LARGE-SCALED INSURANCE ANALYTICS USING TWEEDIE MODELS IN APACHE SPARK

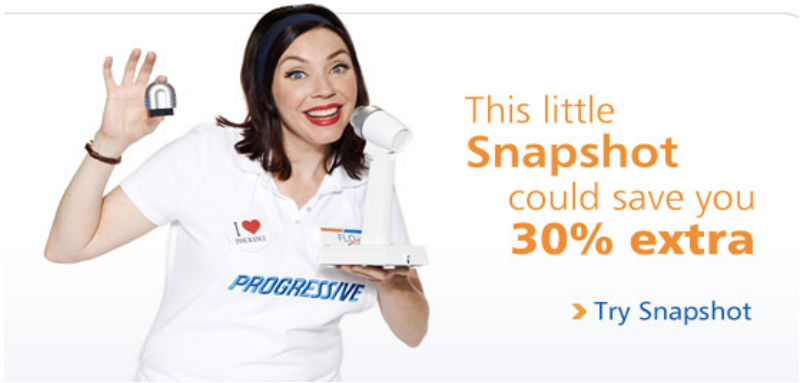
Yanwei (Wayne) Zhang

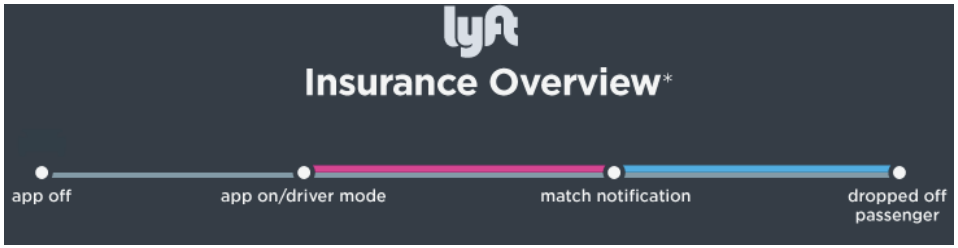
Uber Technologies Inc.

UBER

Usage Based Insurance

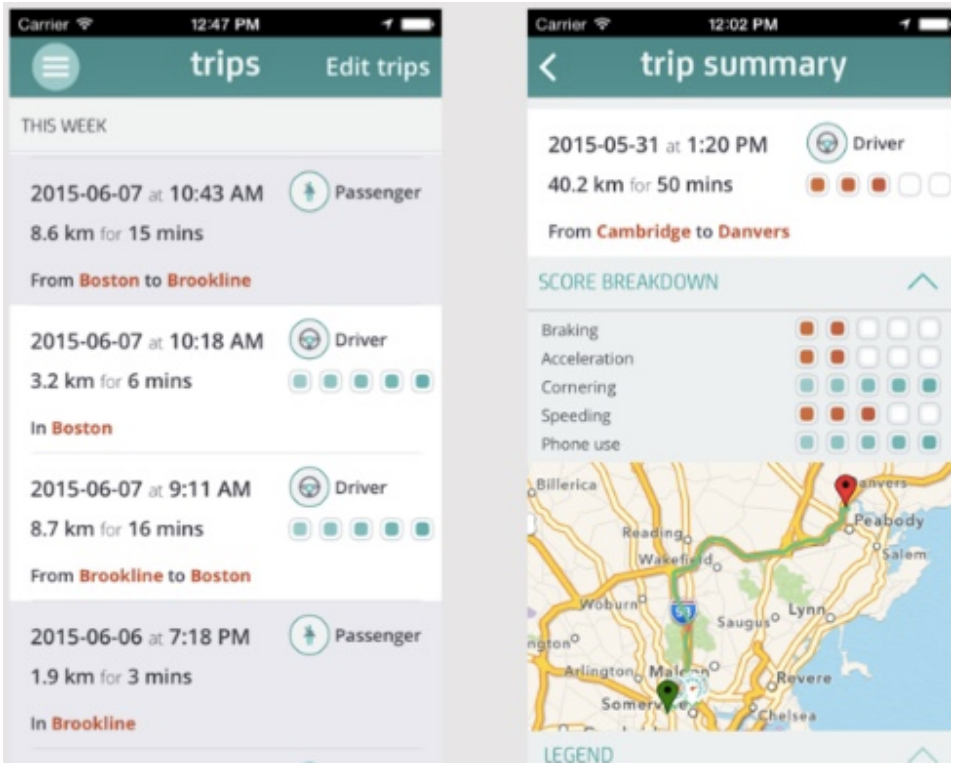
- Pay As You Drive based on miles driven





