A risk management toolkit for actuaries ?

Xavier Conort – Gear Analytics
June 2011

Agenda

- **Increase in computing power** and competitive pressure have **transformed** actuarial science and **the way to do business** in many industries.
- Predictive modelling in insurance **creates opportunities but** increases **complexity and pricing risk**.
- **Reserving risk** is another key concern for insurers.
- **How a statistical software like R can help** the actuary to manage pricing and reserving risks ?

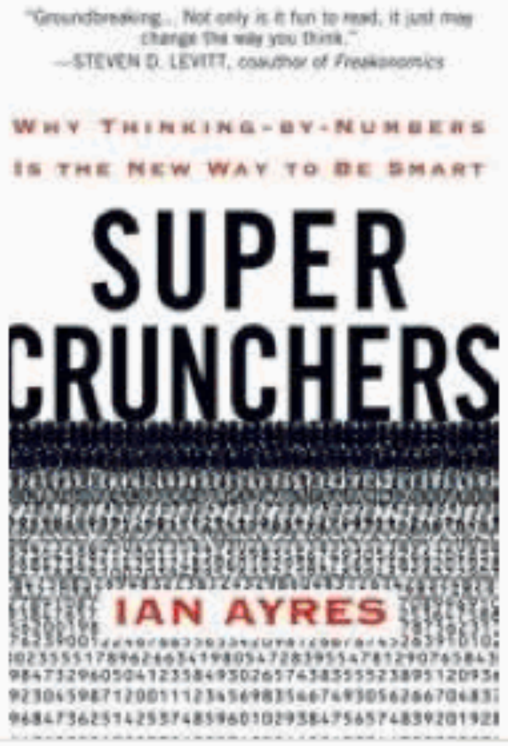
Wikipedia, entry: 'Actuarial science'

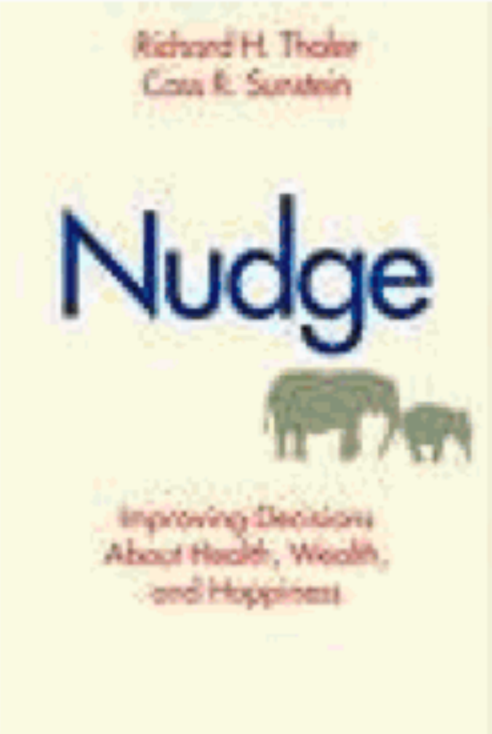
Wikipedia *: "**Actuarial science includes a number of interrelating disciplines**, including probability and statistics, finance and economics. Historically, actuarial science used deterministic models in the construction of tables and premiums*. ***The science has gone through revolutionary changes during the last 30 years due to the proliferation of high speed computers** and the synergy of stochastic actuarial models with modern financial theory."*

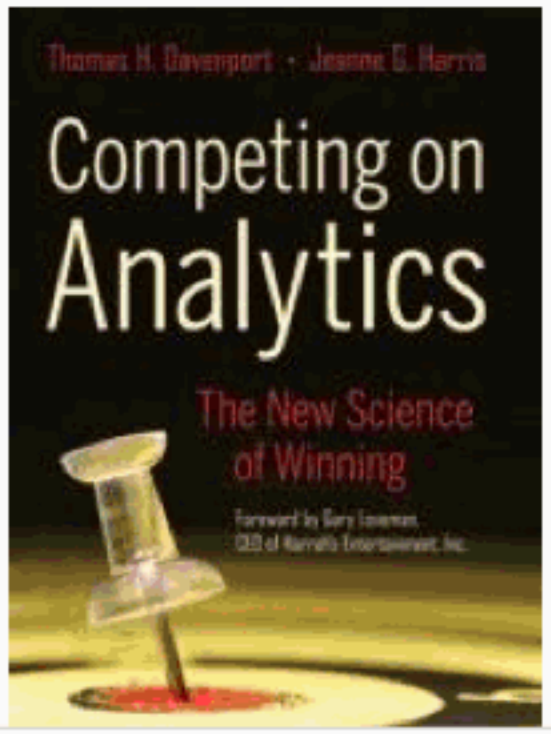
Increase in computing power : predictive modelling is everywhere!

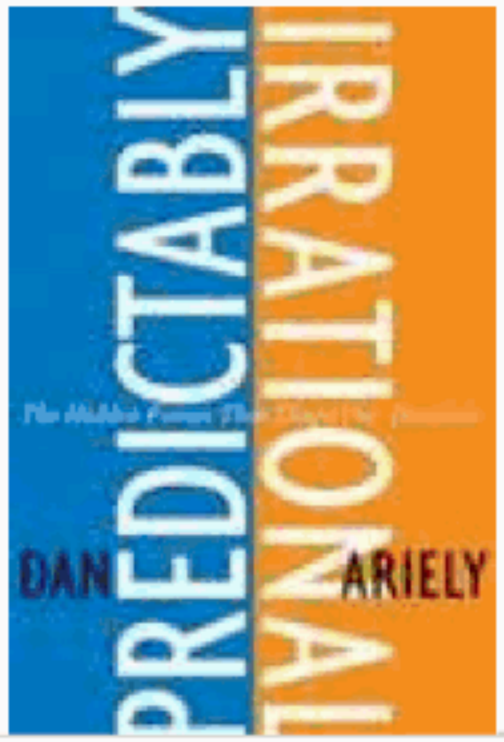
Most industries have taken advantage of increasing computing power and better data:

- Insurers use predictive models to underwrite risk
- Financial institutions determine credit score when you want a loan;
- The post office uses them to decipher your handwriting;
- Meteorologists to predict weather;
- Retailers to decide what to put on their shelves;
- Marketers to improve their products;
- They are even used by sports teams to pick players.









In insurance, predictive modelling is widely used for classification ratemaking

Predictive modelling uncovers opportunities, protects against adverse selection **but increases complexity** each time a rating variable is added.

