
And The Winner Is...?

How to Pick a Better Model

Part 2 – Goodness-of-Fit and Internal Stability

Dan Tevet, FCAS, MAAA



Goodness-of-Fit

- Trying to answer question: How well does our model fit the data?
- Can be measured on training data or on holdout data
- By identifying areas of poor model fit, we may be able to improve our model
- A few ways to measure goodness-of-fit
 - Squared or absolute error
 - Likelihood/log-likelihood
 - AIC/BIC
 - Deviance/deviance residuals
 - Plot of actual versus predicted target

Squared Error & Absolute Error

- For each record, calculate the squared or absolute difference between actual and predicted target variable
- Easy and intuitive, but generally inappropriate for insurance data, and can lead to selection of wrong model
- Squared error appropriate for Normal data, but insurance data generally not Normal

Residuals

- Raw residual = $y_i - \mu_i$, where y is actual value of target variable and μ is predicted value
- In simple linear regression, residuals are supposed to be Normally distributed, and departure from Normality indicates poor fit
- For insurance data, raw residuals are highly skewed and generally not useful

Likelihood

- The probability, as predicted by our model, that what actually did occur would occur
- A GLM calculates the parameters that maximize likelihood
- Higher likelihood → better model fit (very simple terms)
- Problem with likelihood – adding a variable always improves likelihood

AIC & BIC – penalized measures of fit

- Akaike Information Criterion (AIC) =
 $-2 * (\text{Log Likelihood}) + 2 * (\text{Number of Parameters in Model})$
- Bayesian Information Criterion (BIC) =
 $-2 * (\text{Log Likelihood}) + (\text{Number of Parameters in Model}) * \ln(\text{Number of Records in Dataset})$
- Good rule for deciding which variables to include – unless a variables improves AIC or BIC, don't include it

Deviance

- Saturated model – the model with the highest possible likelihood
 - One indicator variable for each record, so model fits data perfectly
- Deviance = $2 * (\text{loglikelihood of saturated model} - \text{loglikelihood of fitted model})$
- GLMs minimize deviance
- Like squared error, but reflects shape of assumed distribution
- We generally fit skewed distributions to insurance data (Tweedie, gamma, etc), and thus deviance is more appropriate than squared error

Deviance – in Math

- Poisson: $2 \sum_i w_i \left(y_i \ln \frac{y_i}{\mu_i} - y_i + \mu_i \right)$
- Gamma: $2 \sum_i w_i \left(-\ln \frac{y_i}{\mu_i} + \frac{y_i - \mu_i}{\mu_i} \right)$
- Tweedie: $2 \sum_i w_i \left(y_i \frac{y_i^{1-p} - \mu_i^{1-p}}{1-p} - \frac{y_i^{2-p} - \mu_i^{2-p}}{2-p} \right)$
- Normal: $\sum_i w_i (y_i - \mu_i)^2$

Deviance Residuals

- Square root of (weighted) deviance times the sign of actual minus predicted
- Measures amount by which the model missed, but reflects the assumed distribution
- Should be approximately Normally distributed, and far departure from Normality indicates that incorrect distribution has been chosen
- Ideally, there should be no discernable pattern in deviance residuals
 - Model should miss randomly, not systemically

Deviance Residual Diagnostics

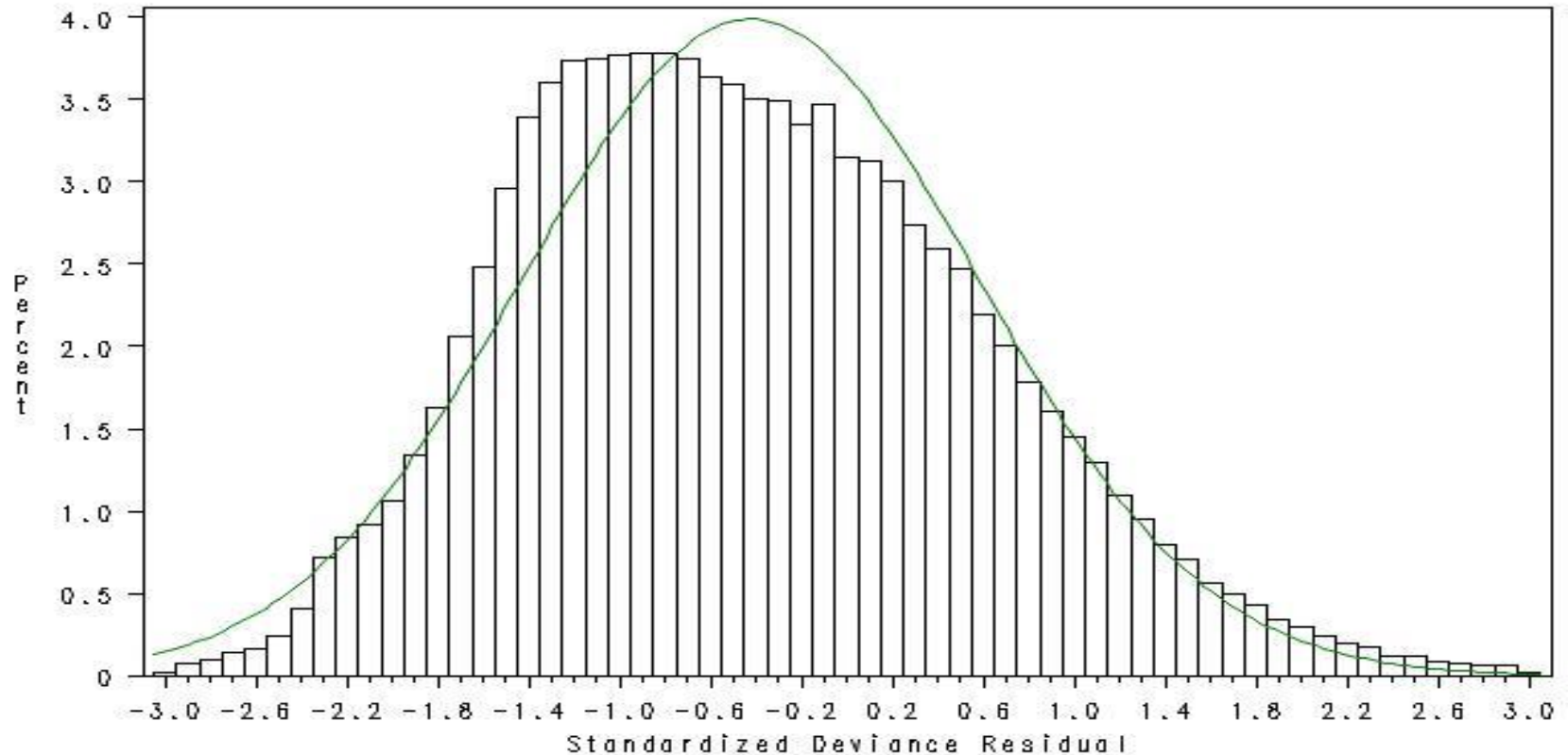
- Histogram of deviance residuals – look for approximate Normality (bell-shape)
 - Far departure from Normality generally indicates that incorrect distribution has been chosen
 - Can also indicate poor fit
- Scatter plot of deviance residuals versus predicted target variable
 - Should be uninformative cloud
 - Pattern in this plot indicates incorrect distribution

