GLM I
An Introduction to Generalized Linear Models

CAS Ratemaking and Product Management Seminar
March 2009

Presented by: Tanya D. Havlicek, Actuarial Assistant

Milliman

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.

Milliman

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

Milliman

Modeling Schematic

Independent vars/ Predictors

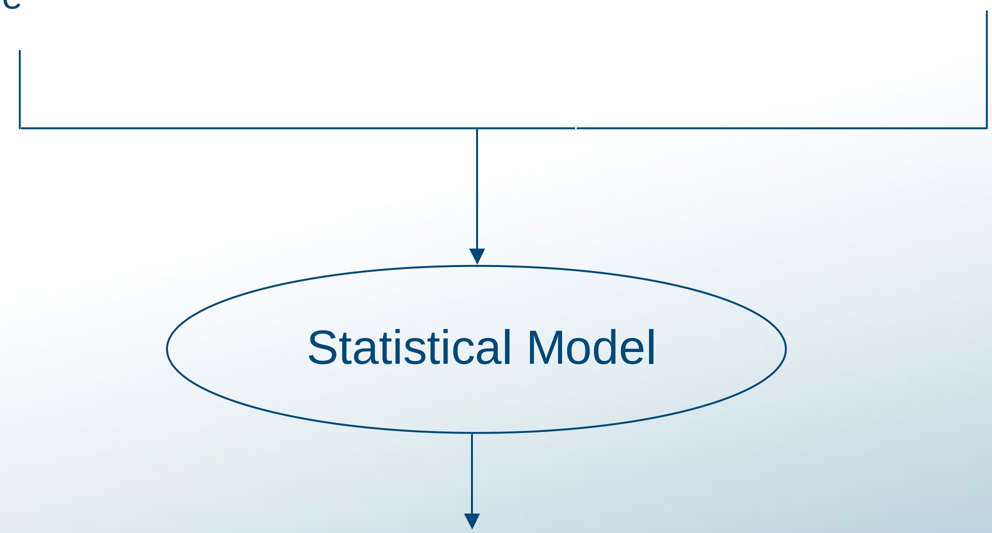
- Loan Age
- Region
- Loan-to-Value (LTV)
- Credit Score

Weights

- Claims
- Exposures
- Premium

Dependent var/ Response

- Losses
- Claims
- Persistency



Model Results

- Parameters
- Validation Statistics

General Steps in Modeling

Goal: Explain how a variable of interest depends on some other variable(s). Once the relationship (i.e., a model) between the dependent and independent variables is established, one can make predictions about the dependent variable from the independent variables.

1. Collect/build potential models and data with which to test models
2. Parameterize models from observed data
3. Evaluate if observed data follow or violate model assumptions
4. Evaluate model fit using appropriate statistical tests
   - Explanatory or predictive power
   - Significance of parameters associated with independent variables
5. Modify model
6. Repeat

Basic Linear Model Structures

- ANOVA : $Y_{ij} = \mu + \psi_i + e_{ij}$
  - Assumptions: errors are independent and follow $N(0,\sigma_e^2)$ – Normal distribution with mean of zero and constant variance σ_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)
- Simple Linear Regression : $y_i = b_o + b_1 x_i + e_i$
  - Assumptions:
    - linear relationship
    - errors are 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 as simple regression, but with n independent random variables (RV's)
- Transformed Regression : transform x, y, or both; maintain assumption that errors are $N(0,\sigma_e^2)$

$$y_i = \exp(x_i)$$

$$\log(y_i) = x_i$$

Milliman

One-way ANOVA

$Y_{ij} = \mu + \psi_i + e_{ij}$

Y_{ij} is the j^{th} observation on the i^{th} treatment

$j = 1, \ldots, n_i$

$i = 1, \ldots, k$ treatments or levels

μ is the common effect for the whole experiment

ψ_i is the i^{th} treatment effect

e_{ij} is random error associated with observation Y_{ij} , $e_{ij} \sim N(0, \sigma_e^2)$

- ANOVAs can be used to test whether observations come from different populations or from the same population

  "Is there a statistically significant difference between two groups of claims?"

  - Is the frequency of default on subprime loans different than that for prime loans?
  - Personal Auto: Is claim severity different for Urban vs Rural locations?

Milliman

One-way ANOVA

$Y_{ij} = \mu + \psi_i + e_{ij}$

- Assumptions of ANOVA model
  - independent observations
  - equal population variances
  - Normally distributed errors with mean of 0
  - Balanced sample sizes (equal # of observations in each group)
- Prediction:
  - (observation) Y = common mean + treatment effect.
  - Null hypothesis is no treatment effect. Can use contrasts or categorical regression to investigate treatment effects.