When specific risk segments are identified as inadequately priced, it represents for a company either:

AN OPPORTUNITY

Use information to attract and select the lower-risk insured

A RISK OF ADVERSE SELECTION if competition is first recognizing it

The company who fails to recognize the inadequacy, will attract and retain the higher-risk insureds and lose the lower-risk insureds to other competing companies where lower rates are available

Implement a new rating variable to price appropriately

See "Basic Ratemaking" by Werner and Modlin available in CAS website to read more on favorable and adverse selection produced by classification ratemaking

More complexity in ratemaking leads to more model risks and operational risks

Risks	How to address them
ESTIMATION RISK (less information available for individual risk segment than for an average rate for all risks)	Use multivariate statistical techniques and smoothing techniques to remove noise from data and identify key risk factors.
SPECIFICATION RISK (limitations in model assumptions, unlike CLT which fits all)	Know model limitations. Relax linear regression assumptions (normal error) by working on transformed data or using GLMs. Relax GLMs assumptions (iid, link-linear, exponential family) by using GAMLSS, GLMM, GEE…
SYSTEM RISK (difficulty to choose the right model and user bias)	Use diagnostic tools, compare models and techniques, document and justify choices.
IMPLEMENTATION COST (due to the complexity of rating model)	Test the model with holdout sample to avoid over-fitting. Make it as simple as possible to balance system cost and benefits of having additional parameters in the rating algorithm.
LAPSE RISK	Model demand side to predict likely behaviour of customers
DEVIATIONS FROM EXPECTATION	Monitor deviations of premium rate from technical price. Restate historical experience to factor expected trends. Quantify prediction errors (estimation and process errors) and retain only risks within tolerance.

Another key driver for more advanced modelling in insurance = **reserving risk**

- **Until early 2000s, not much had been done** to quantify the magnitude of the potential deviations from "Best Estimates" (BEs). **Since then, research on stochastic techniques has been very active.**

- The widest spread techniques are **Chain Ladder based** (Mack, Over-Dispersed Poisson).
  - ⇒ But differ from determinist practices where judgment is a key element in setting BEs

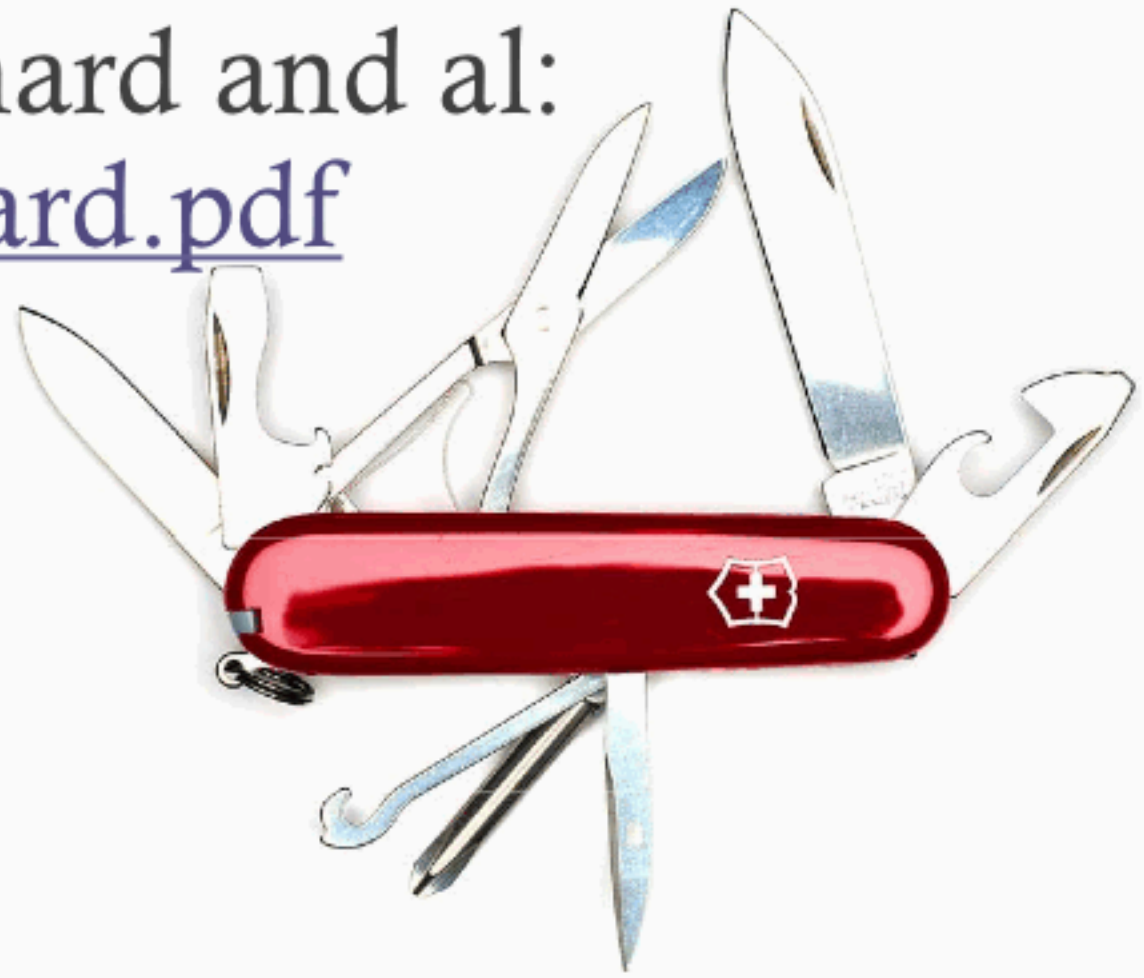
- **Bayesian techniques** are emerging to integrate actuarial judgment into the stochastic framework.
  - ⇒ But use of Markov Chain Monte Carlo techniques can be intimidating
  - ⇒ For an overview of what they can offer : "Stochastic reserving using Bayesian models – can it add value?" *presented by Francis Beens, Lynn Bui, Scott Collings and Amitoz Gill at The Institute of Actuaries of Australia's GI Seminar in November 2010*

- **Micro-level stochastic loss reserving** (individual claims run-off) is an interesting alternative which attempts to make better use of all information available.
  - ⇒ Choice of the CAS Loss Simulation Model Working Party for their open-source model
  - ⇒ May be more intuitive but claim handling is a complex process to model and fit

Is Microsoft Office still an appropriate tool?

- In 2006, **the Actuarial Toolkit Working Party** of UK Institute of Actuaries **likened the Microsoft Office suite for actuaries to a Swiss army knife for a dentist.**

- It can do most of the job but you would rather choose a dentist with a better tool.
  - "An actuarial tookit" by Trevor Maynard and al: http://toolkit.pbworks.com/f/Maynard.pdf



