Introduction to Predictive Modeling Using GLMs

Dan Tevet, FCAS, MAAA, Liberty Mutual Insurance Group

Anand Khare, FCAS, MAAA, CPCU, Milliman

100 Years of Expertise, Insight & Solutions

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.**

100 Years of Expertise, Insight & Solutions

A New England vacation

- The weather
- The car insurance

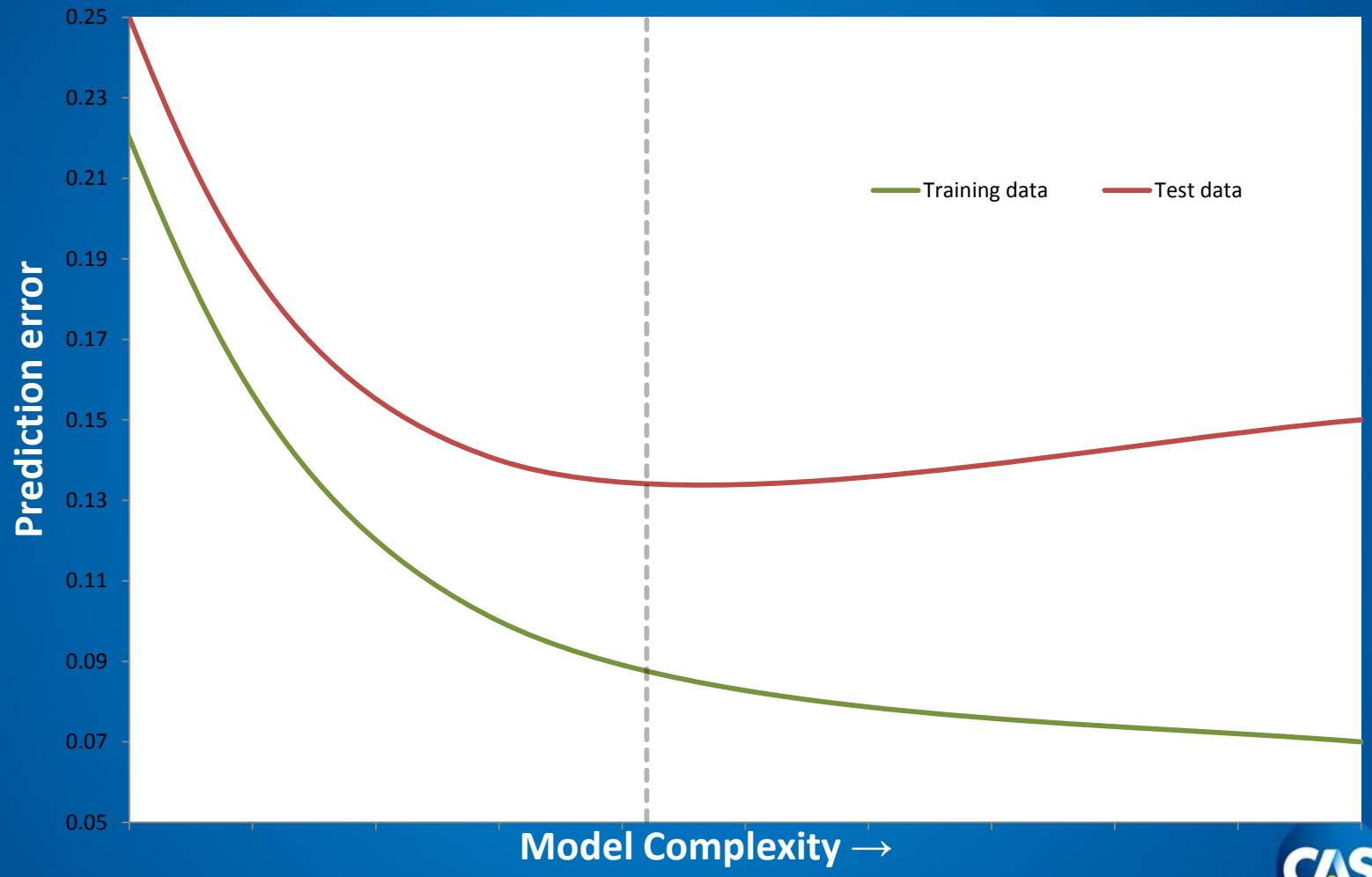
A few words about overfitting -

The more data we have…
The more models we have…
The more sophisticated models become…

The potential for overfitting data never goes away!