Example: Selecting Severity Model

- Goal is to select a distribution to model severity
- Two common choices – Gamma and Inverse Gaussian
 - Gamma: $V(\mu) = \mu^2$
 - Variance of severity is proportional to mean severity squared
 - Inverse Gaussian: $V(\mu) = \mu^3$
 - Variance of severity is proportional to mean severity cubed
- Two lines of business
 - LOB1 is high-frequency, low-severity
 - LOB2 is low-frequency, high-severity

Deviance Residual Histogram

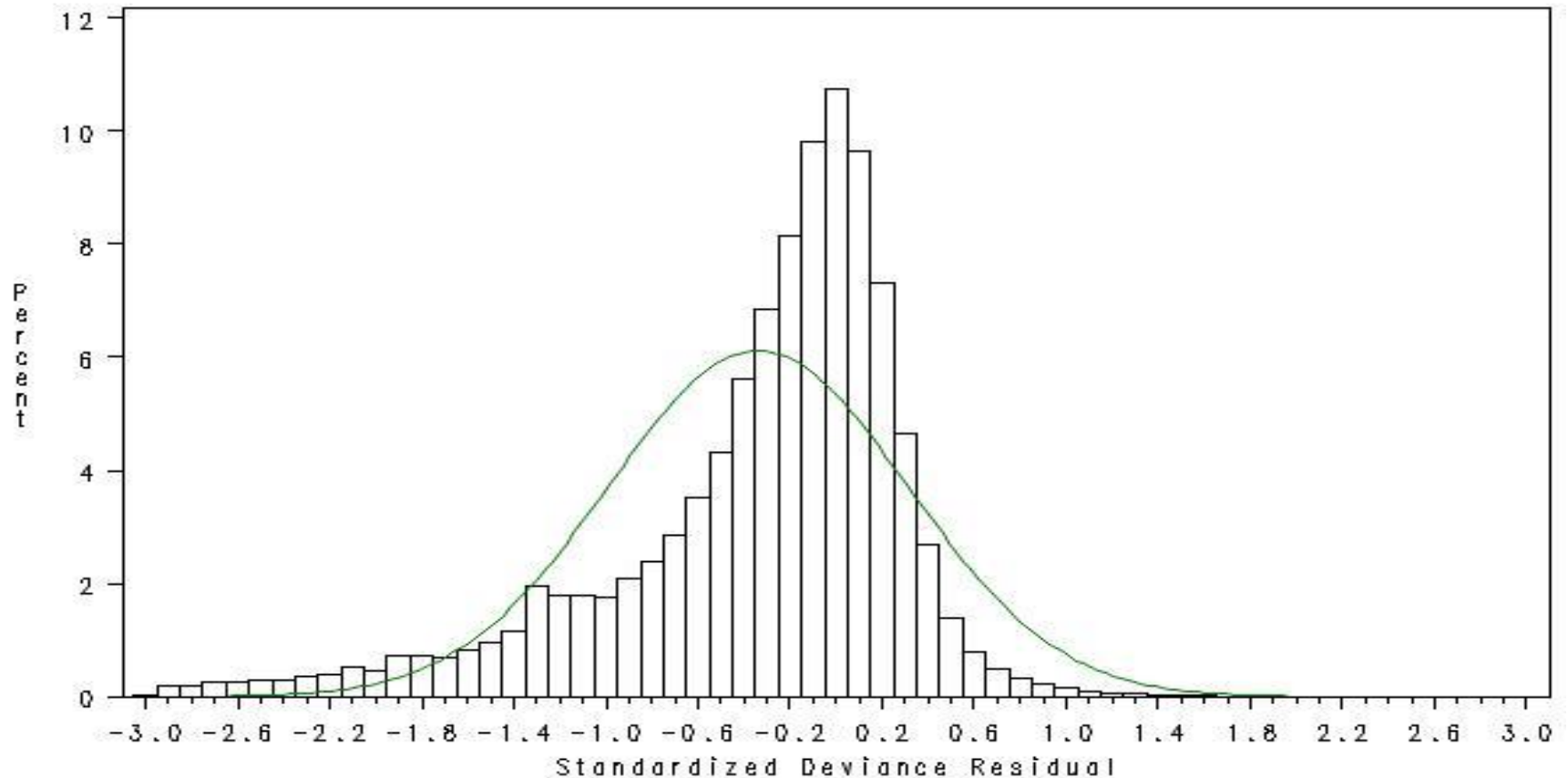
Histogram of Standardized Deviance Residuals
Gamma GLM