Data Collection

Trip-level Driving Data



- Collect trip-level data
  - Time of trip
  - Location
  - Vehicle movement
    - GPS, IMU
    - Speed & acceleration
- Merged data
  - Weather & Traffic
  - Demographics

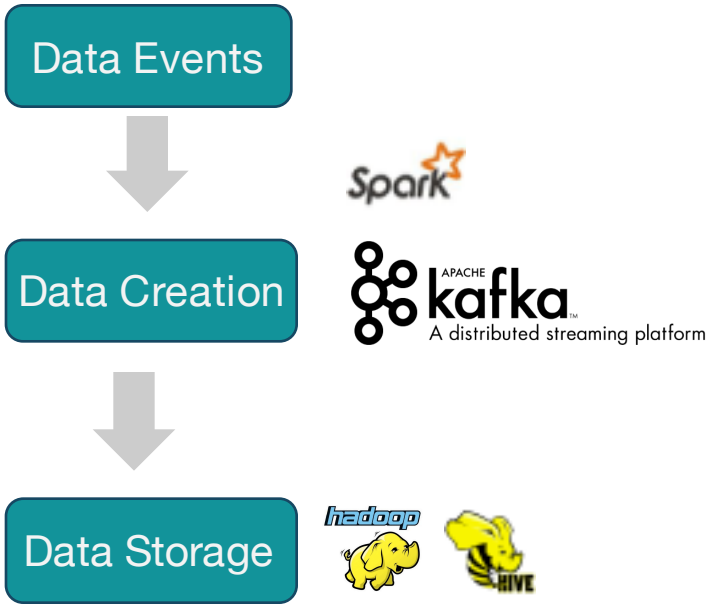
Allows insurers to

- Satisfy consumer demands
- Improve insurance pricing
- Change driving behavior and reduce accident
- However, this creates **MANY CHALLENGES**

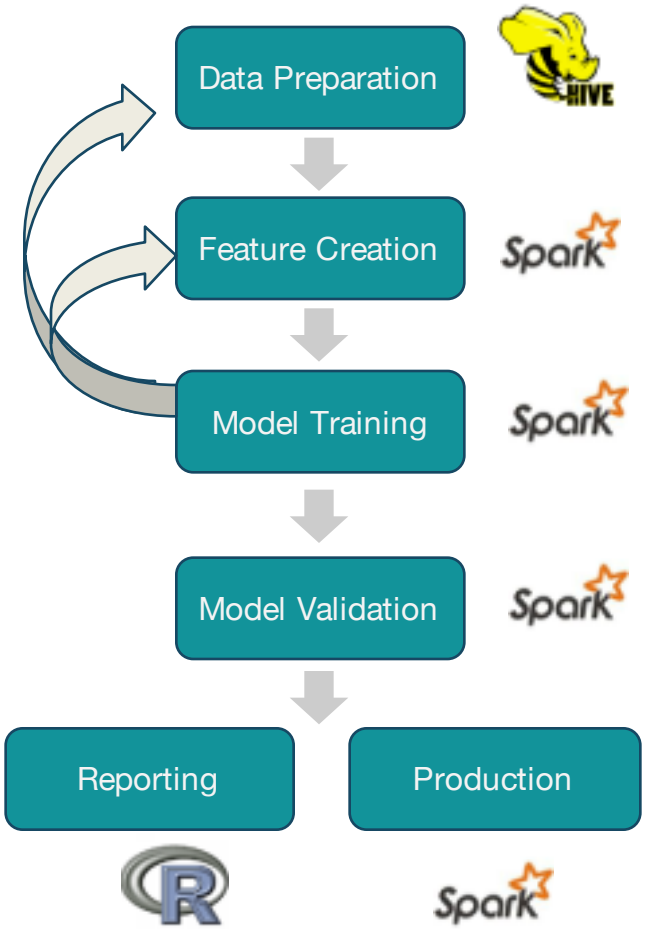
Challenge I: Big Data

- Huge volumes of data
  - Large number of trips
  - High frequency of GPS & IMU
- Two key questions:
  - How to capture & store large volumes of data?
  - How to analyze big data?

