X_2)] = \eta_i = \beta_1 X_1 + \beta_2 X_2$

  - The GLM here estimates logs of multiplicative effects

Examples of Link Functions

- Identity
  - $g(x) = x$ $g^{-1}(x) = x$ additive rating plan
- Reciprocal
  - $g(x) = 1/x$ $g^{-1}(x) = 1/x$
- Log
  - $g(x) = \ln(x)$ $g^{-1}(x) = e^x$ multiplicative rating plan
- Logistic
  - $g(x) = \ln(x/(1-x))$ $g^{-1}(x) = e^x/(1+ e^x)$

Error Structure

- Exponential Family
  - Distribution completely specified in terms of its mean and variance
  - The variance of Y_i is a function of its mean $E[Y_i] = \mu_i$
  - $Var\,(Y_i) = \varphi\, V(\mu_i) \,/\, \omega_i$
  - $V(\mu)$ structure specifies the *distribution* of Y, but
  - $V(\mu)$, the variance function, is not the variance of Y
  - φ is a parameter that scales the variance
  - ω_i is a constant that assigns a weight, or credibility, to observation i

Error Structure

- Members of the Exponential Family
  - Normal (Gaussian) -- used in classic regression
  - Poisson (common for frequency)
  - Binomial
  - Negative Binomial
  - Gamma (common for severity)
  - Inverse Gaussian
  - Tweedie (common for pure premium)
    - *aka* Compound Gamma-Poisson Process
      - Claim count is Poisson distributed
      - Size-of-Loss is Gamma distributed

General Examples of Error/Link Combinations

- Traditional Linear Model
  - response variable: a continuous variable
  - error distribution: normal
  - link function: identity
- Logistic Regression
  - response variable: a proportion
  - error distribution: binomial
  - link function: logit
- Poisson Regression in Log Linear Model
  - response variable: a count
  - error distribution: Poisson
  - link function: log
- Gamma Model with Log Link
  - response variable: a positive, continuous variable
  - error distribution: gamma
  - link function: log

Specific Examples of Error/Link Combinations

Observed Response	Link Fnc	Error Structure	Variance Fnc
Claim Frequency	Log	Poisson	μ
Claim Severity	Log	Gamma	μ^2
Pure Premium	Log	Tweedie	μ^p $(1<p<2)$
Retention Rate	Logit	Binomial	$\mu(1-\mu)$

References

- Anderson, D.; Feldblum, S; Modlin, C; Schirmacher, D.; Schirmacher, E.; and Thandi, N., "A Practitioner's Guide to Generalized Linear Models" (Second Edition), CAS Study Note, May 2005.
- Devore, Jay L. *Probability and Statistics for Engineering and the Sciences 3rd ed.*, Duxbury Press.
- Foote et al. 2008. Negative equity and foreclosure: Theory and evidence. Journal of Urban Economics. 64(2):234-245.
- McCullagh, P. and J.A. Nelder. *Generalized Linear Models,* 2nd Ed., Chapman & Hall/CRC
- SAS Institute, Inc. SAS Help and Documentation v 9.1.3