LOB1, Gamma GLM

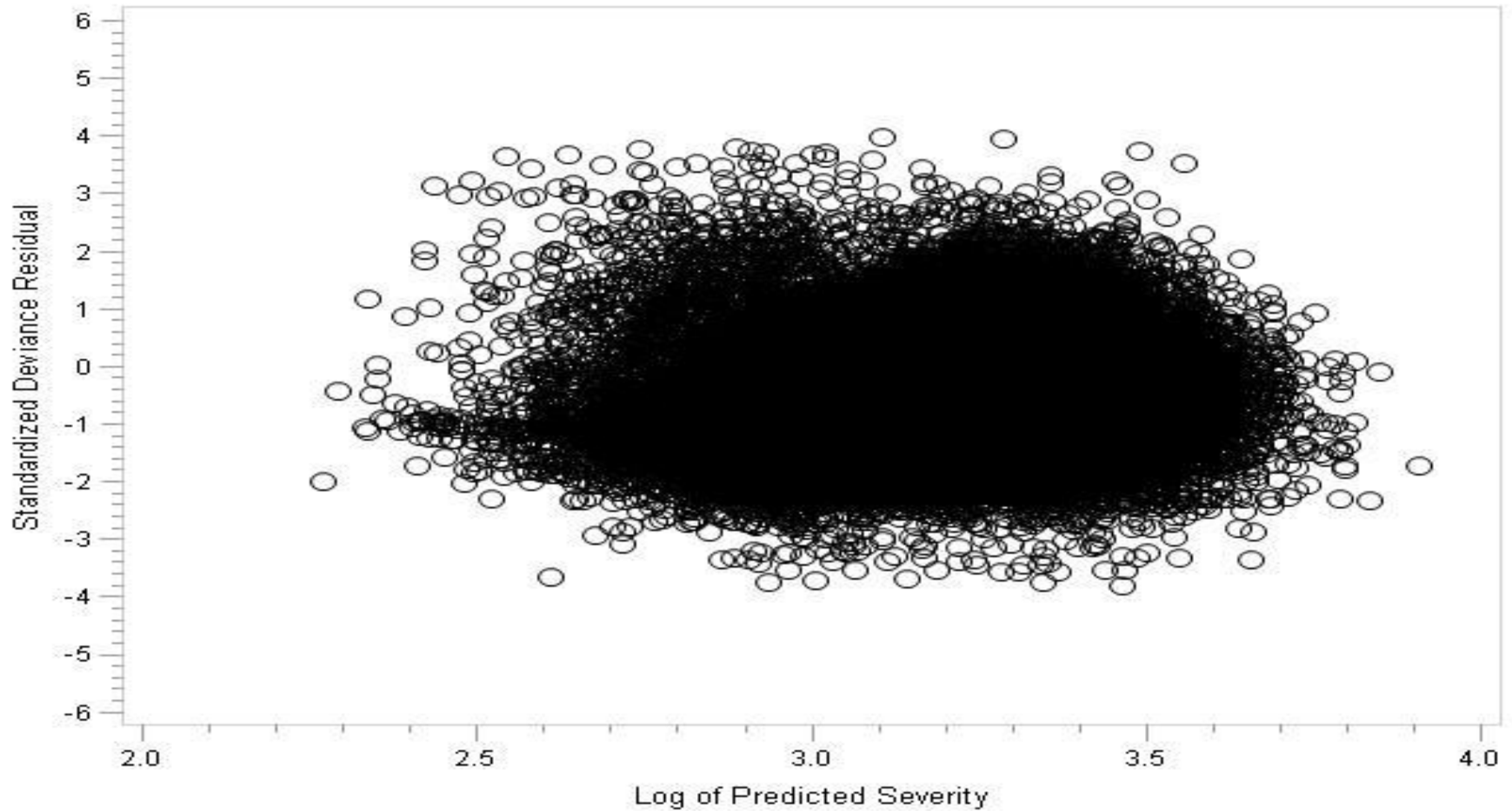
Deviance Residual Histogram

Histogram of Standardized Deviance Residuals
Inverse Gaussian GLM



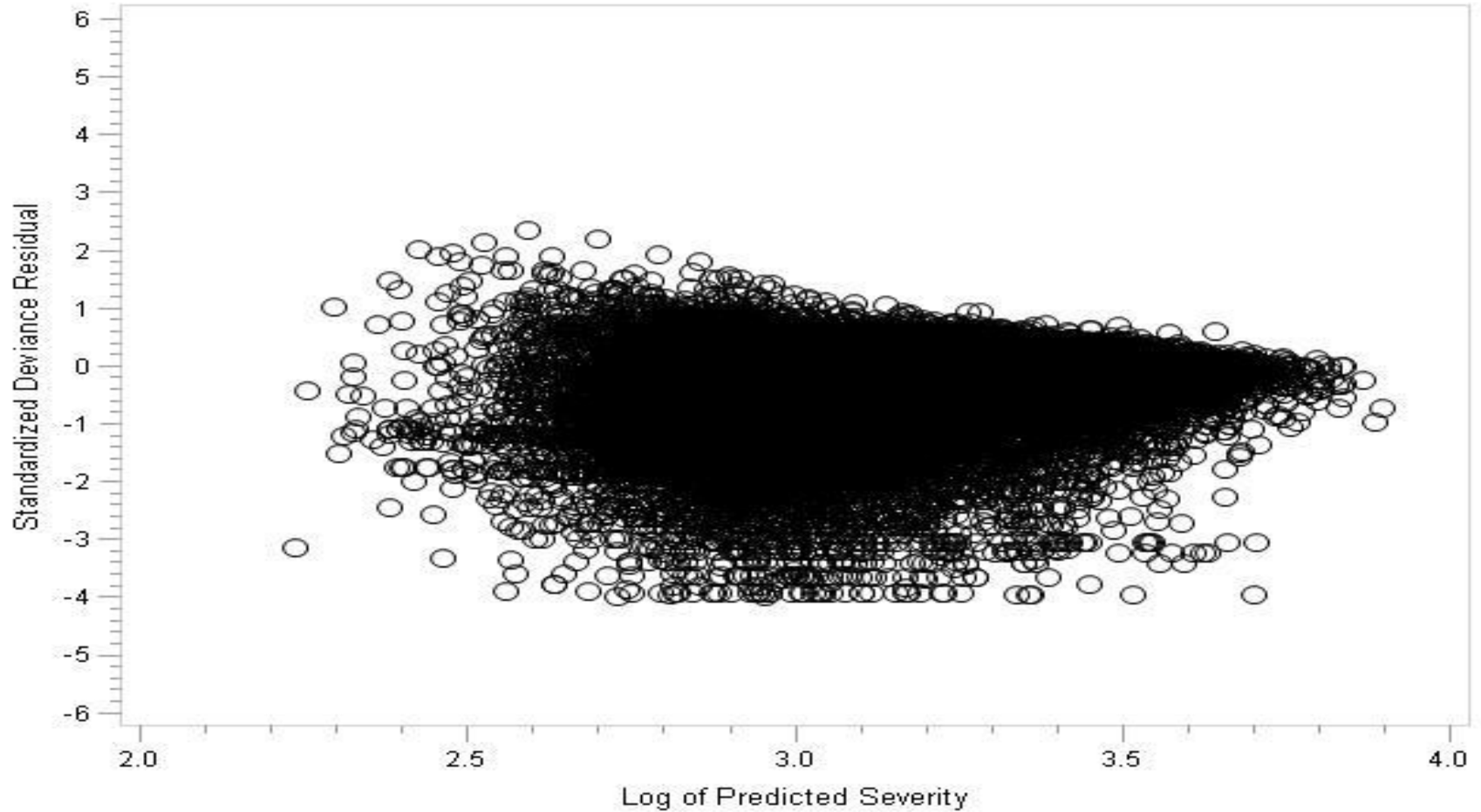
LOB1, IG GLM

Deviance Residual Histogram