One-way ANOVA

- Potential Assumption violations:
  - Implicit factors: lack of independence within sample (e.g., serial correlation)
  - Lack of independence between samples (e.g., samples over time on same subject)
  - Outliers: apparent non-normality by a few data points
  - Unequal population variances
  - Unbalanced sample sizes
- How to assess:
  - Evaluate “experimental design” -- how was data generated? (independence)
  - Graphical plots (outliers, normality)
  - Equality of variances test (Levene’s test)

Milliman

Simple Regression

- Model: $Y_i = b_o + b_1 X_i + e_i$
  - Y is the dependent variable explained by X, the independent variable
    - Y: mortgage claim frequency depends on X: Seriousness of delinquency
    - Y: claim severity depends on X: Accident year
  - Want to estimate 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

- Model: $Y_i = b_o + b_1 X_i + e_i$
  - Assumptions:
    1) model is correct (there exists a linear relationship)
    2) errors are independent
    3) variance of e_i constant
    4) $e_i \sim N(0,\sigma_e^2)$

    In terms of robustness, 1) is most important, 4) is least important

  Parameterize:

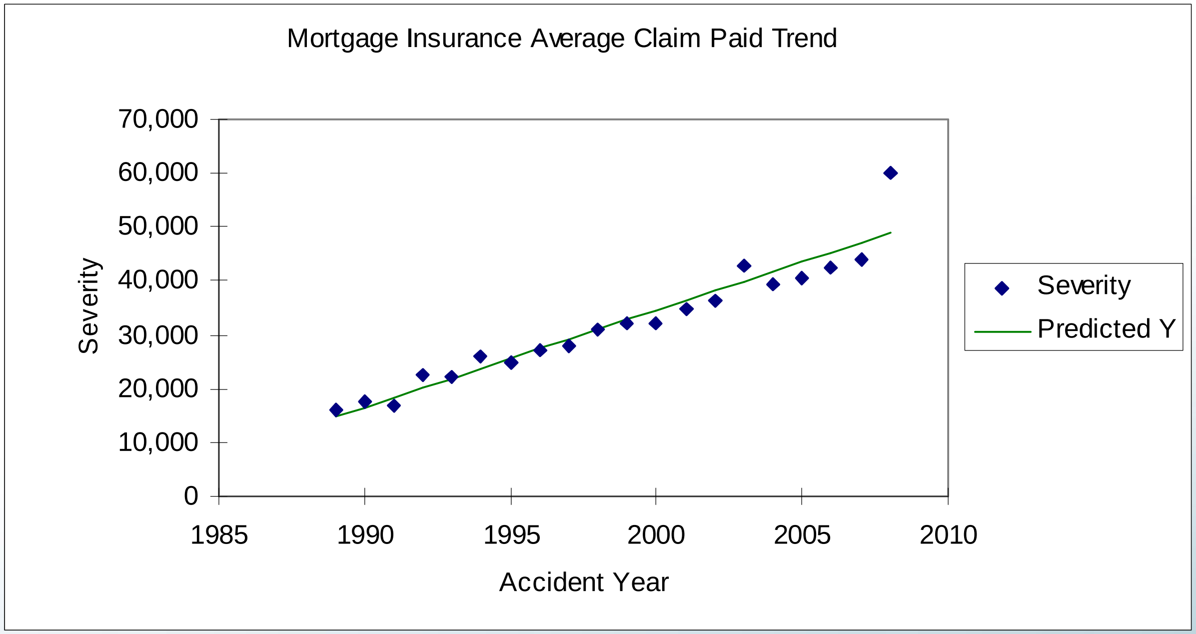
  Fit b_o and b_1 using Least Squares:

  minimize: $\sum [y_i - (b_o + b_1 x_i)]^2$

Milliman

Simple Regression

- The method of least squares is a formalization of best fitting a line through data with a ruler and a pencil
- Based on a correlative relationship between the independent and dependent variables



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

Slope Intercept

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

Milliman

Simple Regression

ANOVA		
	df	*SS*
Regression	1	2,139,093,999
Residual	18	191,480,781
Total	19	2,330,574,780

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

  SS = Sum Squared Errors

  SSTotal = SSRegression + SSResidual (Residual also called Error)

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

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

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

  = SSTotal – SSRegression

Simple Regression

ANOVA

	df	SS	MS	F	Significance F
Regression	1	2,139,093,999	2,139,093,999	201.0838	0.0000
Residual	18	191,480,781	10,637,821		
Total	19	2,330,574,780			