Data Management Pipeline

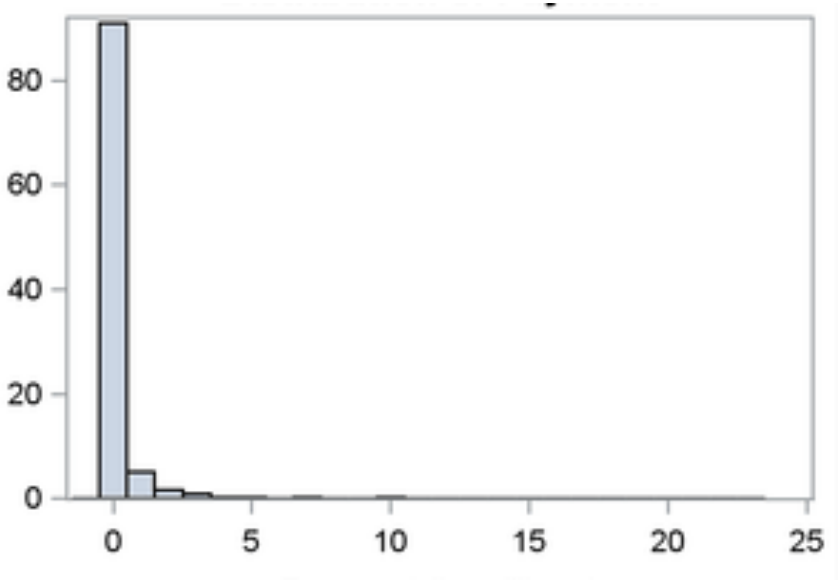


Data Analytics Pipeline



Challenge II: Extreme Sparsity

- Claims are rare events
  - Even rarer on trip level
  - More than 99.9% zeros



- Use Tweedie Compound Poisson distribution
  - Spike at zero
  - Continuous on positives

$$Y = \sum_{i}^{T} X_i$$

$$T \sim Pois(\lambda),\ X_i \overset{\text{iid}}{\sim} Gamma(\alpha, \gamma),\ T \perp X_i.$$

Revolutions

Daily news about using open source R for big data analysis, predictive modeling, data science, and visualization since 2008

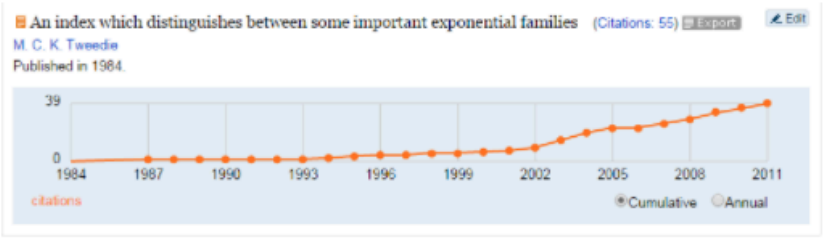
« In case you missed it: September 2014 Roundup | Main | 14 Reasons Why R is better than Excel »

October 09, 2014

A Note on Tweedie

by Joseph Rickert

In a recent post I talked about the information that can be developed by fitting a Tweedie GLM to a 143 million record version of the airlines data set. Since I started working with them about a year or so ago, I now see Tweedie models everywhere. Basically, any time I come across a histogram that looks like it might be a sample from a gamma distribution except for a big spike at zero, I see a candidate for a Tweedie model. (Having a Tweedie hammer makes lots of things look like Tweedie nails.) Nevertheless, apparently lots of people are seeing Tweedie these days. Even the scolarly citations for Maurice Tweedie's original paper are up.



[CITATION] An index which distinguishes between some important exponential families

MCK Tweedie - Statistics: Applications and new directions: Proc. Indian ..., 1984

Cited by 333 Related articles Cite Save

Belongs to the exponential dispersion family: $Var(Y) = \phi \cdot \mu^p$

The Tweedie distributions include a number of familiar distributions

- normal distribution, $p = 0$,
- Poisson distribution, $p = 1$,
- compound Poisson–gamma distribution, $1 < p < 2$,
- gamma distribution, $p = 2$,
- positive stable distributions, $2 < p < 3$,
- inverse Gaussian distribution, $p = 3$,
- positive stable distributions, $p > 3$, and
- extreme stable distributions, $p = \infty$.

For $0 < p < 1$ no Tweedie model exists.