LOB1, Gamma GLM

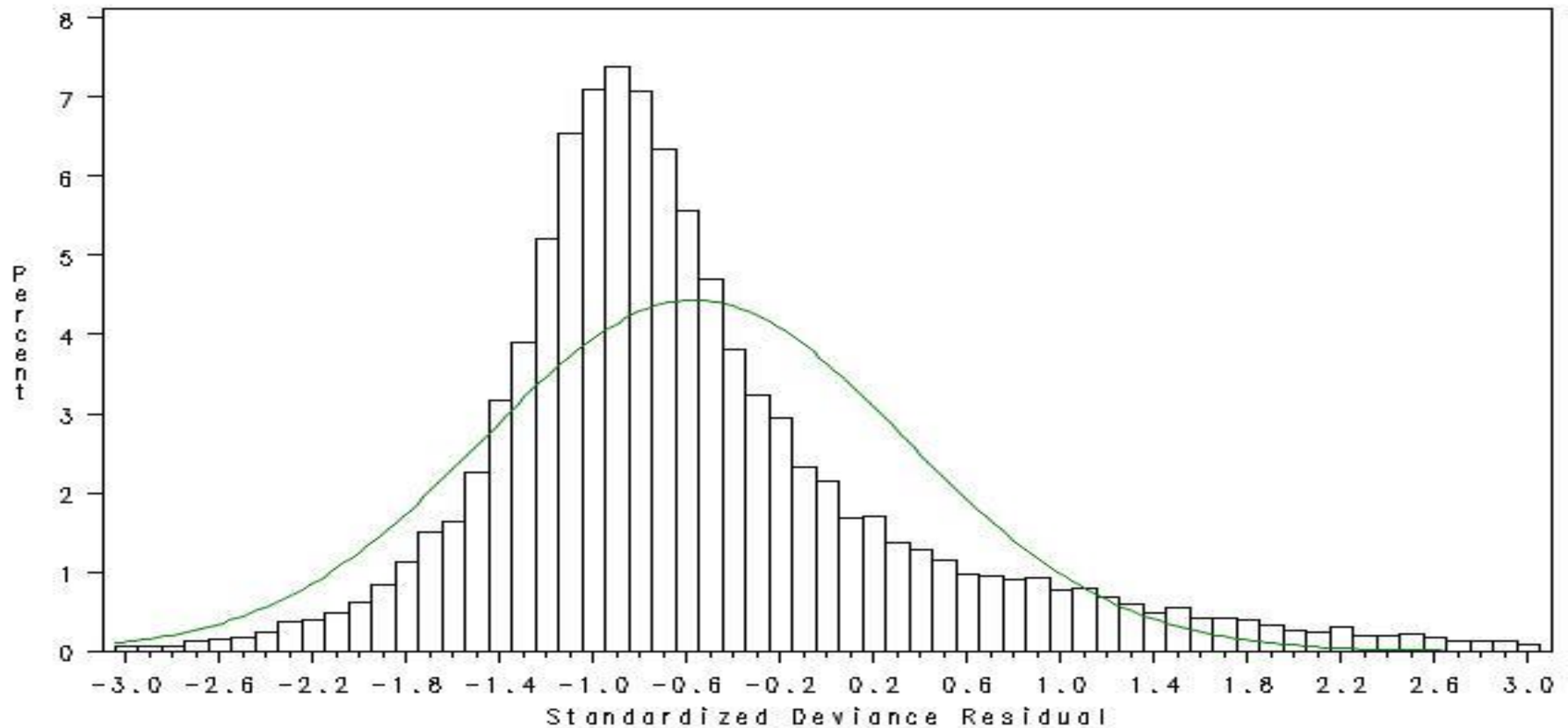
Deviance Residual Histogram



LOB1, IG GLM

Deviance Residual Histogram

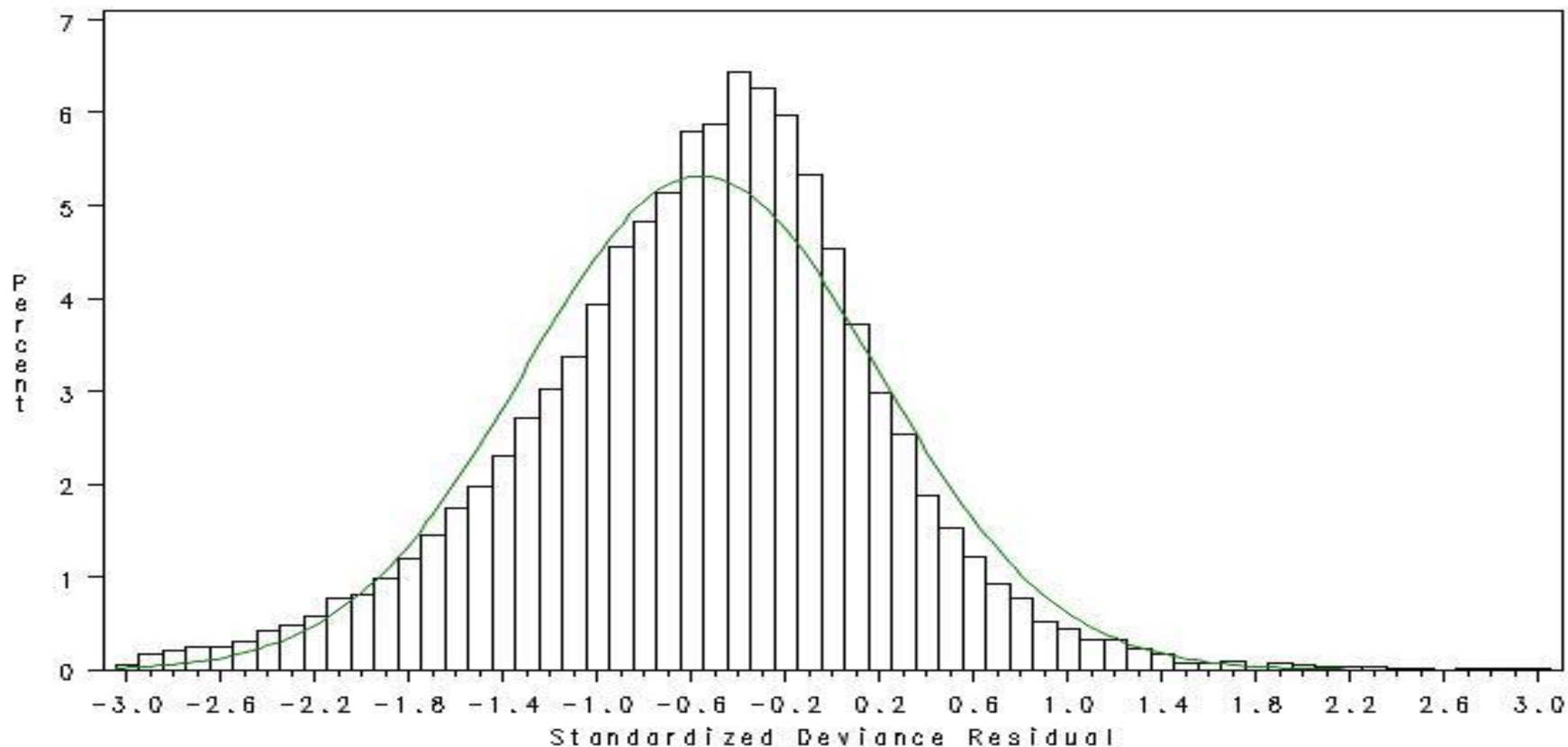
Histogram of Standardized Deviance Residuals
Gamma GLM