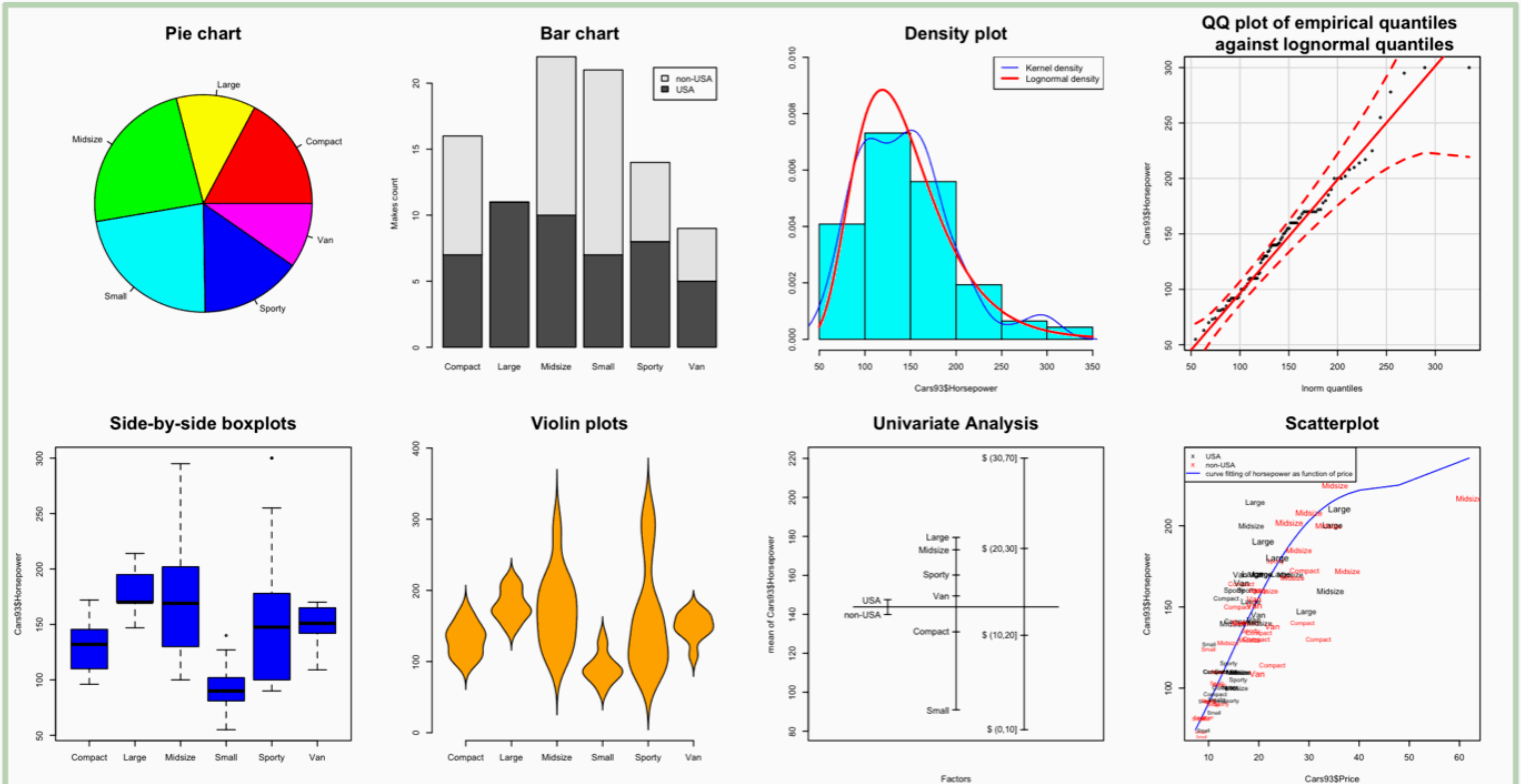
Features of a more appropriate tool for actuaries

- **A more appropriate tool** shall include
  - a **wide variety of graphical and statistical techniques** (GLM, GLMM, GAMLSS, CART, survival models…) to visualize patterns, fit models, identify key risk factors and produce diagnostics about the certainty of results and the appropriateness of the model fitted.
  - a platform to implement **stochastic reserving** techniques.
  - a **simulation** framework where a large range of distributions can be simulated and dependencies between simulated variables can be applied in several ways, including the use of copulas.
- We will present here R which addresses all the needs listed above, but other solutions such as SAS, EMB suites or other statistical tools are also appropriate candidates

What is R ?

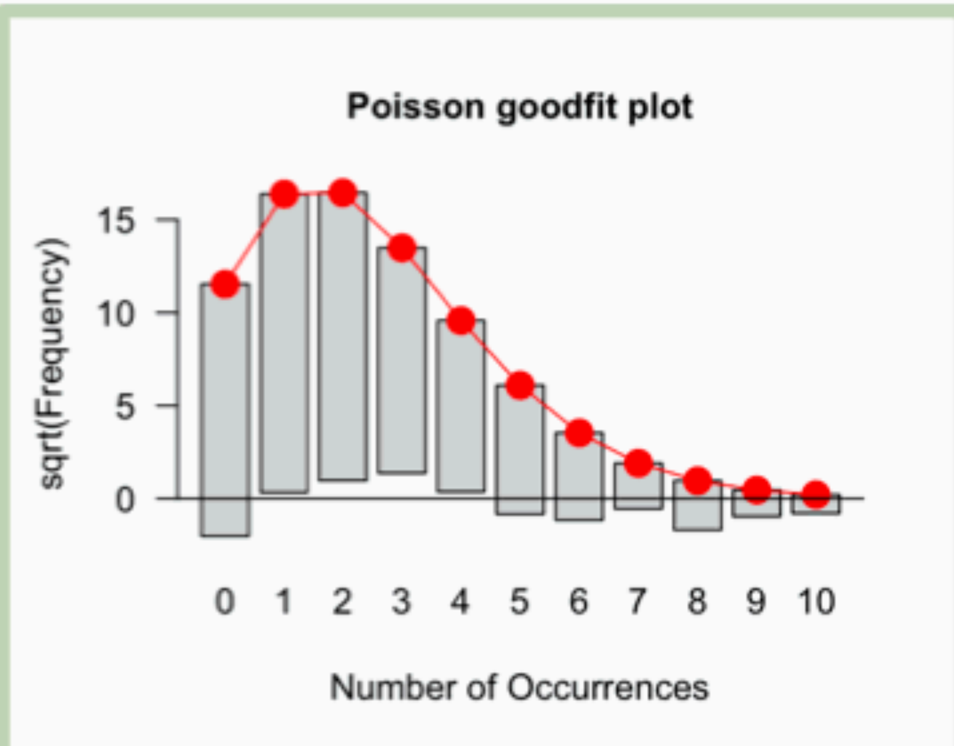
- It is open-source and **free** to all. www.r-project.org/
- It is the most common statistical package used in **universities** and more and more students in Actuarial Science are trained in R
- It is gaining exponential popularity in a **wide variety of industries**, including insurance, pharmaceuticals, finance, telecom, websites and oil and gas. Google, Merck, Shell, Bell, AT&T, Intercontinental, Oxford and Stanford are among R's benefactors and donors.
- It has attracted the attention of key **actuarial institutions** in Europe and North America who have already run and promoted courses on R.
  - http://www.actuaries.org.uk/events/one-day/predictive-modelling-using-r-fully-booked
  - http://www.the-actuary.org.uk/827719 (R u Ready?)
  - http://actuaries.zynex.ch/00_home/willkommen.htm/Programme_SAA_En.pdf
  - http://www.casact.org/education/webinar/2010/index.cfm?fa=rintro
  - http://www.caritat.fr/formation-statistique-assurance-logiciel-R-295.html