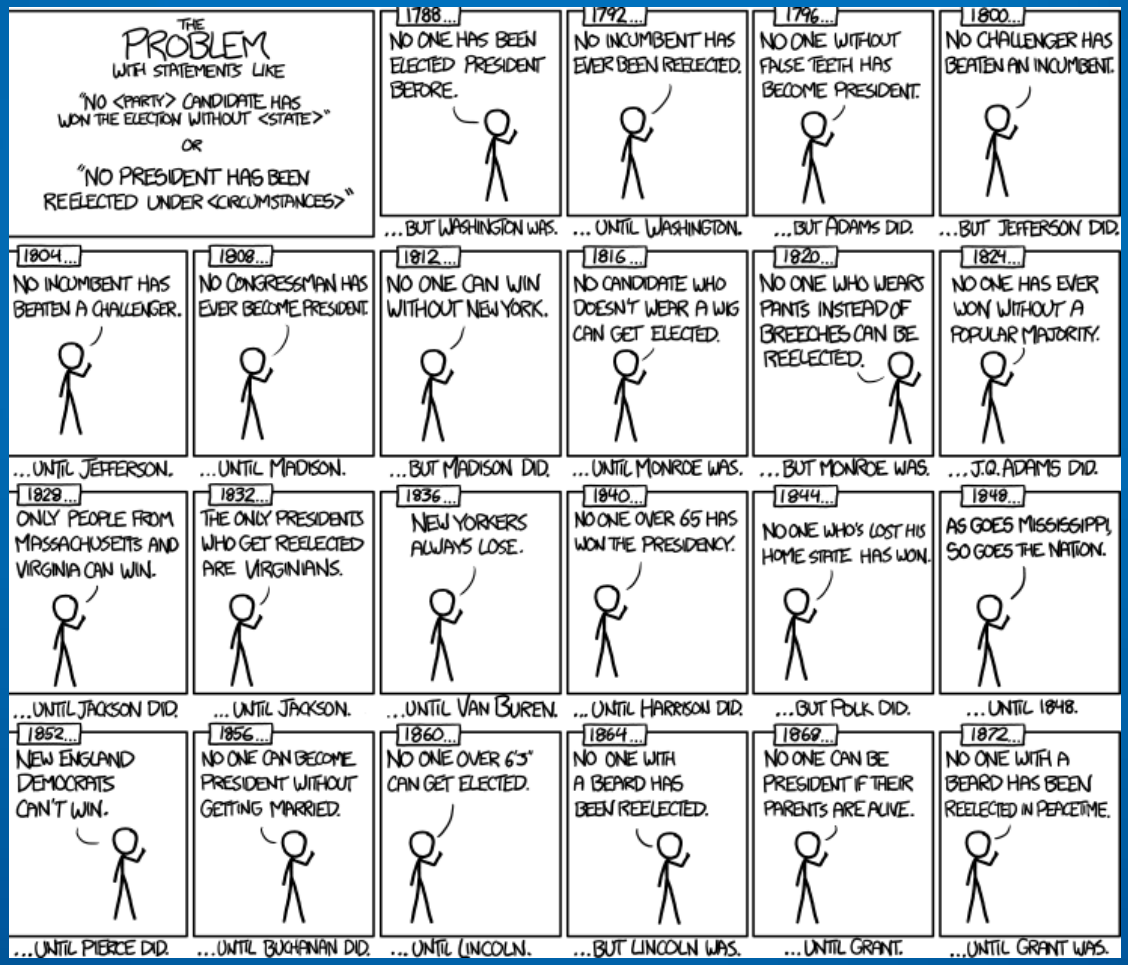
That's one big reason to validate models.

In a nutshell



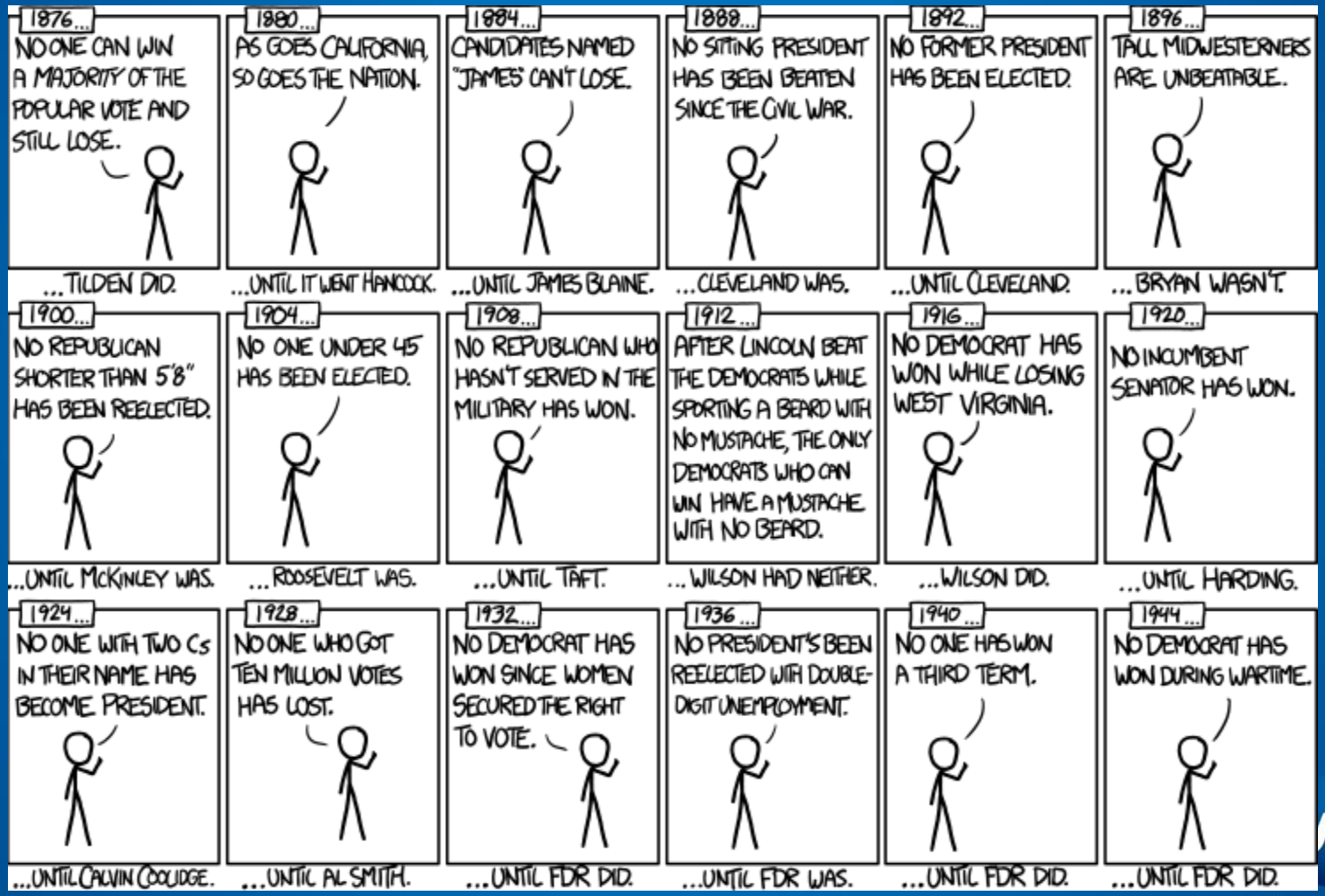
100 Years of Expertise, Insight & Solutions

Overfitting U.S. Elections (courtesy of Randall Munroe, writer of “xkcd” and What If)