LOB2, Gamma GLM

Deviance Residual Histogram

Histogram of Standardized Deviance Residuals
Inverse Gaussian GLM

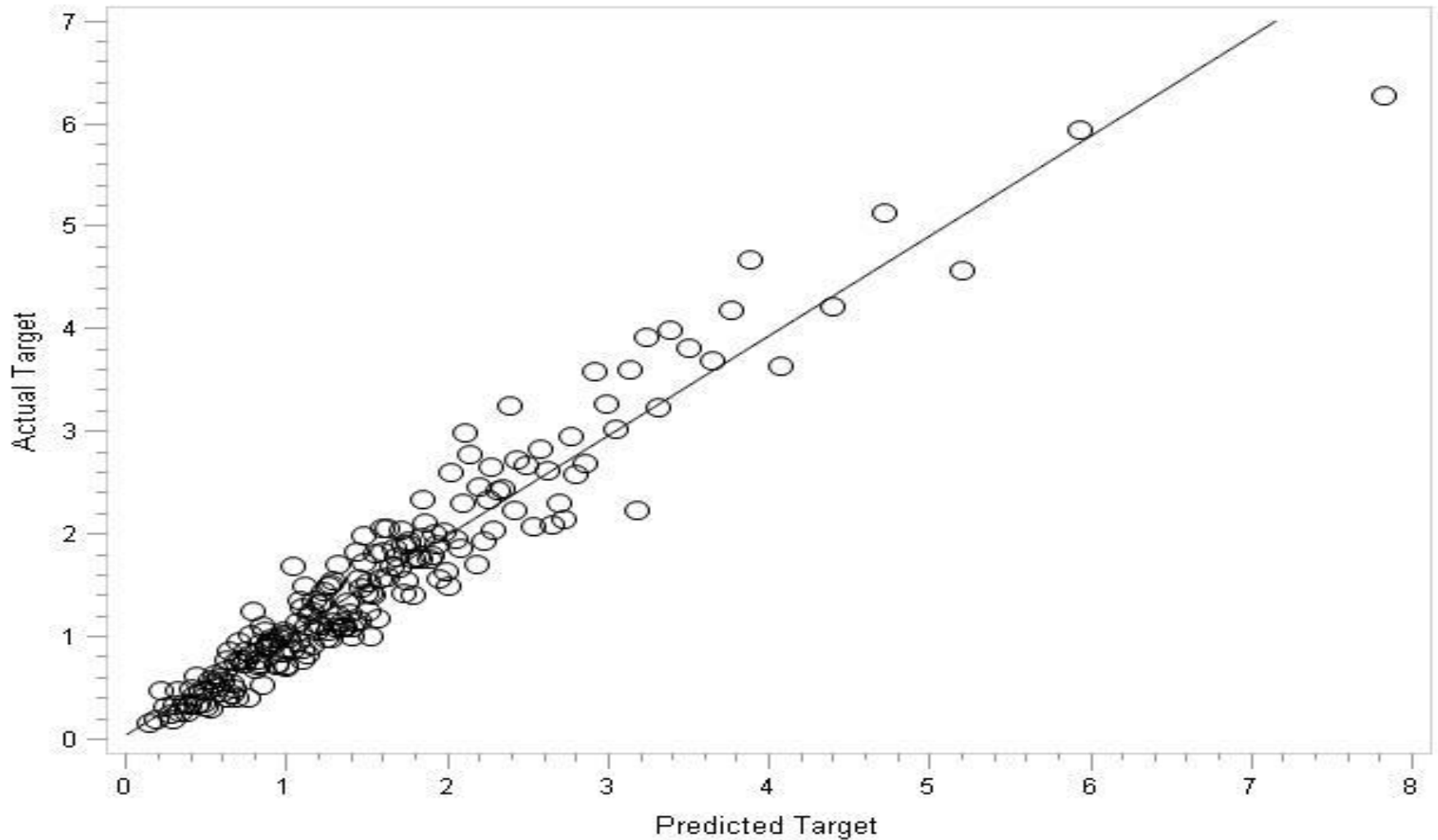


LOB2, IG GLM

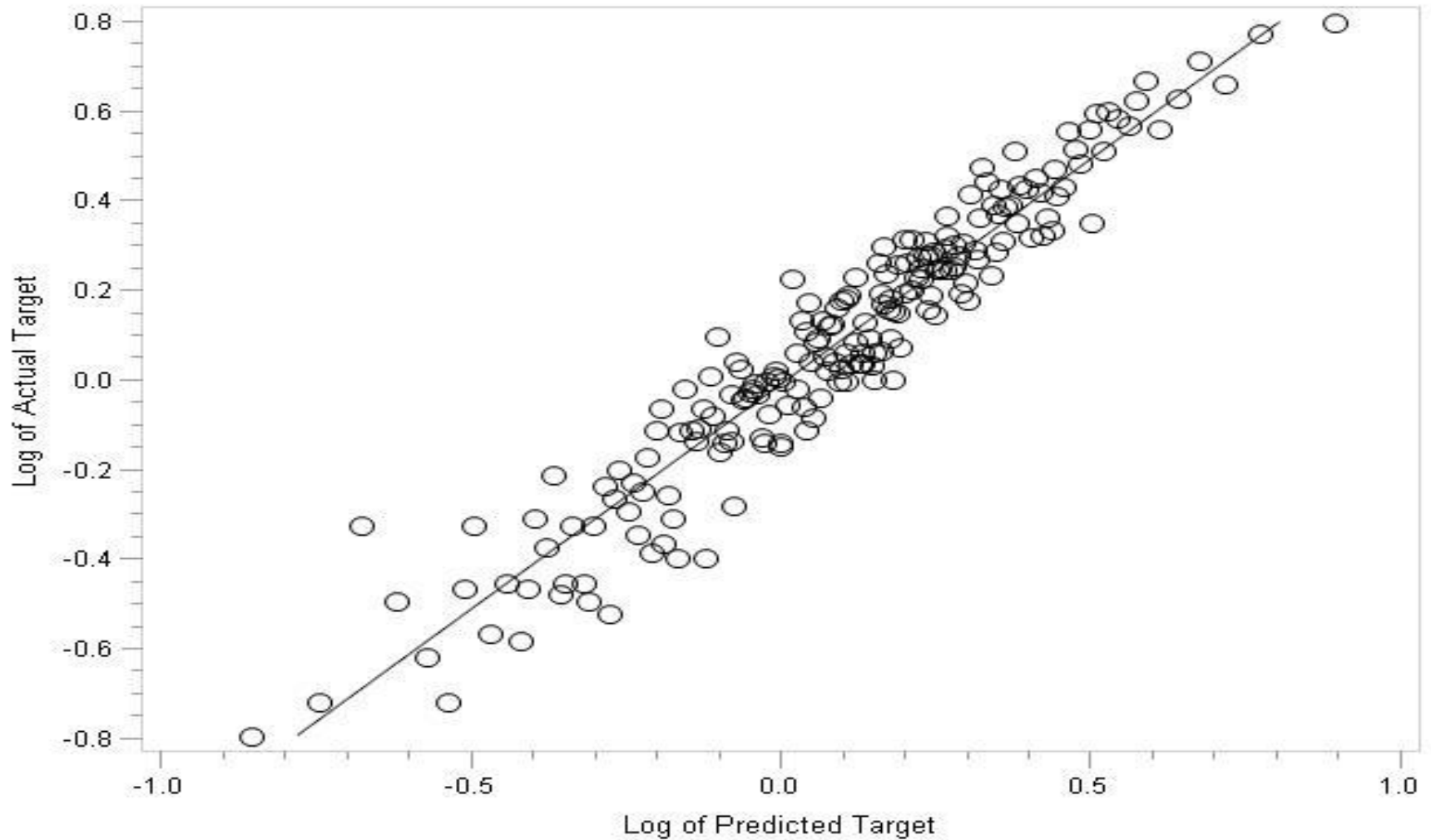
Actual vs Predicted Target

- Scatter plot of actual target variable (on y-axis) versus predicted target variable (on x-axis)
- If model fits well, then plot should produce a straight line, indicating close agreement between actual and predicted
 - Focus on areas where model seems to miss