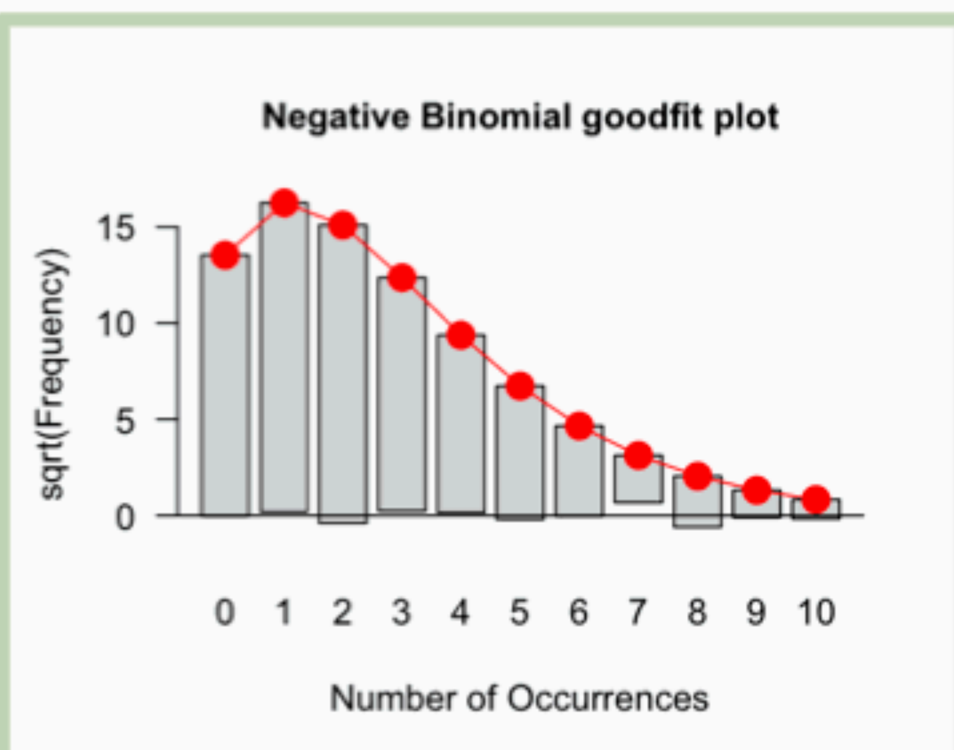
Visualize data



R script for this slide in appendix

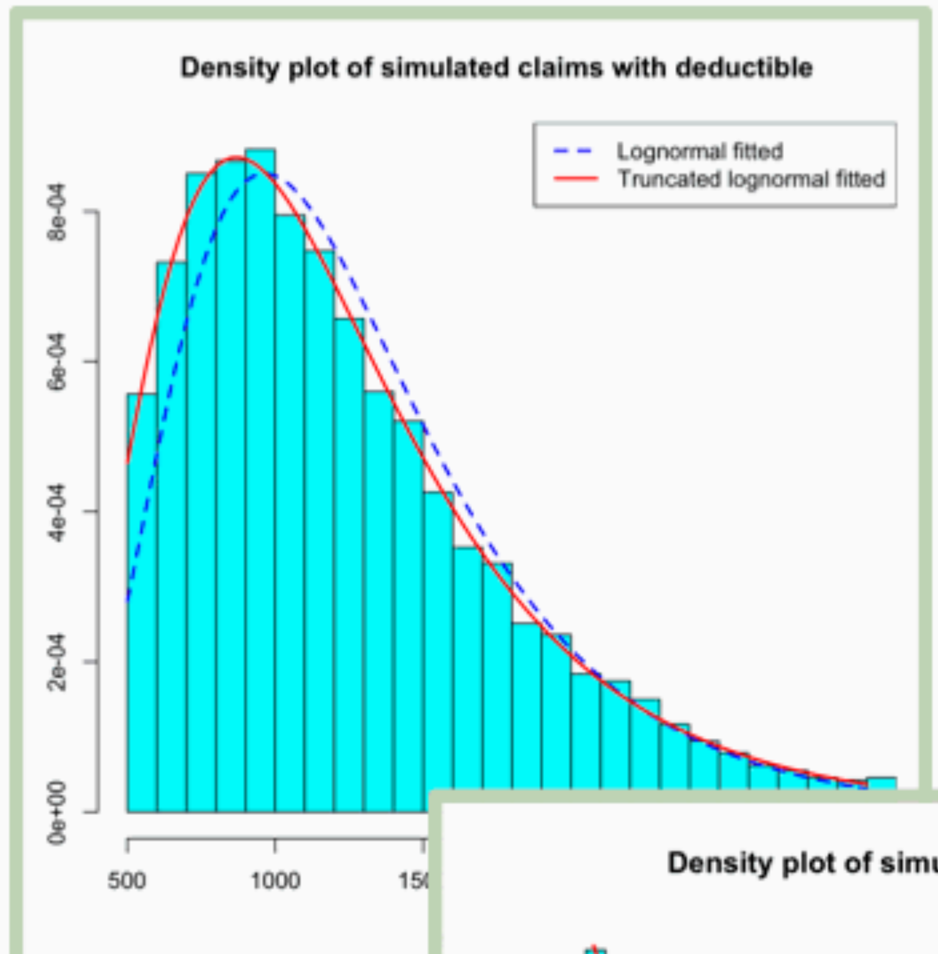
Fit large range of distributions