[SPARK-18929][ML] Add Tweedie distribution in GLM #16344

Edit

Closed actuaryzhang wants to merge 24 commits into `apache:master` from `actuaryzhang:tweedie`

Conversation 124 | Commits 24 | Files changed 2 | +571 −85

actuaryzhang commented on Dec 19, 2016 • edited — Contributor

What changes were proposed in this pull request?

I propose to add the full Tweedie family into the GeneralizedLinearRegression model. The Tweedie family is characterized by a power variance function. Currently supported distributions such as Gaussian, Poisson and Gamma families are a special case of the Tweedie https://en.wikipedia.org/wiki/Tweedie_distribution.

@yanboliang @srowen @sethah

actuaryzhang added some commits on Dec 15, 2016

Add Tweedie family to GLM	952887e
Fix calculation in dev resid; Add test for different var power	4f184ec
Merge test into GLR	7fe3910

Reviewers

srowen

yanboliang

Assignees

No one assigned

Labels

None yet

Projects

None yet

Milestone

SPARK SUMMIT 2017

Challenge III: Dependency

- Repeated measures lead to correlated data
- Need hierarchical/random-effects models
  - Better inference
  - Regularization

$$y_i \sim \mathrm{Tw}(\mu_i, \phi, p), \tag{1}$$

$$\eta(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta} + u_{g_i}, \tag{2}$$

$$u_j \sim N(0, \sigma^2). \tag{3}$$

Large-scaled Random Effects Model

	L2 Regularization	Numerical Integration	Bayesian Formulation
Estimation	Blockwise coordinate decent	Laplace approximation + LBFGS	Markov chain Monte Carlo
Pros	- Easy to implement	- No hyperparameters - Well established theory	- Easy to implement - No hyperparameters - Well established theory - Bayesian predictive distribution
Cons	- Hard/impossible to choose hyperparameters - No standard errors	- Spark lacks sufficient support on sparse matrix factorization	- Requires vast computation time - Not well suited for large analysis
Spark implementation	Zhang et al 2016 KDD		Zhang 2017, ASTIN Bulletin

Bayesian Analysis of Big Data

- To appear at *ASTIN Bulletin*
- Presents efficient Bayesian computation via distributed computing
- Boosts speed by 65 times over standard method
- Expands scope of applicability of Bayesian methods in practice
- Applies Bayesian hierarchical Tweedie model to 13M records
- Demonstrates value of Bayesian methods in large-scaled insurance predictive modeling

Existing Research

- Improve sampling efficiency & convergence speed
  - E.g., Langevin drift, Hamiltonian MCMC, Adaptive MCMC
  - Reduce # iterations needed in Markov chain
  - But still cannot handle big data
- Develop scalable algorithm using parallel processing
  - Parallel computation of data likelihood
  - Stochastic approximation of gradient
  - Parallel independent MCMC on data partitions

Distributed Bayesian Simulation

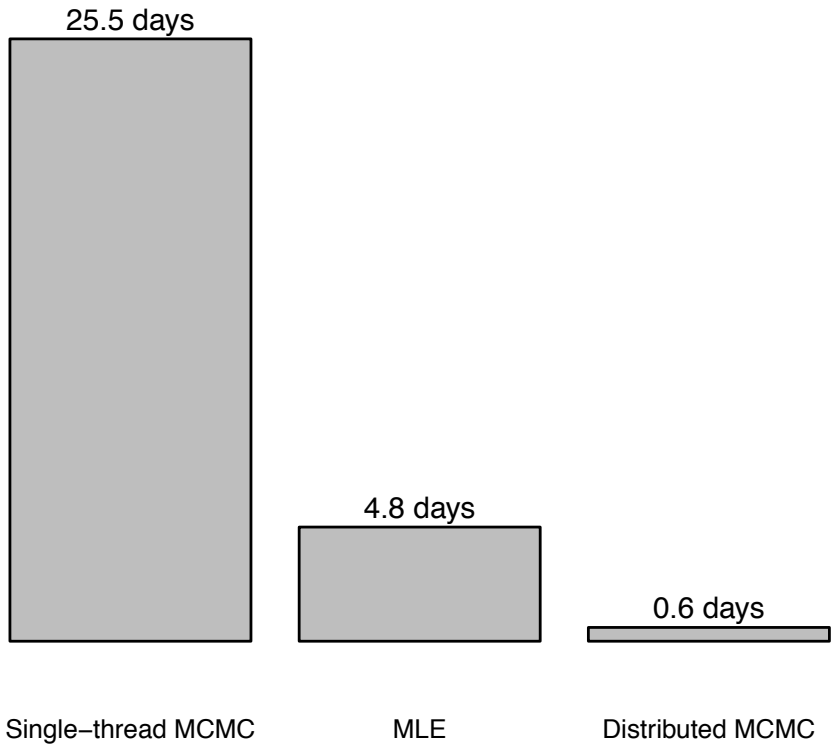
- Distributed computing of data likelihood
  - Divide large task into smaller splits
  - Process many splits simultaneously by separate processors
  - Can use large number of processors