- If have many records, may need to bucket (such as into percentiles)
- Depending on scale, may need to plot on a log-log scale

Example of Actual vs Predicted



Example of Log of Actual vs Log of Predicted

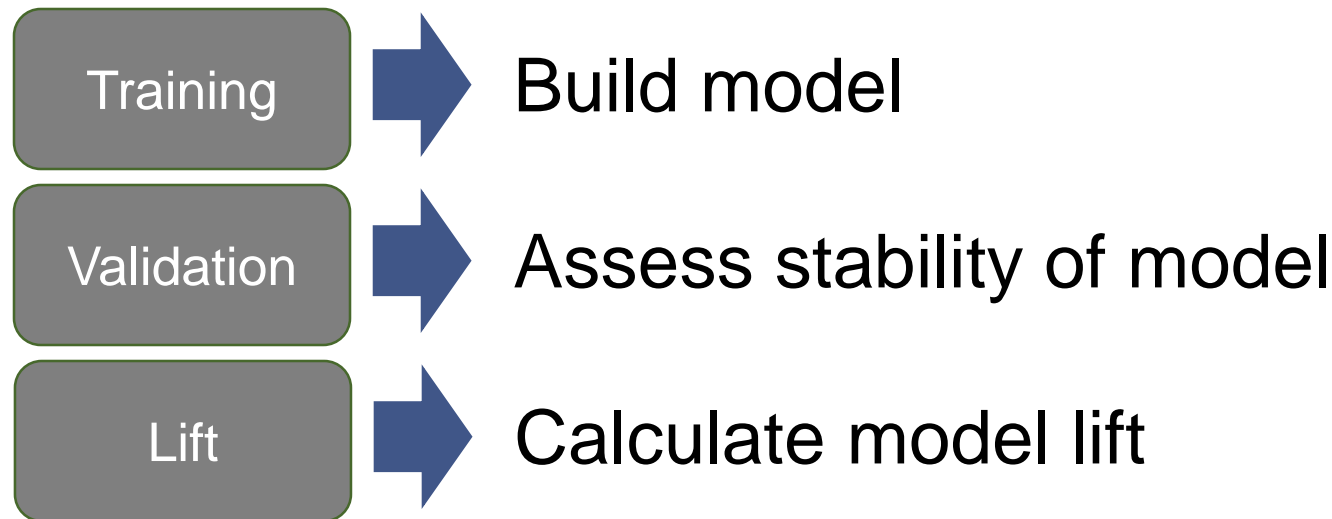


Measuring Internal Stability

- Process of determining how robust model results are
- Getting a second opinion (and third and fourth and fifth) on how well the model performs
- Goals
 - Guard against overfitting
 - Select models that are more stable
 - Better understand inherent volatility of results