MS = SS divided by degrees of freedom

R^2: (SS Regression/SS Total)

- percentage of variance explained by linear relationship

F statistic: (MS Regression/MS Residual)

- significance of regression:
  - tests H_o: $b_1=0$ v. H_A: $b_1 \neq 0$

Milliman

Simple Regression

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-3,552,486.3	252,767.6	-14.054	0.0000	-4,083,531	-3,021,441
Accident Year	1,793.5	126.5	14.180	0.0000	1,528	2,059

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

- significance of coefficients
- $T^2 = F$ for b_1 in simple regression

Milliman

Simple Regression

- p-values test the null hypotheses that the parameters b_1 =0 or b_o = 0. If b_1 is 0, then there is no linear relationship between the independent variable Y (severity) and the dependent variable X (accident year). If b_o is 0, then the intercept is 0.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.958
R Square	0.918
Adjusted R Square	0.913
Standard Error	3261.57
Observations	20

- 92% of the variance is explained by the regression
- The probability of observing this data given that b_1 =0 is <0.00001, the significance of F
- Both parameters are significant
- $F = T^2$ for X Variable 1 $201.08 = (14.1804)^2$

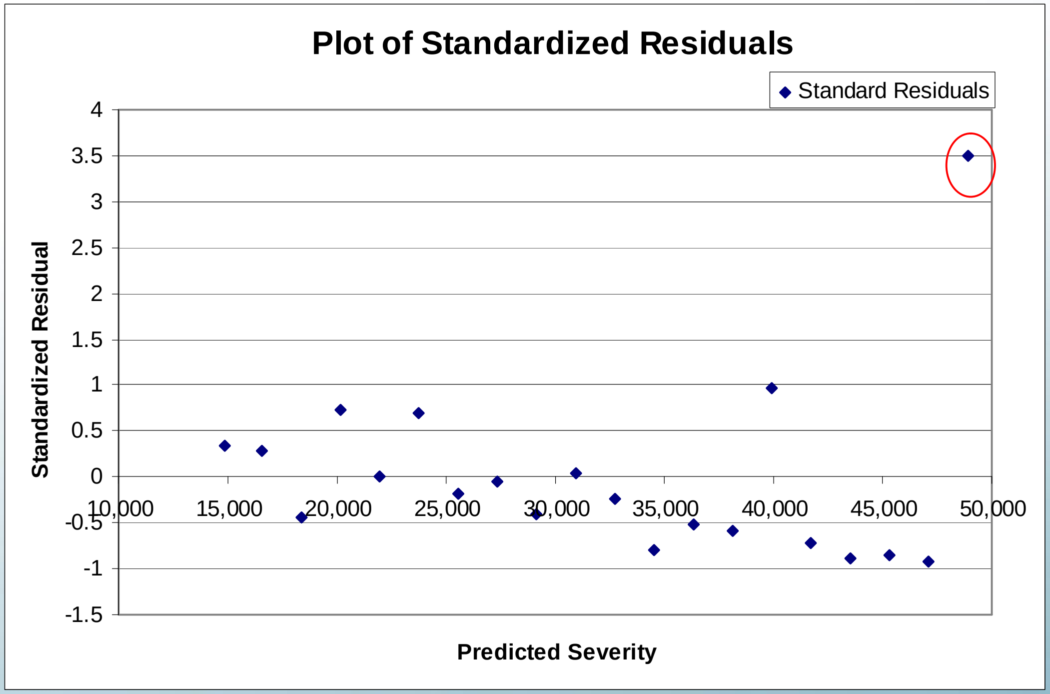
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	2,139,093,999	2,139,093,999	201.0838	0.0000
Residual	18	191,480,781	10,637,821		
Total	19	2,330,574,780			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-3,552,486.3	252,767.6	-14.054	0.0000	-4,083,531	-3,021,441
Accident Year	1,793.5	126.5	14.180	0.0000	1,528	2,059