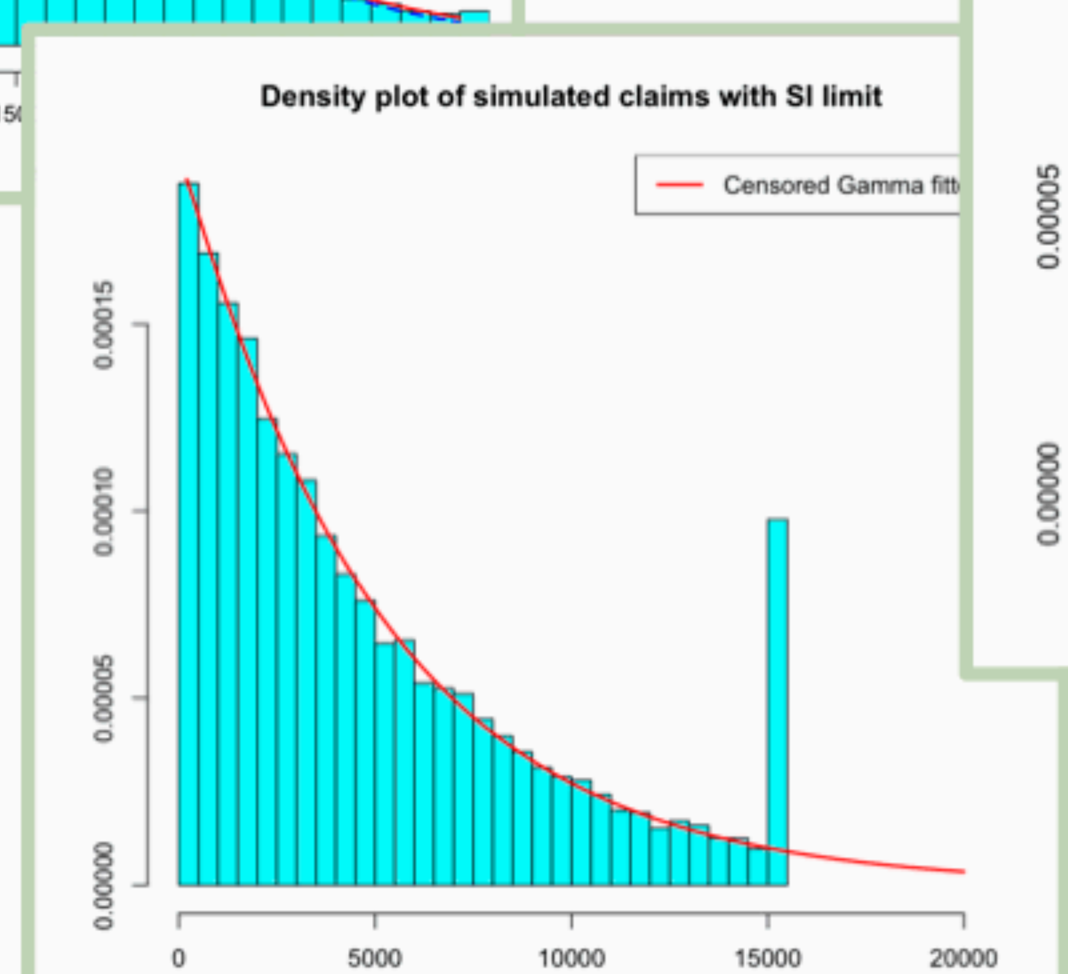




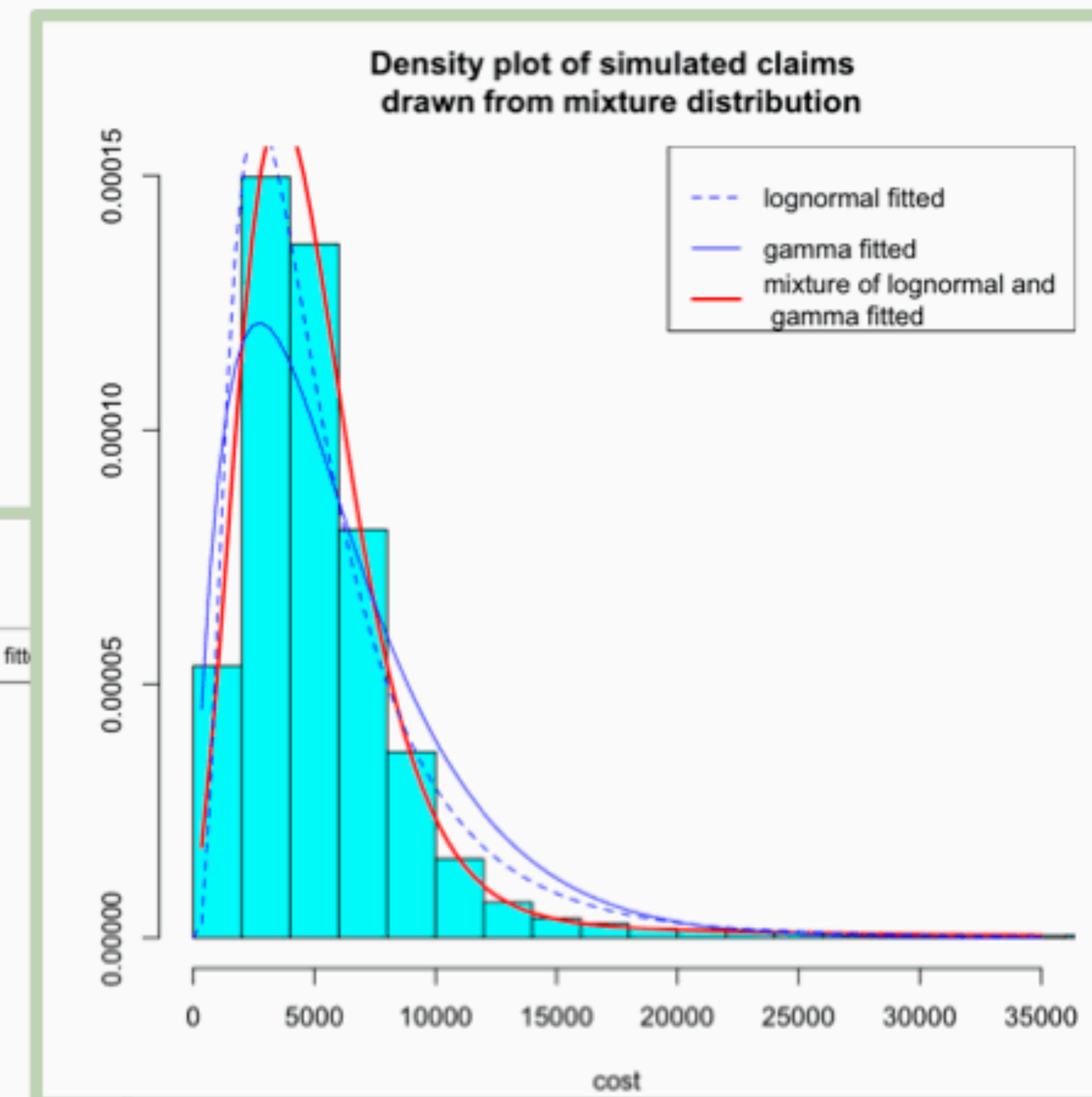
Over-dispersed count data

Truncated data (deductibles)