Validation 101: Assess model on holdout data

Split data into training-test-validation

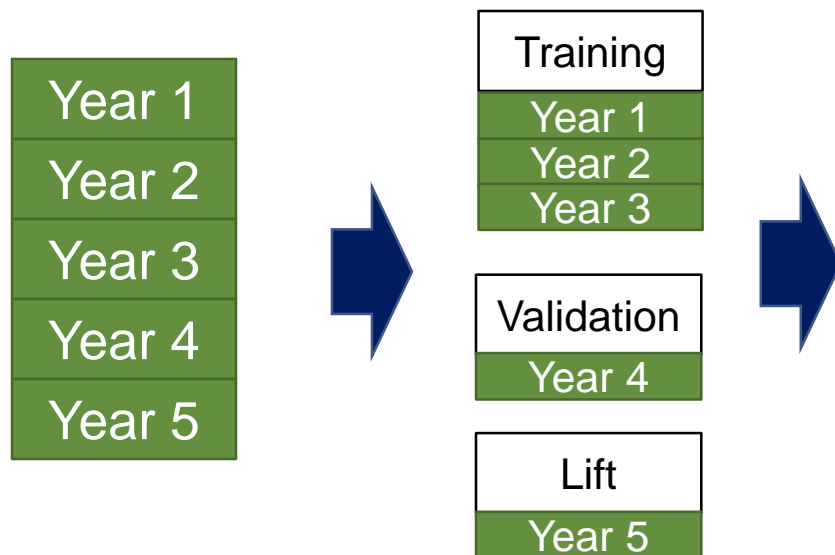


Why is it more complex than this?

- Randomly splitting data doesn't necessary guard against overfitting
- Data may be too thin for such a rigid split
- Doesn't provide a great diversity of opinions


Overfitting can happen if models aren't validated out-of-time

- The same storm hits all homes in an area, the same bad winter impacts auto claims in a region, etc
- Through out-of-time validation, we can help guard against overfitting



How do we use this?