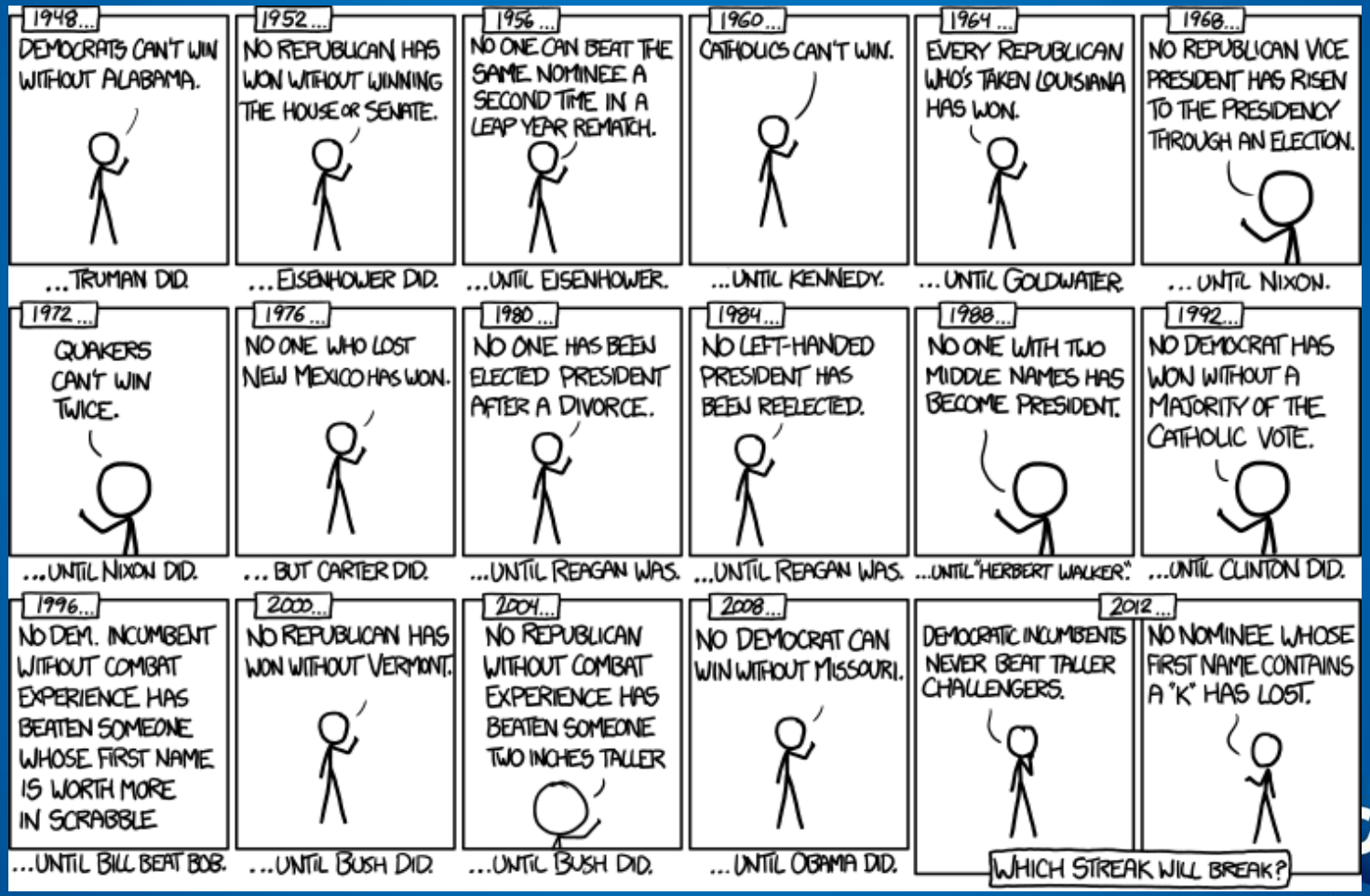
100 Years of Expertise, Insight & Solutions

1876-1944



100 Years of Expertise, Insight & Solutions

1948-Present



100 Years of Expertise, Insight & Solutions

Outline

- Overview of predictive modeling
- Predictive modeling in the actuarial world
- Simple linear models vs generalized linear models (GLMs)
- Specification of GLMs
- Interpretation of GLM output
- Frequency/severity vs pure premium modeling
- Model validation
- Modeling process and important considerations
- The next 100 years

100 Years of Expertise, Insight & Solutions

What is Predictive Modeling

- Model – an abstraction of reality, generally with a random or probabilistic component
  - Simplification of a real world phenomenon
- Model types include:
  - Linear models – predict target variable using linear combination of predictor variables
  - Trees – split dataset, one variable at a time, into subgroups that behave similarly
  - Neural networks – “self-learning” algorithms that adapt to best predict a quantity of interest

How Do Actuaries Use Modeling?

- Rating plans – model insurance loss data to build plans that charge actuarially fair rates
- Underwriting plans – knowing relative riskiness of policyholder can inform underwriting decisions
- Enterprise risk management – model correlations between lines of business or probability of ruin
- Product analytics – customer retention, conversion, lifetime value

Simple Linear Model

- $Y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \ldots + \varepsilon$
  - Y is the *target* or *response* variable – it is what we are trying to predict (e.g. pure premium)
  - X_1, X_2, etc are the *explanatory* (e.g. age of driver, type of vehicle) variables – we use them to predict Y
  - ε is the *error* or *noise* term – it is the portion of Y that is unexplained by X
- $\mu = E(Y) = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \ldots + \beta_n * X_n$
- In general, we are modeling the mean of Y

Simple Linear Model Assumptions

- Assumptions of simple linear models
  - Target variable Y does not depend on the value of Y for any other record, only the predictors
  - Y is normally distributed
  - Mean of Y depends on the predictors, but all records have same variance
  - Y is related to predictors through simple linear function
- Unfortunately, these assumptions are often unrealistic
  - Target variables of interest, such as pure premium, frequency, and severity, are not normally distributed and have non-constant variance

Generalized Linear Models

- Generalized linear model: $g(\mu) = \beta_0 + \beta_1*X_1 + \beta_2*X_2 + \ldots + \beta_n*X_n$
- Assumptions of generalized linear models
  - Target variable Y does not depend on the value of Y for any other record, only the predictors
  - Distribution of Y is a member of the exponential family of distributions
  - Variance of Y is a function of the mean of Y
  - $g(\mu)$ is linearly related to the predictors. The function g is called the link function
- The exponential family of distributions include the following: Normal, Poisson, Gamma, Binomial, Negative Binomial, Inverse Gaussian, Tweedie