Residuals Plot

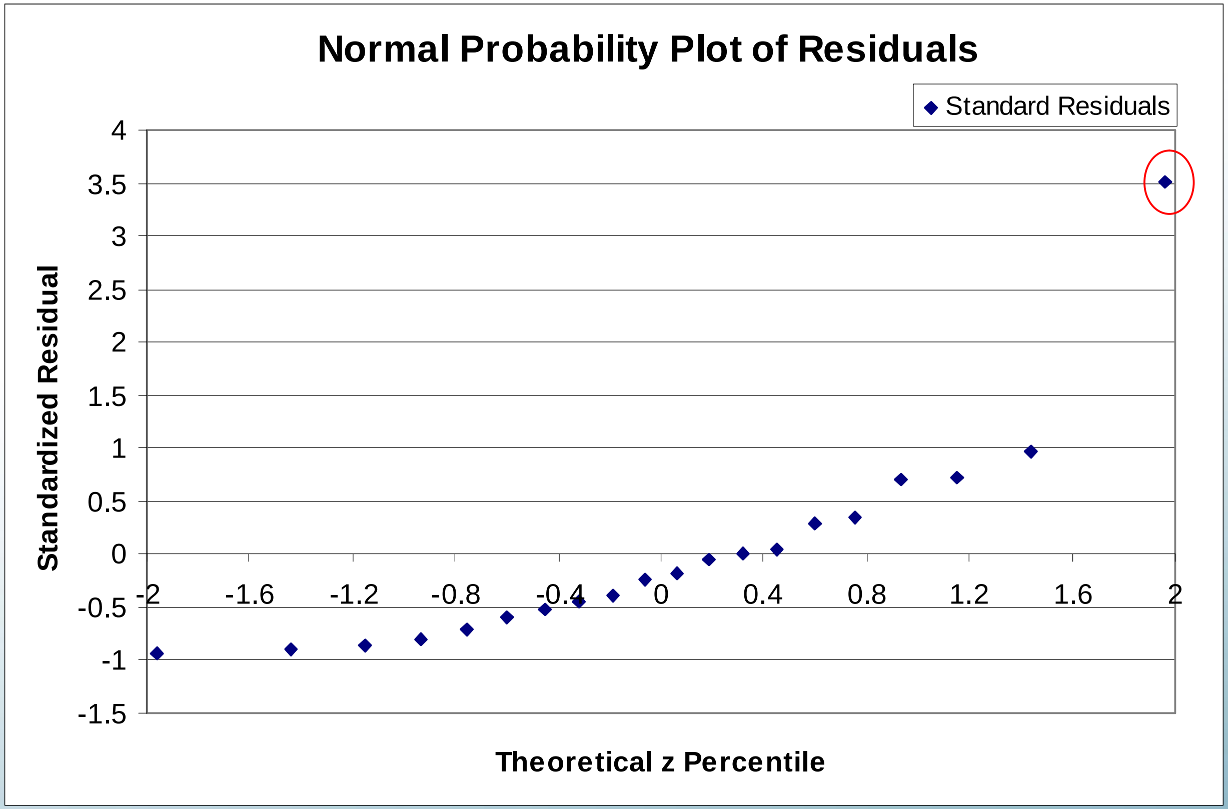
- Looks at $(y_{obs} - y_{pred})$ vs. y_{pred}
- Can assess linearity assumption, constant variance of errors, and look for outliers
- Residuals should be random scatter around 0, standard residuals should lie between -2 and 2
- With small data sets, it can be difficult to asess



Milliman

Normal Probability Plot

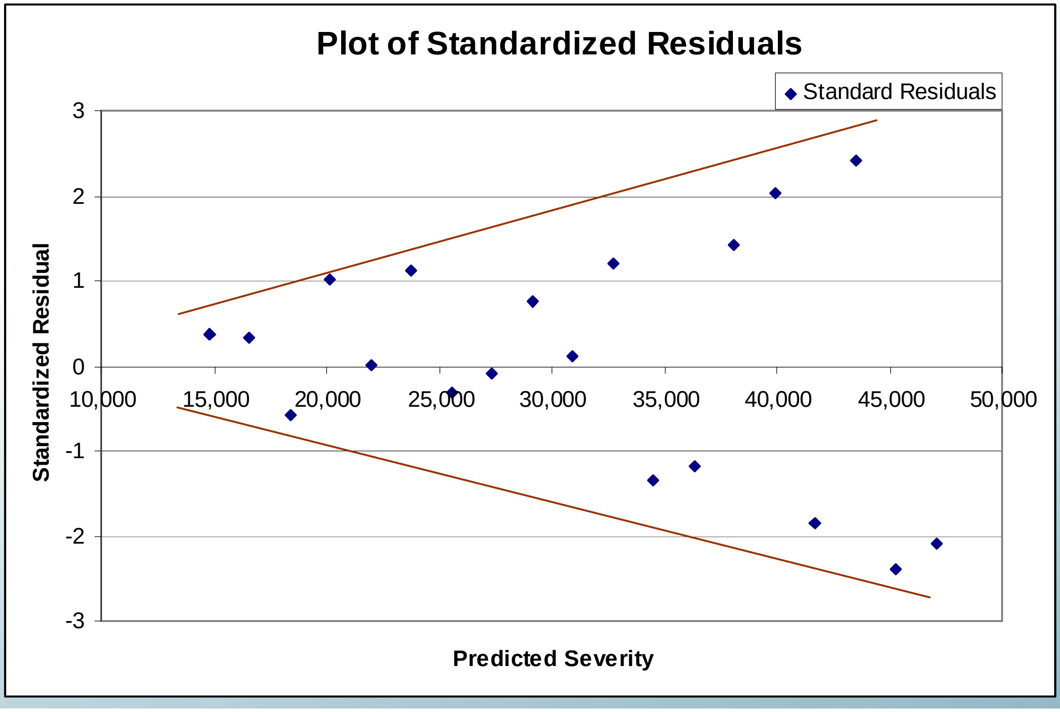
- Can evaluate assumption $e_i \sim N(0,\sigma_e^2)$
  - Plot should be a straight line with intercept μ and slope σ_e^2
  - Can be difficult to assess with small sample sizes