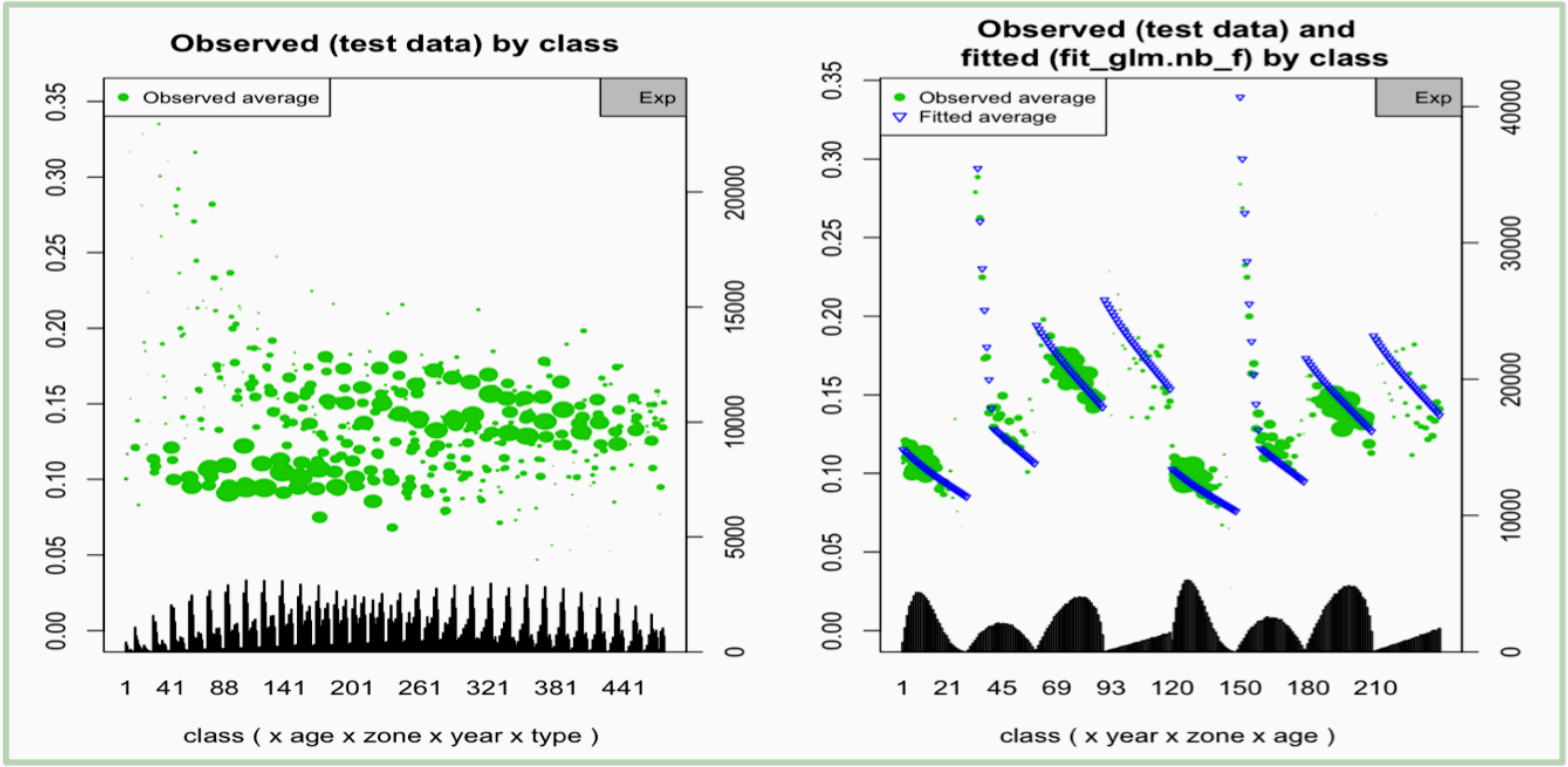


Mixed data (typical and extreme)



Censored data (SI limits)

Fit GLMs to **remove noise** & capture signal of multidimensional data



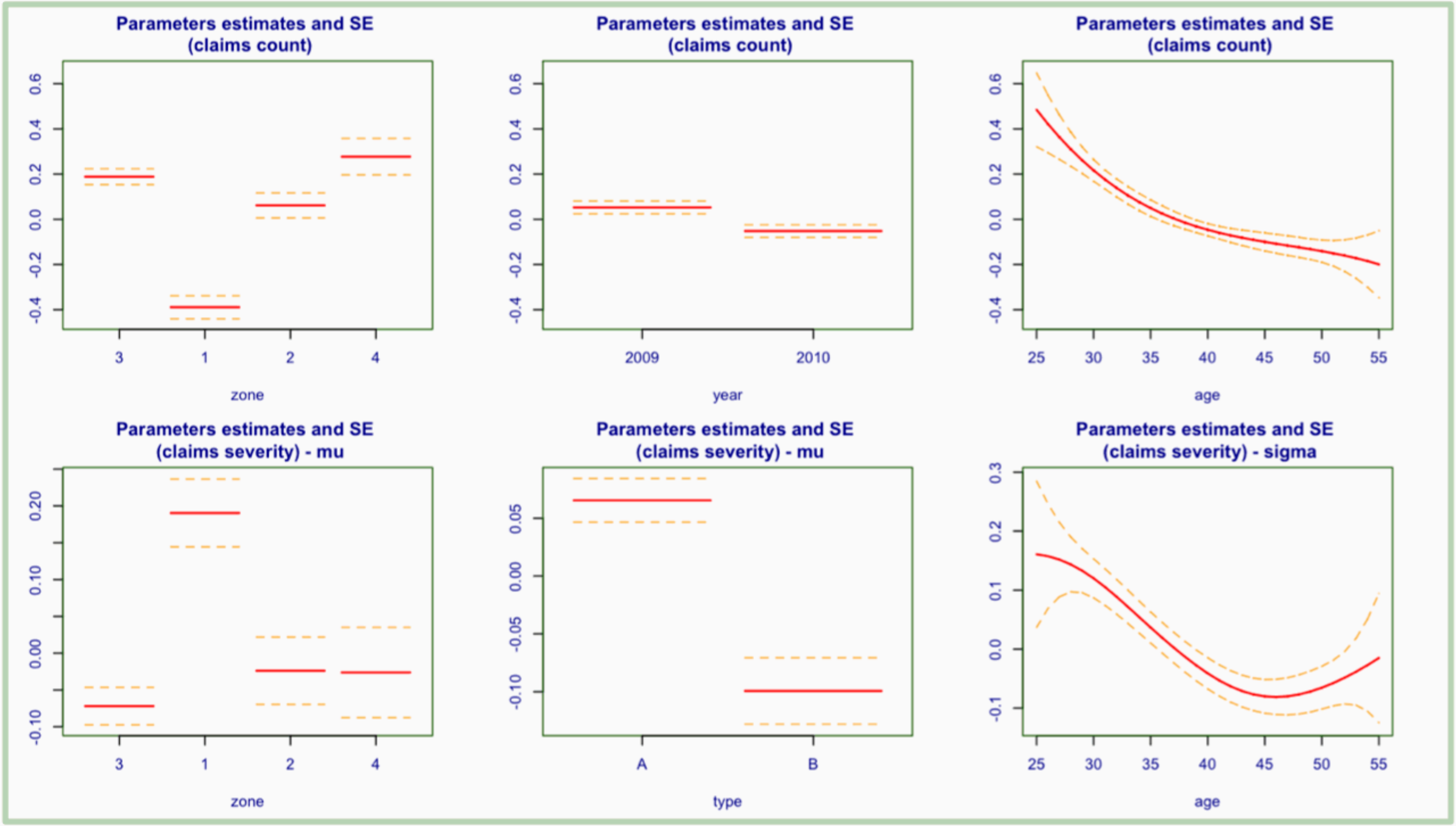
Identify **key risk factors**

Null hypothesis: models with and without a factor have the same statistical significance

```
> dropterm(fit_glm.nb,test="Chisq")
Single term deletions

Model:
N ~ zone + type + year + bs(age) + offset(log(Exp))
        Df   AIC     LRT   Pr(Chi)
<none>     38038
zone     3 38289 256.964 < 2.2e-16 ***
type     1 38036   0.171 0.6790991
year     1 38050  13.889 0.0001939 ***
bs(age)  3 38118  86.163 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> fit_glm.nb_r<- glm.nb(N ~ zone + year + bs(age) + offset(log(Exp)),
data=policy)
```