GLM Variance Function

- $Var(Y) = \varphi * V(\mu)/w$
- φ is the dispersion coefficient, which is estimated by the GLM
- w is the weight assigned to each record
  - GLMs calculate the coefficients that maximize likelihood, and w is the weight that each record gets in that calculation
- $V(\mu)$ is the GLM Variance Function, and is determined by the distribution
  - Normal: $V(\mu) = 1$
  - Poisson: $V(\mu) = \mu$
  - Gamma: $V(\mu) = \mu^2$

GLM Link Function

- $g(\mu) = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \ldots$
- Common choices for link function
  - Identity: $g(\mu) = \mu$
  - Log: $g(\mu) = \ln(\mu)$
  - Logit: $g(\mu) = \ln(\mu/(1- \mu))$
- Log link commonly used to model rating plans because it produces multiplicative relativities
  - $\ln(\mu) = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2$
    $\rightarrow \mu = \exp(\beta_0 + \beta_1 * X_1 + \beta_2 * X_2)$
    $\rightarrow \mu = \exp(\beta_0) * \exp(\beta_1 * X_1) * \exp(\beta_2 * X_2)$
- Logit link used to model probability of an event occurring

Offsets

- An effect in a model that is fixed by the modeler
- Variable offsets – fix the effect of variables that are not being modeled
  - Example: Constructing a rating plan and not modeling base territory rates
  - Solution: Offset for current territory rates
- Volume offsets – reflect fact that different records have different volumes of data and thus have different expected values
  - Example: Modeling claim counts. Some records have a single exposure, other have many exposures
  - Solution: Offset for exposure volume of each observation

100 Years of Expertise, Insight & Solutions

Interpreting GLM Coefficients w/ Log Link

- Discrete variables: exponentiate GLM coefficient
  - Example: coefficient for youth drivers is 0.52
    - → Rating factor = exp(0.52) = 1.68
    - → Youth drivers have 68% surcharge relative to base level of adult drivers (who, by definition, have rating factor of 1.00)
- Continuous variables with no transformation
  - Example: modeling pure premium, and annual miles driven is a continuous variable
  - As miles driven increases by 1 unit, expected pure premium is scaled by a factor of exp(β), regardless of whether mileage goes from 1,000 to 2,000 or 20,000 to 21,000
- Continuous variables with log transformation
  - Pure premium ~ (annual mileage)^β
  - If β < 1, then as mileage increases, pure premium increases at decreasing rate

Uncertainty in Parameter Estimates

- GLMs allow us to quantify uncertainty in parameter estimates
- Wald 95% confidence interval for mean of parameter estimate = Mean +/- 1.96*(Standard Error)
- Test for the significance of an individual parameter
  - Wald Chi Square = (Parameter Estimate/Standard Error)^2

    Approximately follows a Chi Squared distribution with 1 degree of freedom
  - P-value is probability of obtaining a Chi Square statistic of given magnitude by pure chance

    Lower p-value → more significant

Two Modeling Approaches

- Pure Premium Approach: Build a single model for pure premium
  - Generally straightforward to implement
- Frequency-Severity Approach: Build one model for claim frequency and another for claims severity
  - Additional work for additional insight

Pure Premium Approach

- Advantages:
  - Only a single model needs to be built
  - No need to split variable offsets
  - Results often very similar to frequency-severity approach
- Disadvantages:
  - Yields less insight than frequency-severity
  - Tweedie distribution is only good choice
    - Relatively new and mathematically complex
    - Includes implicit assumptions that may not hold

Frequency-Severity Approach

- Advantages:
  - May yield meaningful insights about data
  - Can choose from several well-known and well-understood distributions
- Disadvantages:
  - Two models to build, run, and validate
  - Requires splitting variable offsets
  - Often produces limited additional benefit

Tweedie Distribution

- Mixed Poisson-Gamma process – number of claims follow a Poisson distribution, and the size of each claim follows a Gamma distribution
- The Tweedie is a 3-parameter distribution:
  - Mean (μ), equal to the product of the means of the underlying Poisson and Gamma distributions
  - Power (p), which depends on the coefficient of variation of the underlying Gamma distribution
  - Dispersion (φ), a measure of variance

Frequency Distribution Options

- Poisson
  - The Coca Cola of claim count distributions
- Overdispersed frequency distributions
  - Overdispersed Poisson
  - Zero-Inflated Poisson
  - Negative Binomial
  - Zero-Inflated Negative Binomial

Severity Distribution Options

- Several reasonable distributions
- Criteria
  - Member of exponential family
  - $p \geq 2$, where $V(\mu)=\mu^p$
- In order of increasing variance:
  - Gamma ($p=2$)
  - Tweedie ($2<p<3$)
  - Inverse Gaussian ($p=3$)
  - Tweedie ($p>3$)

Three Pillars of Model Validation

- Tests of Fit
- Tests of Lift
- Tests of Stability

Fit Statistics

- Traditional: Absolute/Squared Error
- Alternatives: Likelihood, Deviance, Pearson's Chi-Squared
- Penalized: AIC, BIC
- Per Observation: Residuals, Leverage

Absolute/Squared Error

- Only appropriate if data is normally distributed
- Inappropriate to use on disaggregate claim frequency, severity, or pure premium data
- Useful to assess model fit within buckets
  - Bucket data into percentiles, or similar quantiles, and calculate squared difference between actual and predicted for each bucket

Better Alternatives to Squared Error

- Likelihood: chance of observation, given model
  - Always increases as parameters are added to model
- Deviance: twice the difference in loglikelihoods between the saturated and fitted models
  - GLMs are fit so as to minimize deviance
  - Accounts for the shape of the distribution
- Pearson's chi-squared: squared error divided by the variance function of the distribution
  - Accounts for the skew of the distribution

Penalized Measures

- Akaike Information Criterion (AIC): Penalizes loglikelihood for additional model parameters
- Bayesian Information Criterion (BIC): Penalizes loglikelihood for additional model parameters, and this penalty increases as the number of records in the dataset increases
  - Can be too restrictive
- Used primarily for variable selection