Milliman

Residuals

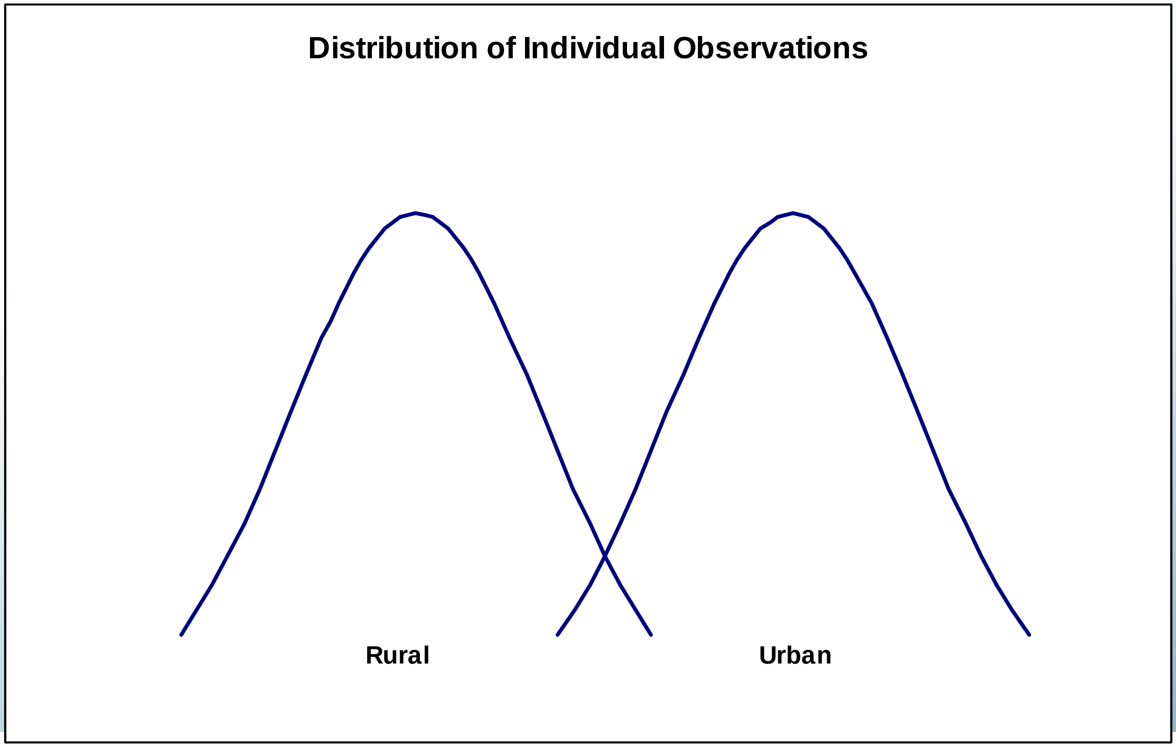
- If absolute size of residuals increases as predicted value increases, may indicate nonconstant variance
- May indicate need to transform dependent variable
- May need to use weighted regression
- May indicate a nonlinear relationship