Plot parameter estimates and **standard errors**



Overcome **limitations of GLMs**

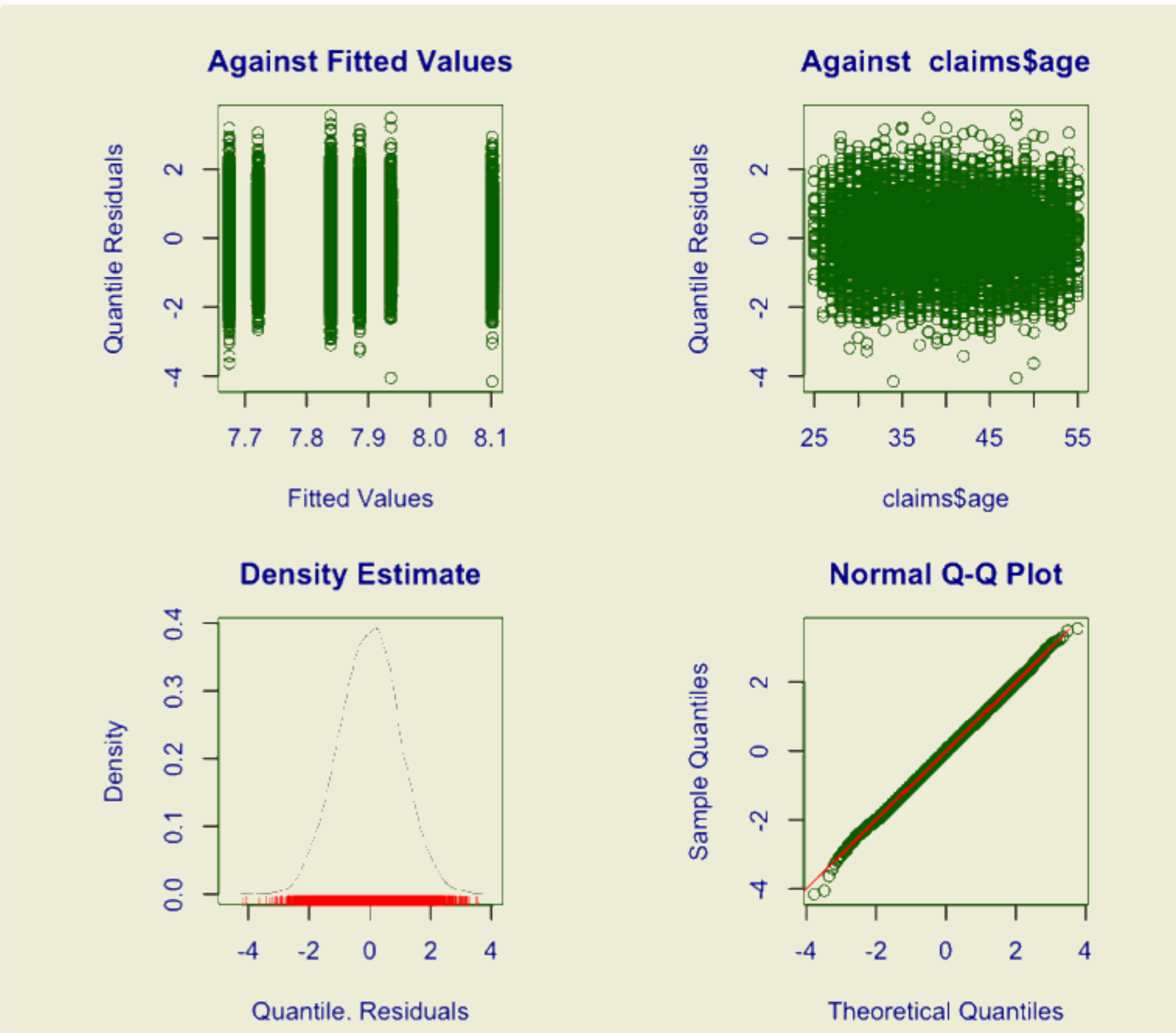
<table>
<tr><th>GLM limitations</th><th>How to address</th></tr>
<tr><td>The error structure may be different from that imposed by the choice of the exponential family</td><td rowspan="2">Fit the normal model to transformed data, but in this case, it is the median which is modeled as function of risk factors and not the mean!
Use GAMLSS techniques (Generalized Additive Models for Location, scale and shape):
exponential family distribution assumption is relaxed and replaced by a general distribution family, including truncated, censored, highly skew and/or kurtotic continuous and discrete distributions
other parameters of the distribution can be modeled as a function of risk factors</td></tr>
<tr><td>Model only the mean as a function of risk factors while the scale or shape of the distribution of the response variable might change with risk factors.</td></tr>
</table>