Heavy computation!

Empirical Analysis

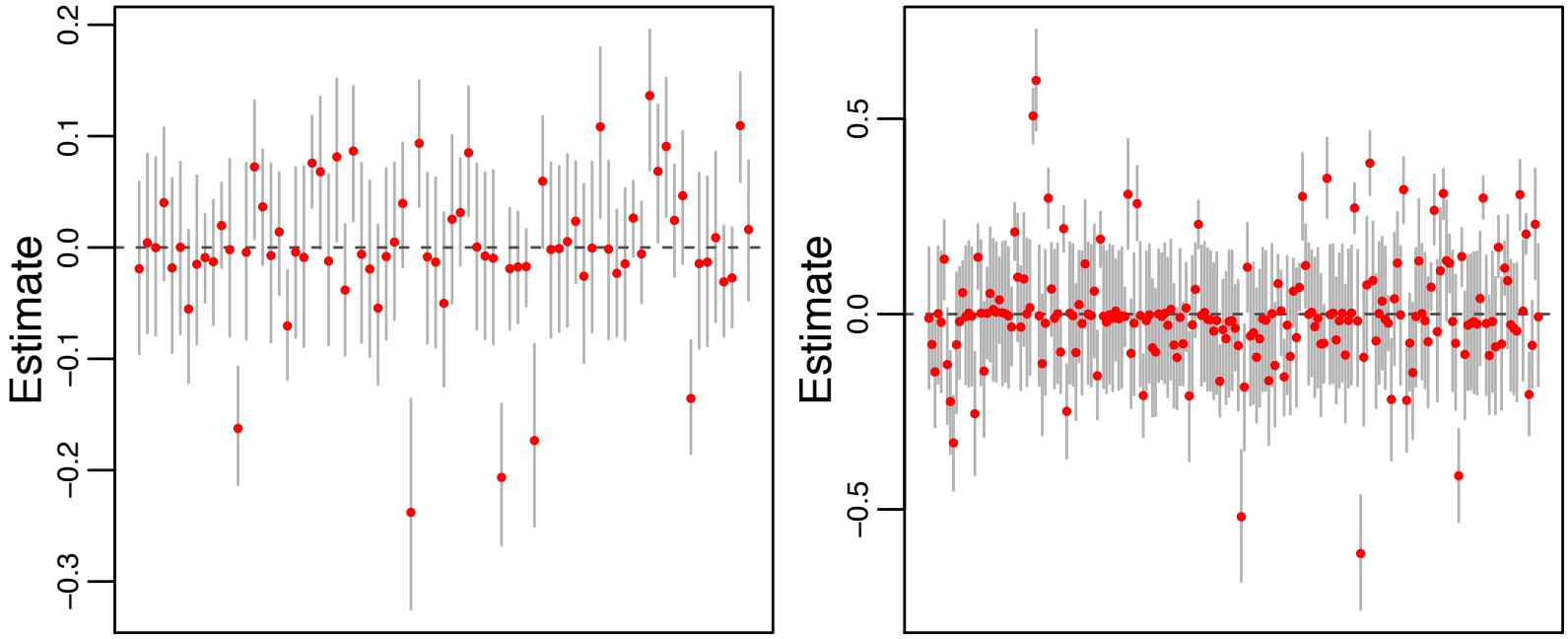
- Public data
  - Personal auto BI coverage
  - 13.2 million records (1GB)
  - 99% records no payments
  - Car make, model, and other predictors (values masked)
- Fit Bayesian hierarchical Tweedie model
  - Tweedie for spike at zero
  - Hierarchical on car make and car model