Milliman

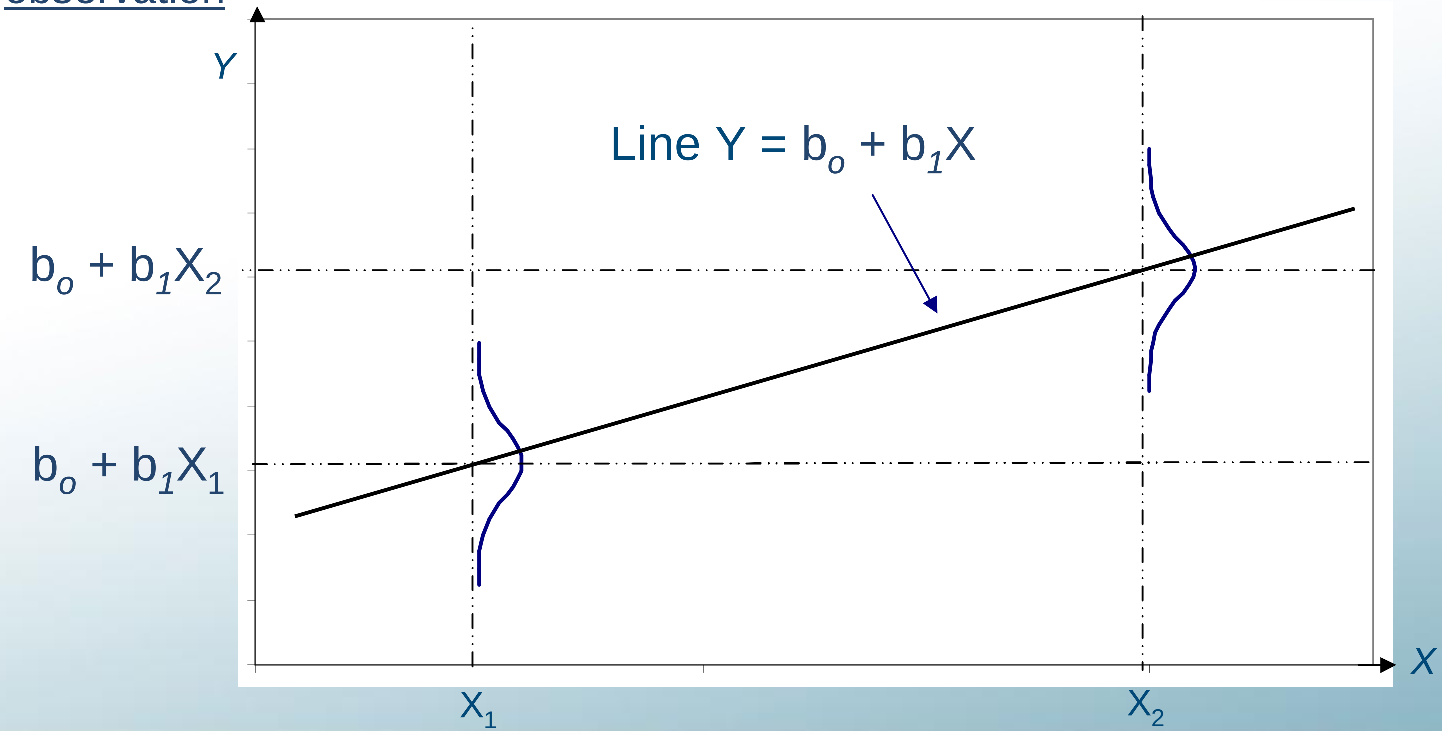
Distribution of Observations

- Average claim amounts for Rural drivers is 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 is not Normally distributed
  - 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>



Milliman

Predicting with regression

- MSResiduals (also called s_e^2) is an estimate of σ_e^2, the variability of the errors
- Estimated Y has a lower standard error than Predicted Y, but both have the same point estimate - μ_i
  - $Y_{pred} = b_o + b_1 x^*$ for some new x*
  - $Y_{est} = b_o + b_1 x^*$ for some new x*
- standard error for both use s_e in formula
- Y_{pred} standard error accounts for random variability around the line in addition to the uncertainty of the line
- Typically give a confidence interval around the point estimate (e.g. 95%)
- $(Y_{pred} \pm se(Y_{pred}) * T_{0.025,\ DF})$
- Predictions should only be made within the range or slightly outside of observed data. Extrapolation can lead to erroneous predictions

Milliman

Multiple Regression

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

- 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 met
  - Model assumes all predictor variables independent of one another– as add more, they may not be (multicollinearity—strong linear relationships among the X's)
  - As you increase the number of parameters (one for each variable in regression) you lose degrees of freedom
    - want to keep df as high as possible for general predictive power
    - problem of over-fitting