Per Observation

- Traditional residual: actual minus predicted
- Deviance residual: square root of weighted deviance times sign of actual minus predicted
  - Reflects the shape of the distribution
  - Plotting deviance residual against weight or any predictor should yield an uninformative cloud
  - Should be approximately normally distributed
- Leverage: used to identify extreme outliers
  - Does not necessarily measure impact

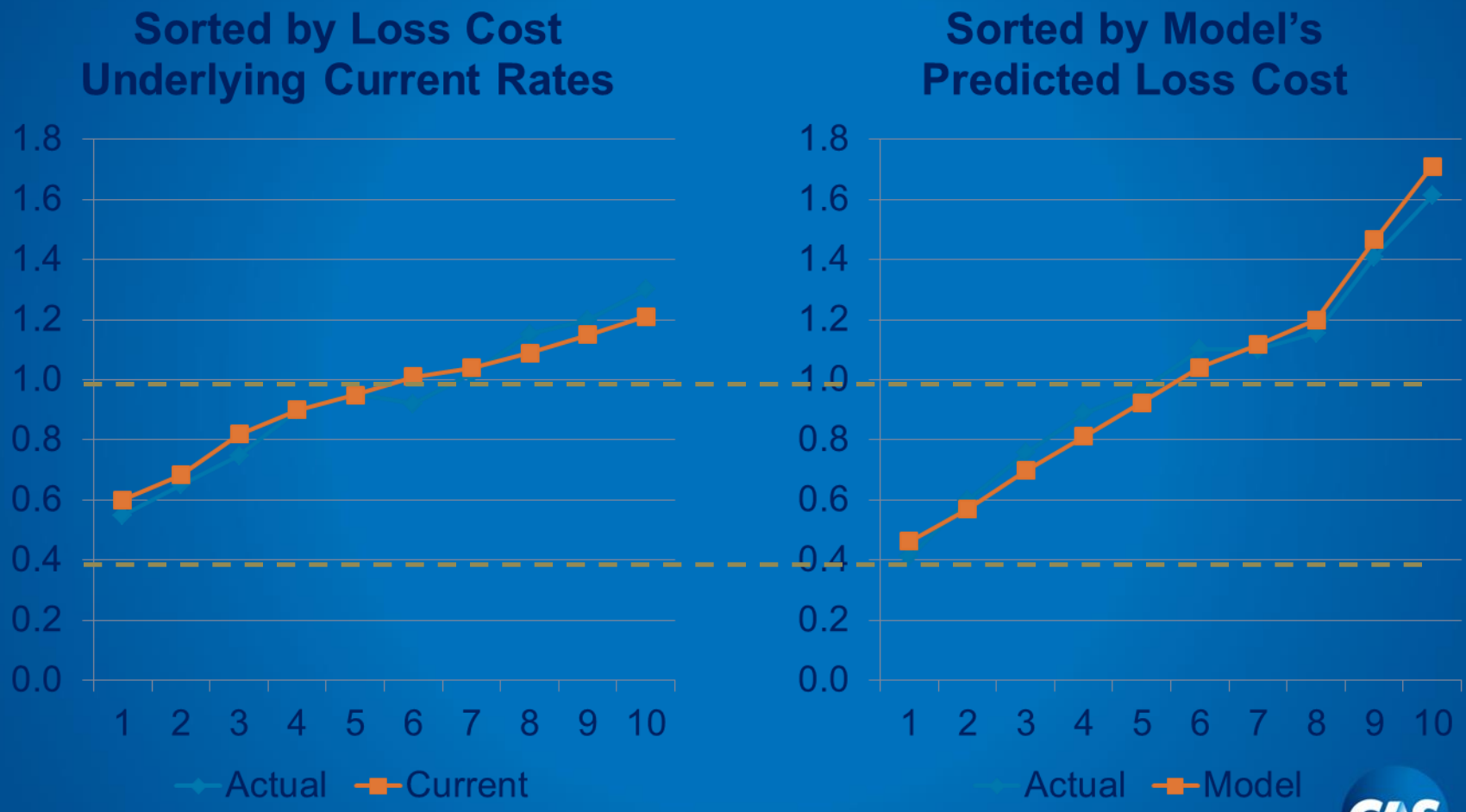
Model Lift

- Lift is meant to approximate economic value
  - Fit has no relationship with economic value
- Economic value is produced by comparative advantage in avoidance of adverse selection
  - Lift is a *comparative* measure, i.e. the lift of one model over another, or the lift of a model over status quo
- Lift should always be measured on holdout data