Performance Analysis



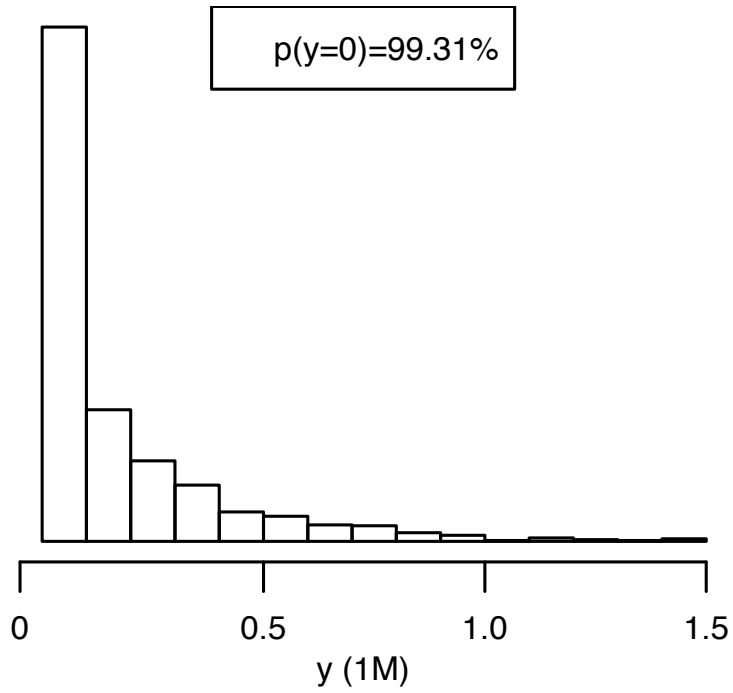
- MCMC with 20,000 iterations
- Single-thread algorithm infeasible
- Distributed algorithm 43x faster
- Run as over-night job
- Even faster than MLE

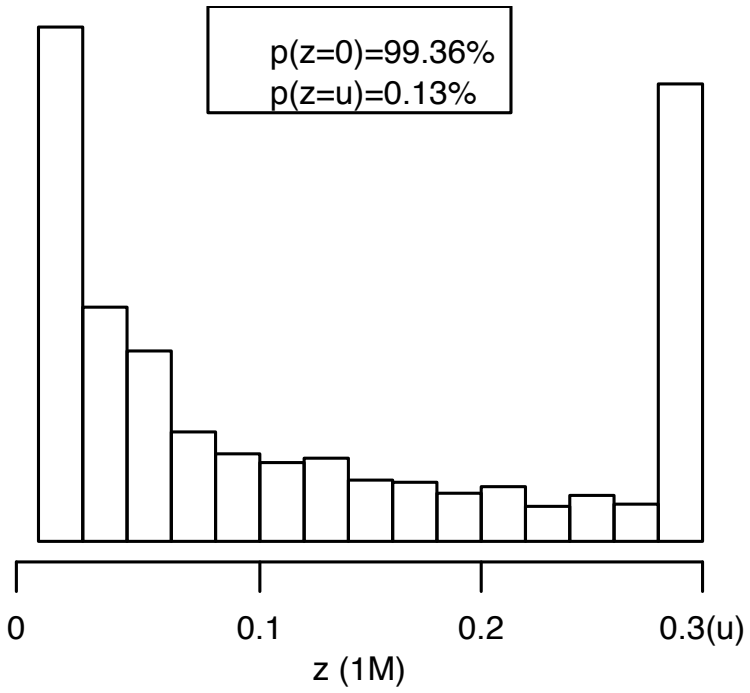
Hierarchical Modeling



Prediction Under Policy Modifiers

- Policies issued with deductibles and/or limits
- Leads to modified loss distribution $E(y \wedge u) = \int_0^u S(y)dy,$
- Easy to derive using Bayesian methods
  - For each posterior sample, simulate predicted loss
  - Apply modifiers to predicted loss
  - Summarize using modified predictive distribution
    - Mean loss cost under policy modifiers


p(y=0)=99.31%
0
0.5
1.0
1.5
y (1M)


p(z=0)=99.36%
p(z=u)=0.13%
0
0.1
0.2
0.3(u)
z (1M)

Summary & Conclusion

- Challenges in practical UBI analytics
  - Computational burden of big data
  - Extreme rare events
  - Correlated data from repeated measures
- Suggests distributed Bayesian methods on Spark
- Valuable in many areas of data analytics

Thank You.

Email *actuary_zhang@hotmail.com* if you have any questions.