Multiple Regression

- Multicollinearity arises when there are strong linear relationships among the x's
- May see:
  - High pairwise correlations amongst the x's
  - Large changes in coefficients when another variable added or deleted
  - Large change in coefficients when data point added or deleted
  - Large standard deviations of the coefficients

- Some solutions to combat overfitting and multicollinearity
  - Stepwise Regression (Forwards, Backwards, Exhaustive) -- Order matters
  - Drop one or more highly correlated variables
  - Use Factor Analysis or Principle Components Analysis to combine correlated variables into a smaller number of new uncorrelated variables

Milliman

Multiple Regression

- F significant and Adj R-sq high
- 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

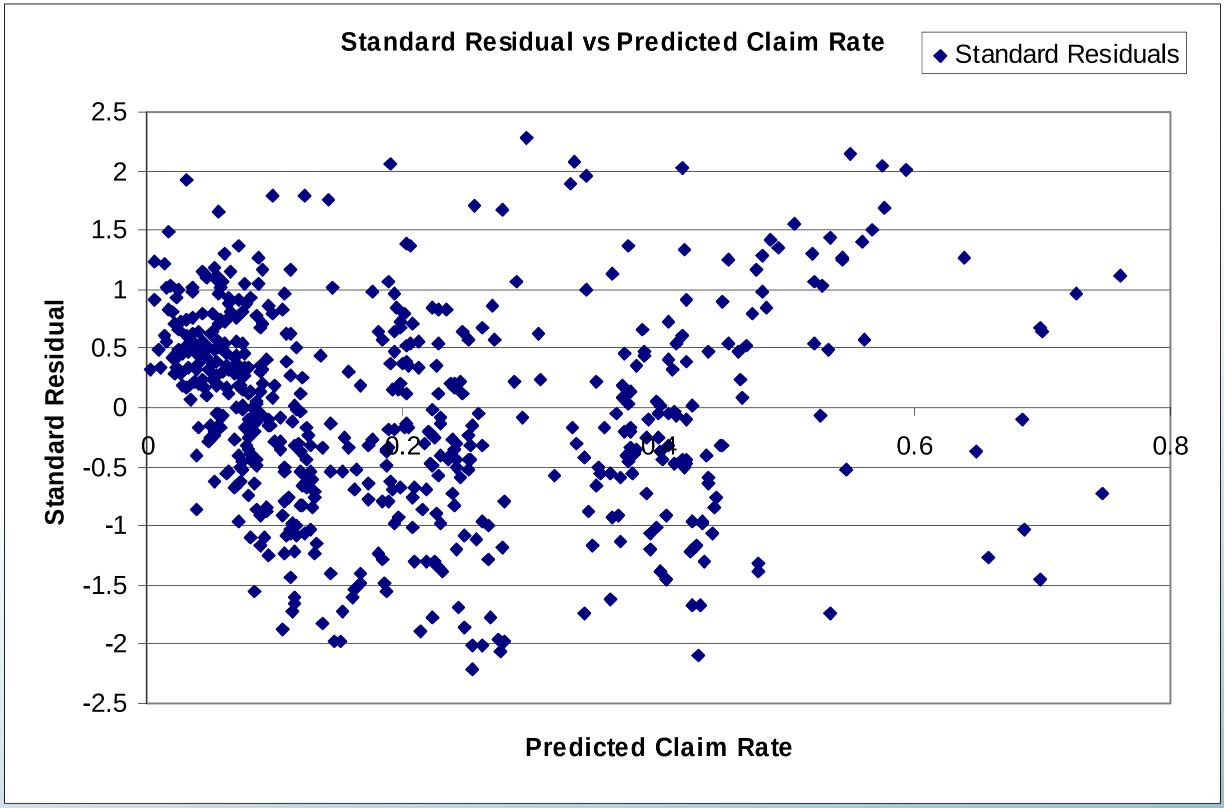
ANOVA

	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

Multiple Regression

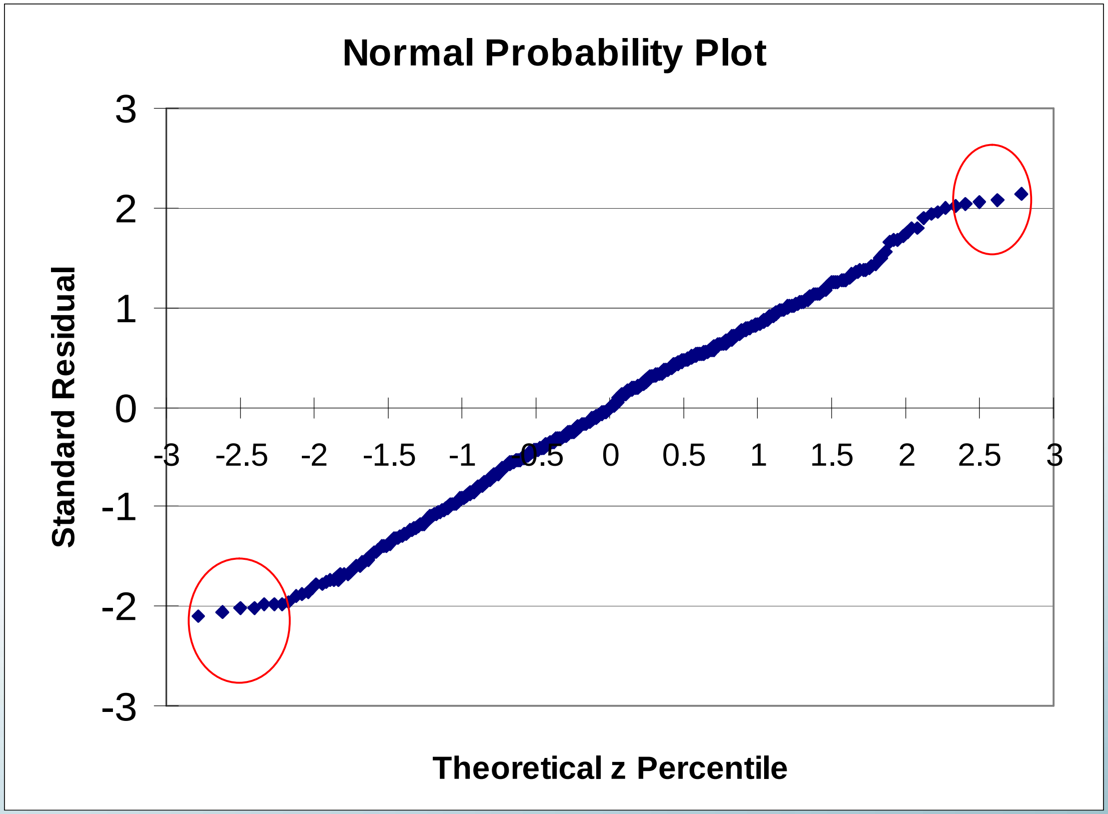
- Residuals Plot



Milliman

Multiple Regression

- Normal Probability Plot



Categorical Variables

- 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 discrete except HPA (embedded home price appreciation)

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 postively correlated with claim frequency
  - Allowing each level it's own dummy variable allows for the possiblity 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

Milliman

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 $N(0,\sigma_e^2)$

Milliman

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