Lift Measures

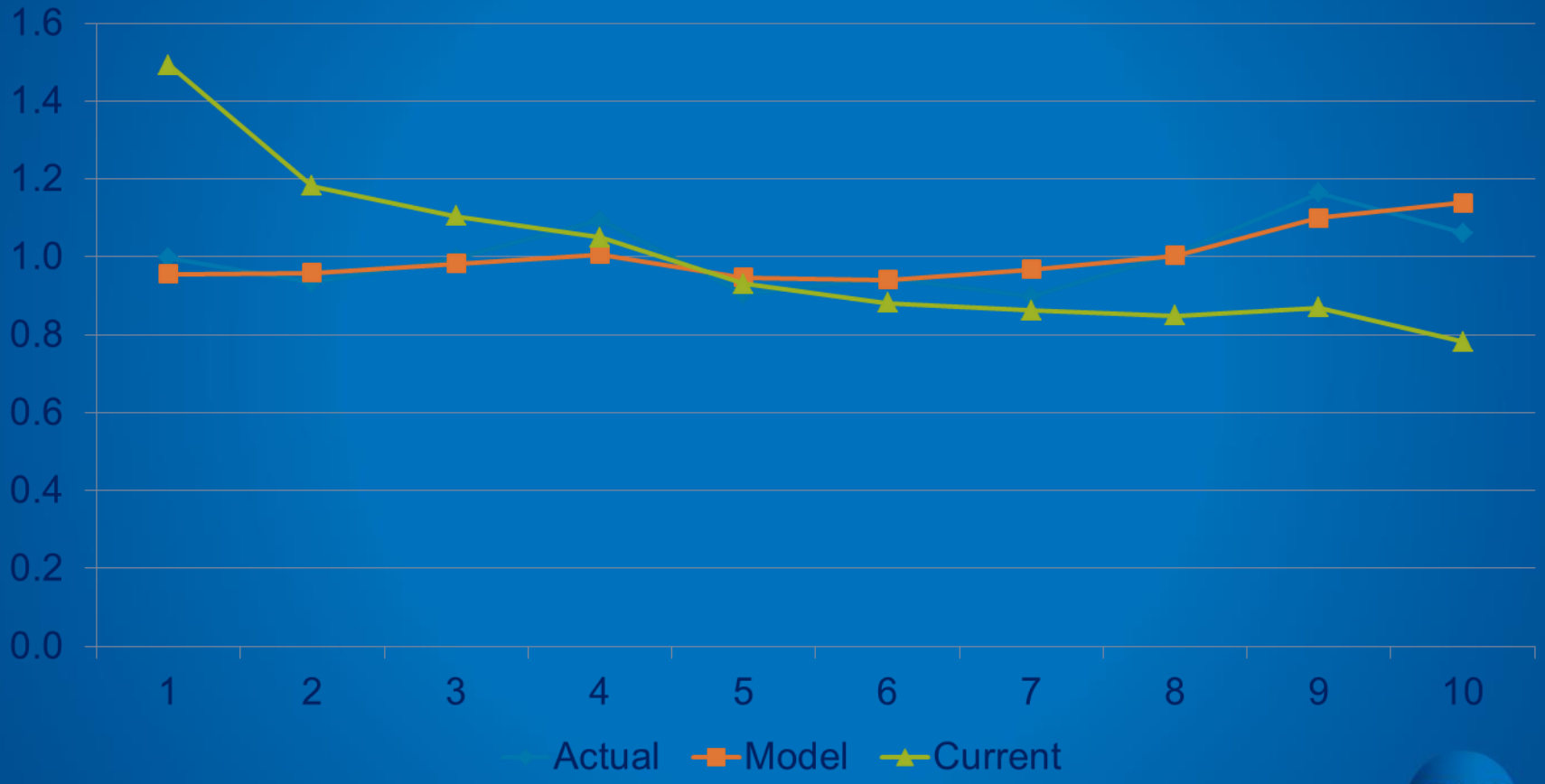
- Simple Quantile Plot
- Double Lift Chart
- Loss Ratio Chart
- Gini Index

Model Lift – Simple Quantile Plots



100 Years of Expertise, Insight & Solutions

Double Lift Chart

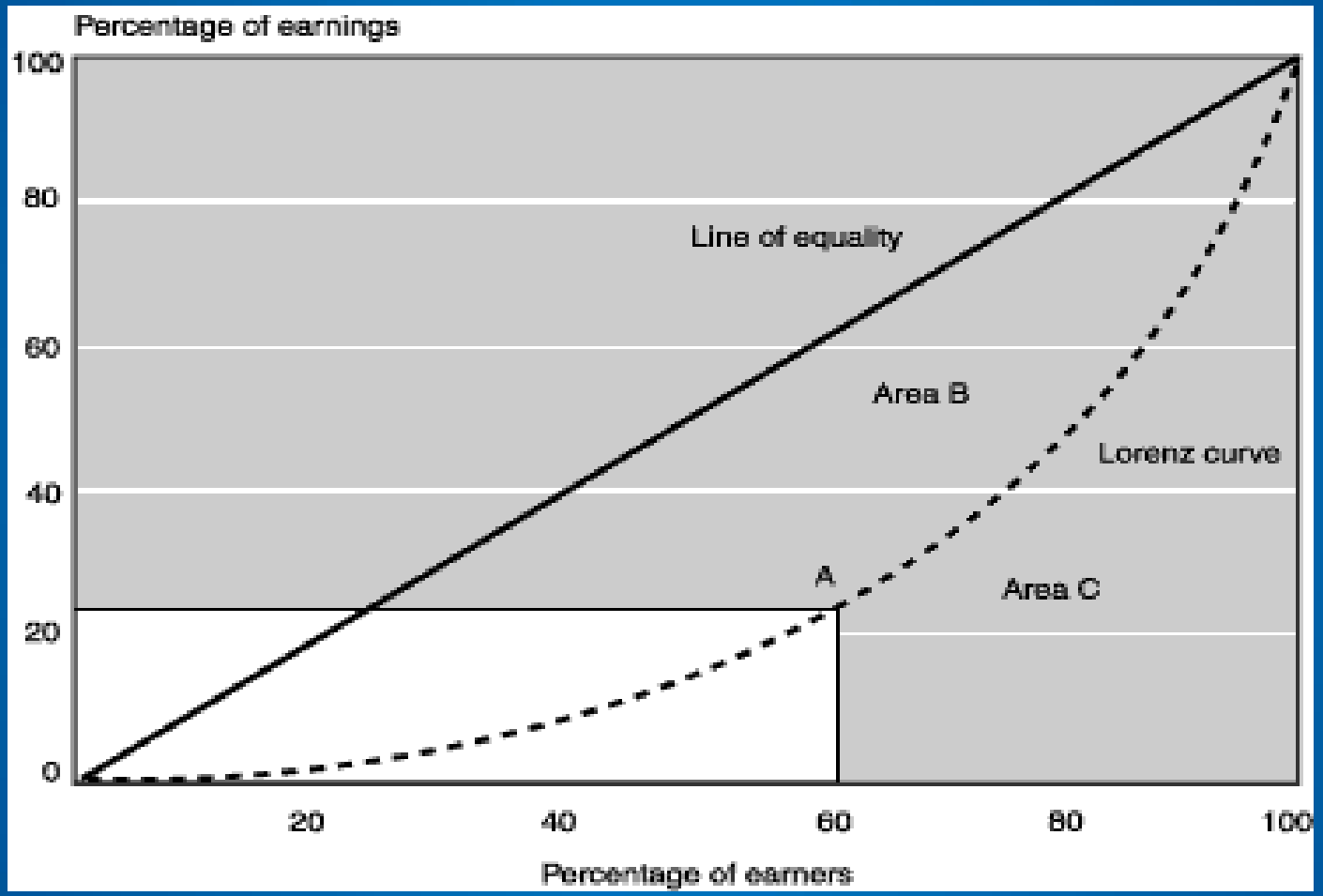


100 Years of Expertise, Insight & Solutions

Loss Ratio Chart

100 Years of Expertise, Insight & Solutions

Economic Gini Index



100 Years of Expertise, Insight & Solutions

Gini Index of Rating Plan

- Model should differentiate lowest and highest loss cost policyholders
- Creation of Gini index:
  - Order policyholders by model prediction, from best to worst
  - X-axis is cumulative percent of exposures
  - Y-axis is cumulative percent of losses
- Had model produced Gini index in prior slide, would have identified 60% of exposures that contribute only 20% of losses