Examine model fit on validation set

If reasonable → 

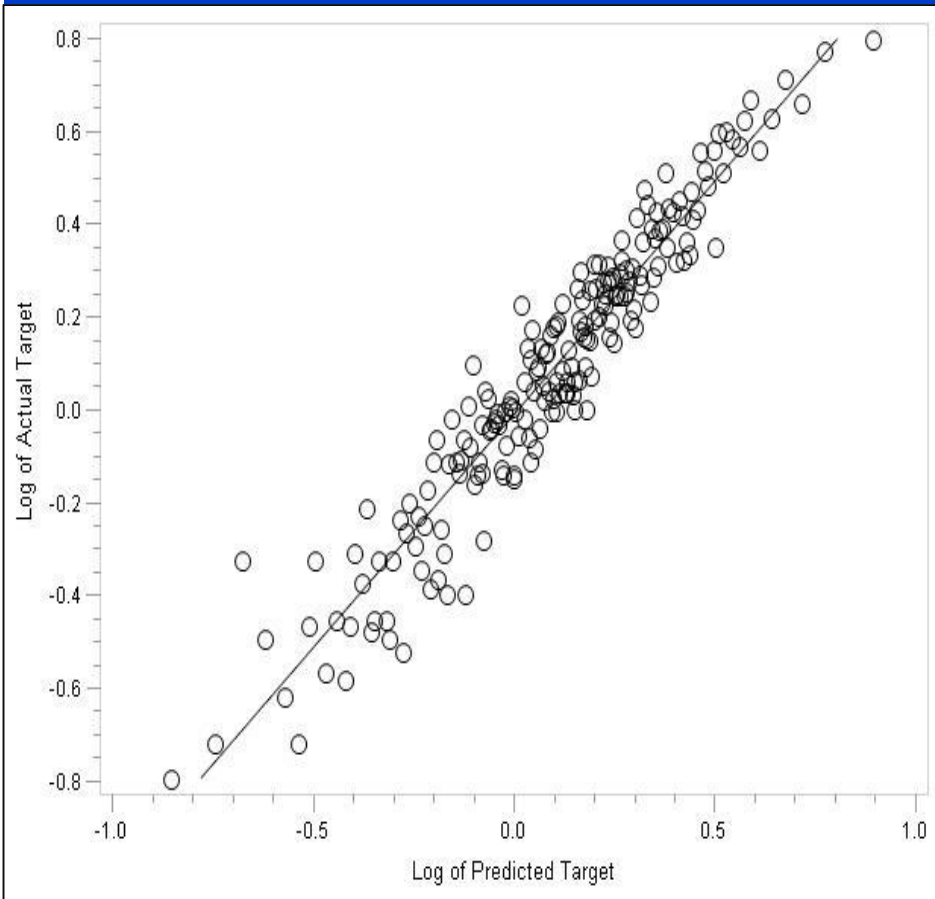
If not →



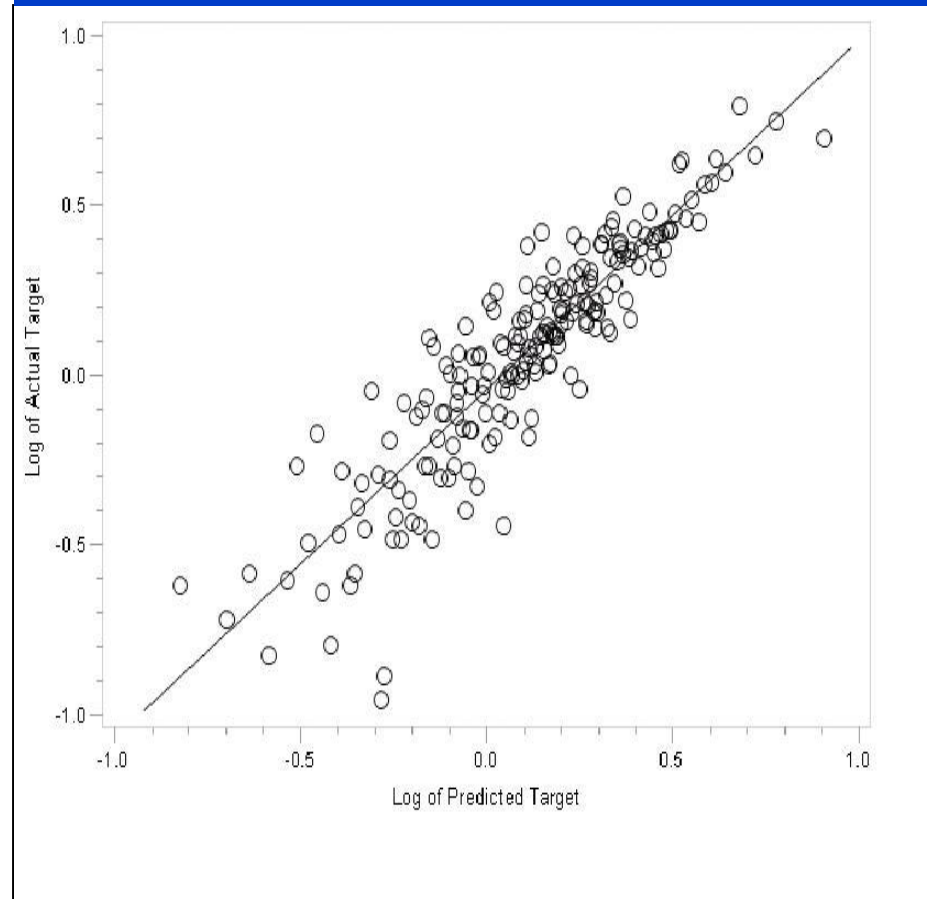
Determining reasonableness often more art than science

Example of Plot of Actual vs Predicted on Holdout

Training

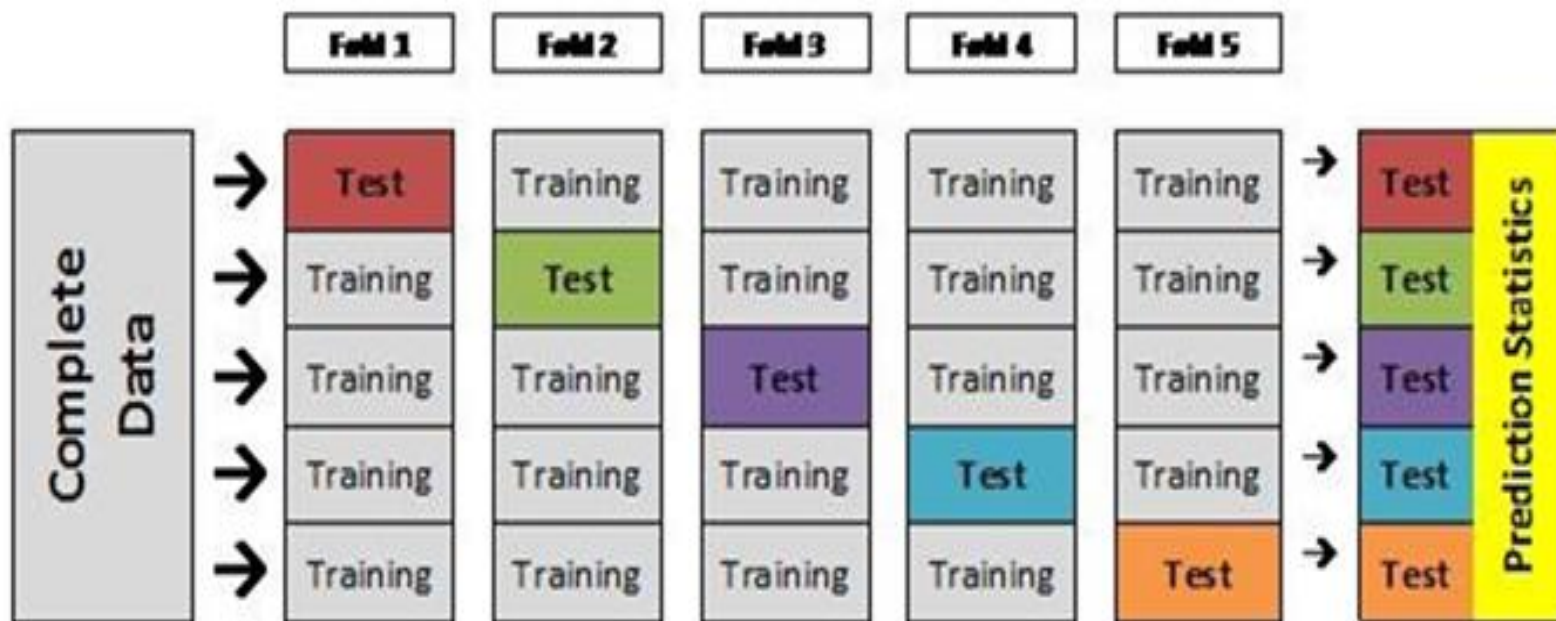


Validation (Out of Time)



Cross-validation is very useful when data is then and can give us more confidence in results

- Split data into subsets
- Refit model on each subset and compare results across subsets



Bootstrapping

- Re-sampling technique that allows us to get more out of our data
- Start with a dataset and sample from it with replacement
 - Some records will get pulled multiple times, and some will not get pulled at all
- Generally, we create a dataset with the same number of records as our original dataset
- Can create many bootstrap datasets, and each dataset can be thought of as an alternate reality
 - Since each bootstrap is an alternate reality, we can use bootstrapping to construct confidence intervals and get more opinions on model performance

We can use bootstrapping to put confidence intervals around lift measures

1

Understand how significant the 'victory' is

Model A currently in production, with Gini of 35.4

Challenger Model B has Gini of 36.9

Should we implement Model B?

2

Better understanding of uncertainty

New model expected to generate \$1M in additional revenue in first 3 months

Actual revenue is \$850K

Did model fail, or is this normal variation?

References

- De Jong and Heller, *Generalized Linear Models for Insurance Data*, Cambridge University Press, 2008
- Efron and Tibshirani, *An Introduction to the Bootstrap*, Chapman & Hall, 1994
- Goldburd, Khare, Tevet, *Generalized Linear Models for Insurance Rating*, CAS Monograph Series, 2016
- McCullagh and Nelder, *Generalized Linear Models*, 2nd Ed., Chapman & Hall, 1989
- Werner and Modlin, *Basic Ratemaking*, Casualty Actuarial Society, Fourth Edition, October 2010.



Liberty Mutual[®]

INSURANCE