- 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

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

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

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

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

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

$\varepsilon \sim$ exponential family

Milliman

Classic Linear Model to Generalized Linear Model

- LM:
  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.

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

- GLM:
  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"

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

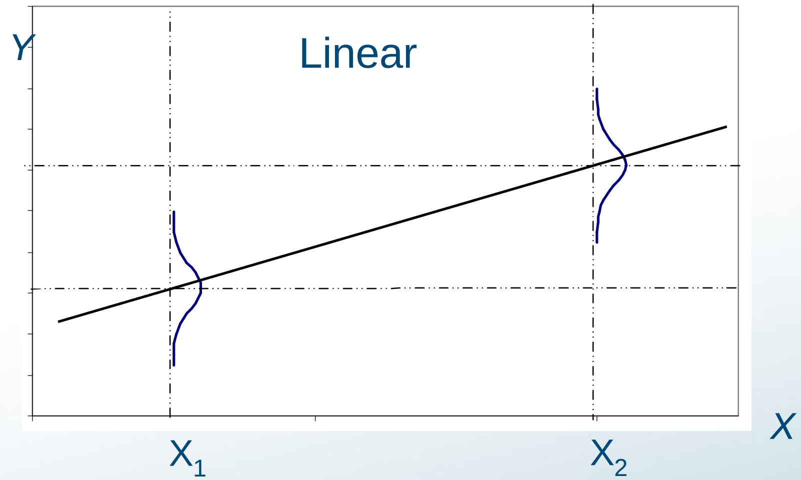
  3) *Link Function* : The relationship between the random and systematic components is specified via a link function g, that is differentiable and monotonic

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

Milliman

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 $g(x) = ln\ (x)$; $g^{-1}(x) = e^{x}$
- Suppose Premium (Y) is an 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

Milliman

Examples of Link Functions

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

Milliman

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
- 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)

Milliman

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

Milliman

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.
- 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

Milliman