Methods for Testing Model Stability

- Cross-validation
  - Split data into subsets (e.g. by time period)
  - Refit model on each subset
  - Compare model parameter estimates
- Bootstrapping
  - Refit model on many bootstrapped samples
  - Calculate variability of parameter estimates
- Deletion of influential records

Measures of Influence

- Cook's Distance: Statistical measure of the impact each record has on the *overall* model
  - Excellent tool for identifying errors or anomalies
  - Deletion of records with high Cook's Distance may significantly change model results, and so this procedure can be used to test stability
- DFBETA: Influence on a *certain parameter*
- Influence is not to be confused with leverage

Predictive Modeling Process

- Collect Data
- Exploratory Data Analysis
  - Examine univariate distributions
  - Examine relationship of each variable to target
- Specify Model
- Evaluate Output
- Validate Model
- “Productize” Model
- Maintain Model
- Rebuild Model

Some important considerations

- What are you trying to predict?
- Which explanatory variables to use?
  - Legal concerns
  - Reputation risk
  - Explainability
  - Data integrity
  - Cost
- What components of the product are not included in the model?
- Does the model need regulatory approval?
- What do you expect to find?

100 Years of Expertise, Insight & Solutions

Some important considerations

- Who will work on the model, and how will you set clearly defined roles for each member of the team?
- Does the model significantly outperform the existing one?
- How will you sell the results to each of the key stakeholders?

Predictive modeling…the next 100 years!

- As GLMs become ingrained in actuarial work, what will be the next development in actuarial predictive modeling?
- What challenges will we face in adopting new modeling frameworks?
  1. Gaining acceptance from key stakeholders
  2. Regulatory approval
  3. Conforming to actuarial standards
  4. Finding/developing the necessary talent

100 Years of Expertise, Insight & Solutions

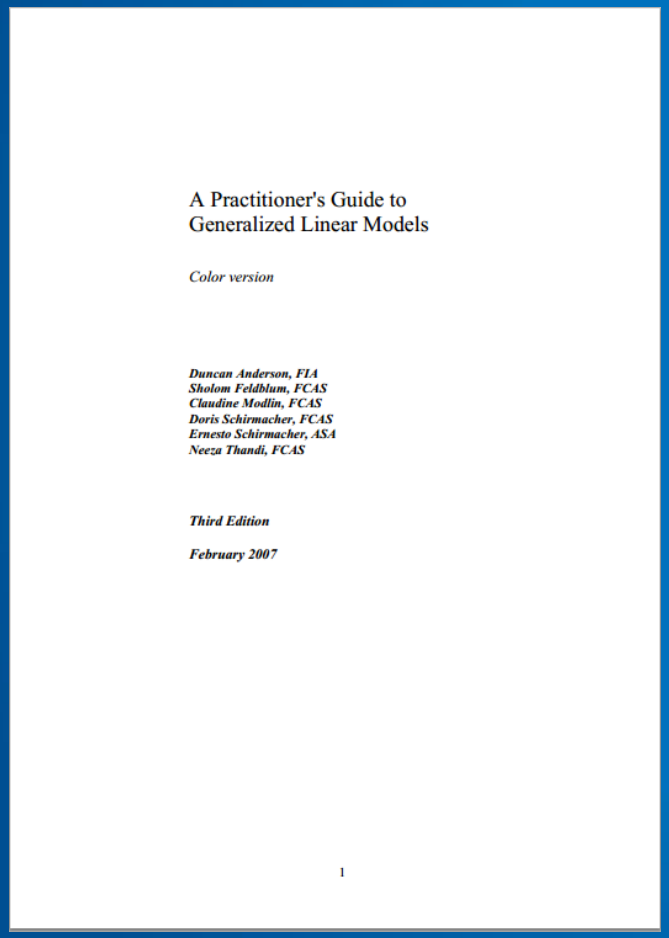
The future of actuarial predictive modeling

- Extended linear models
  - Generalized Linear Mixed Models (GLMMs): include both fixed and random effects
  - Generalized Additive Models (GAMs): linear predictor depends on smooth function of predictor variables
- Bayesian statistics
  - Parameters of the distribution are random and have an a priori distribution
  - Data is sampled to create posterior distribution
  - Produces better estimates of uncertainty than “traditional” statistics
  - Potential to be heavily used in reserving

The black box

- "Black box" methods (e.g. machine learning) are gaining popularity among many statisticians
  - Little to no knowledge of the underlying data is required
- Use of such methods in actuarial work raises several questions:
  1. Can we accept an accurate but unexplainable answer?
  2. Do actuaries need to be involved in the construction of such models?

For Further Reference



A Practitioner's Guide to
Generalized Linear Models

Color version

Duncan Anderson, FIA
Sholom Feldblum, FCAS
Claudine Modlin, FCAS
Doris Schirmacher, FCAS
Ernesto Schirmacher, ASA
Neeza Thandi, FCAS

Third Edition

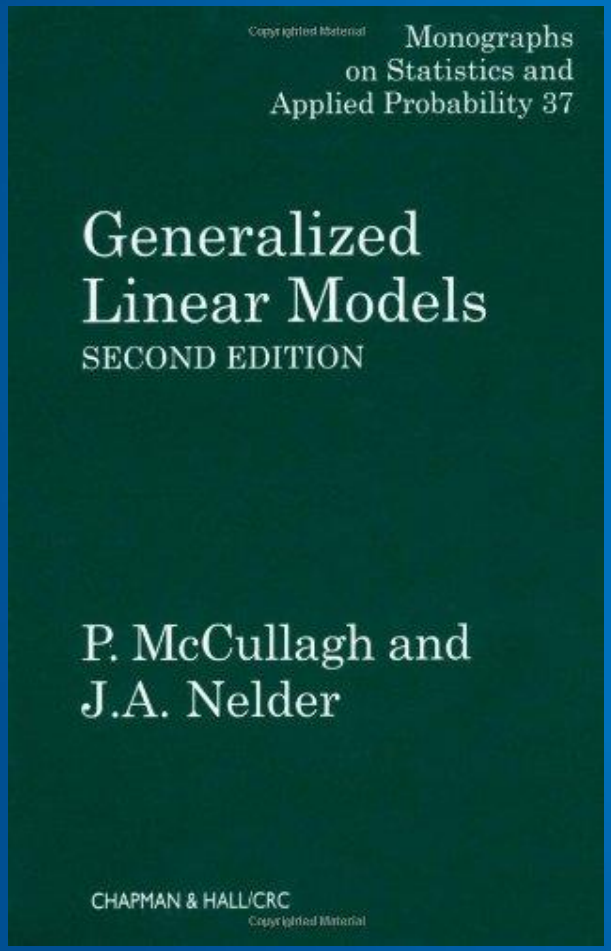
February 2007

1

- Anderson, Duncan, et. al., *A Practitioner's Guide to Generalized Linear Models*, CAS Discussion Paper Program, 2004

100 Years of Expertise, Insight & Solutions

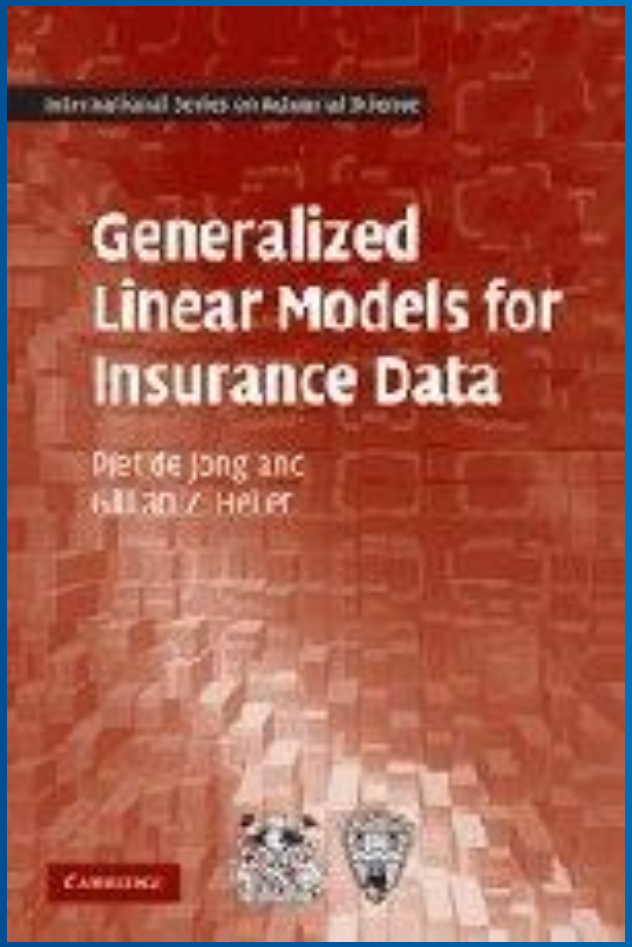
For Further Reference



- McCullagh, Peter and Nelder, John A., *Generalized Linear Models*, 2nd Ed., Chapman & Hall, 1989

100 Years of Expertise, Insight & Solutions

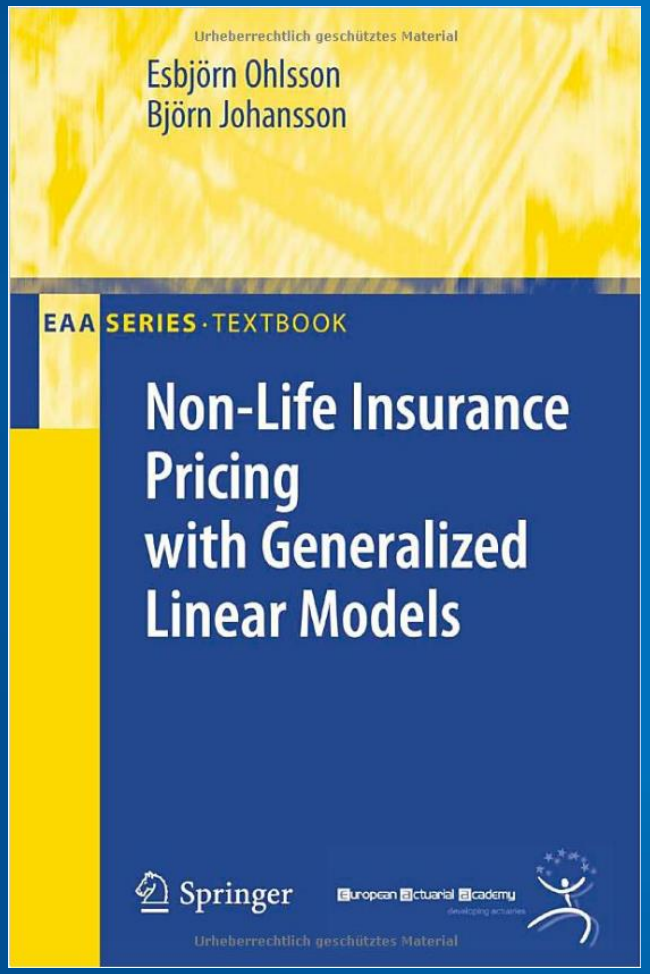
For Further Reference



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

100 Years of Expertise, Insight & Solutions

For Further Reference



- Ohlsson, Esbjörn and Johansson, Björn, *Non-Life Pricing with Generalized Linear Models*, Springer, 2010

Questions?