Overcome **limitations of GLMs**

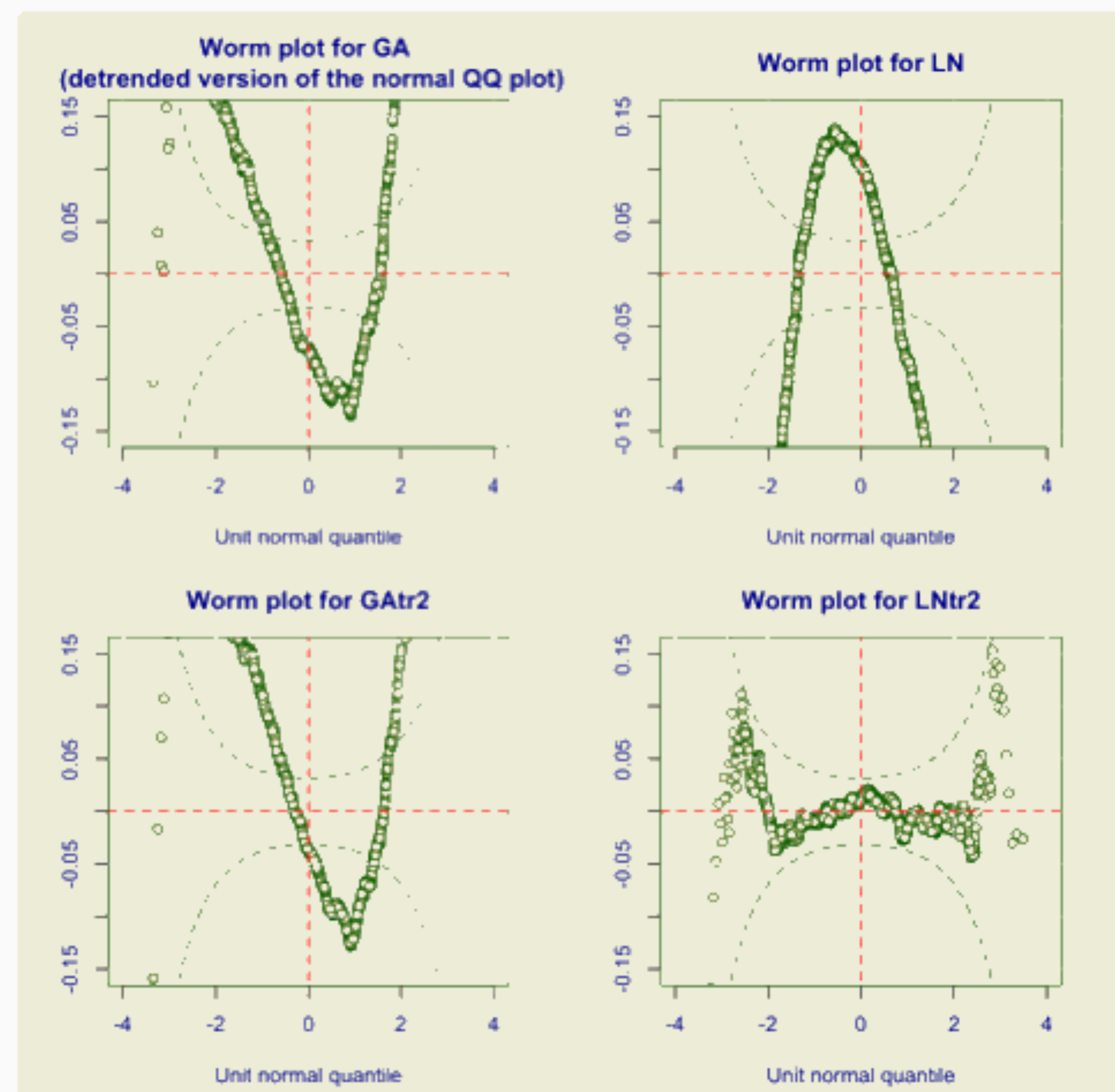
GLM limitations	How to address
In real life, **independence assumption is violated** in presence of: - repeated measurements - sub-groups of samples with high degree of correlation (e.g. patients treated in the same hospital)	Generalized estimating equations (GEEs) and Generalized Linear Mixed Models (GLMMs) are used to model these longitudinal or clustered data in the GLM framework. • GEEs modify estimation procedures to account for correlation in the data and estimate population-average effects as GLMs. • GLMMs model a transformation of the mean as a linear function of both fixed and random effects. They provide credibility estimates at the subject level (individual or group of policyholders).

Produce **diagnostics** of the appropriateness of model fitted

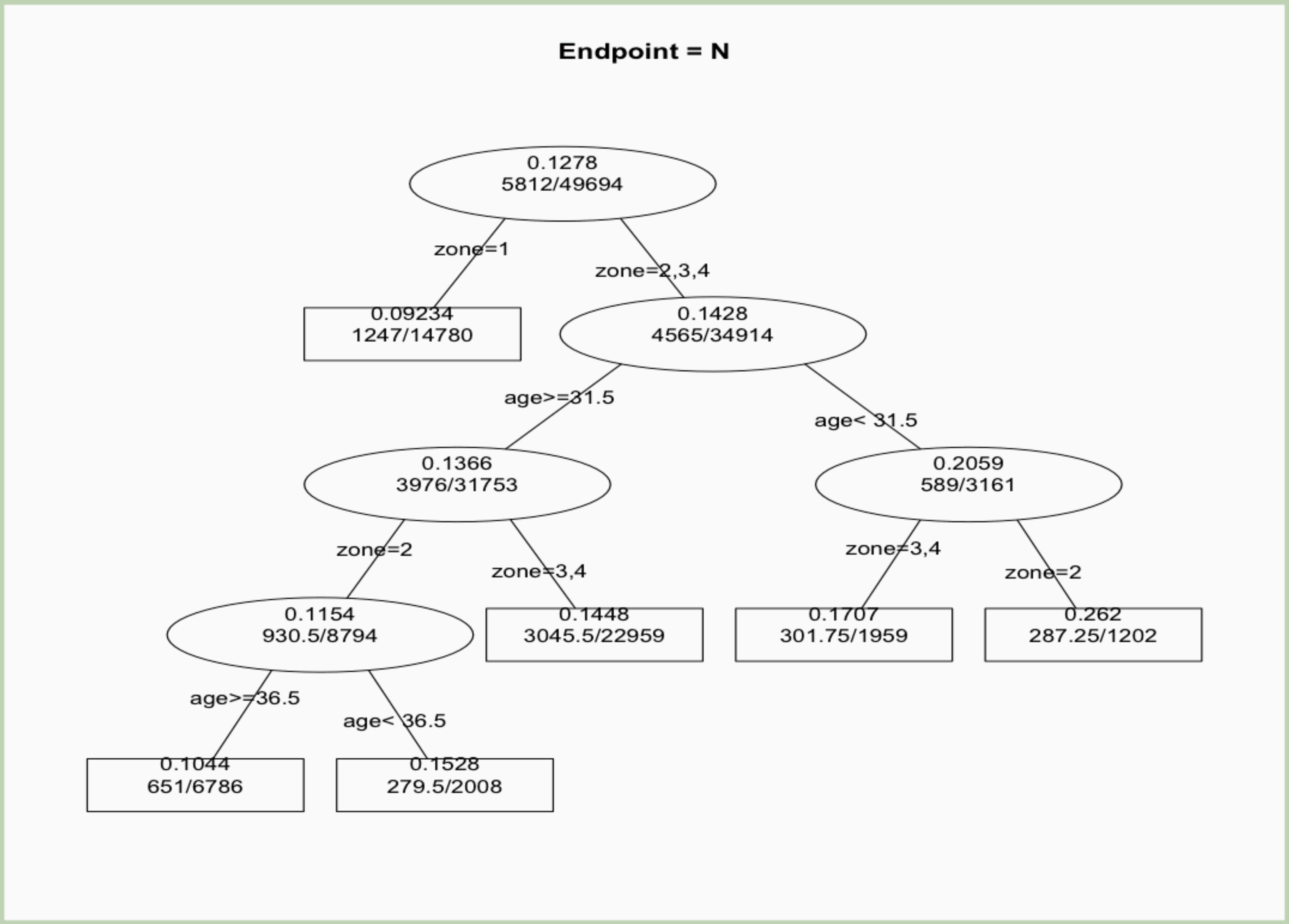
Test normality of residuals




```
> plot(fit_LNtr2_r,xvar=claims$age)
******************************************************************
           Summary of the Quantile Residuals
                           mean   =  -0.0008353087
                       variance   =  1.002379
               coef. of skewness  =  -0.01979015
               coef. of kurtosis  =  3.015783
Filliben correlation coefficient  =  0.9998495
******************************************************************
```

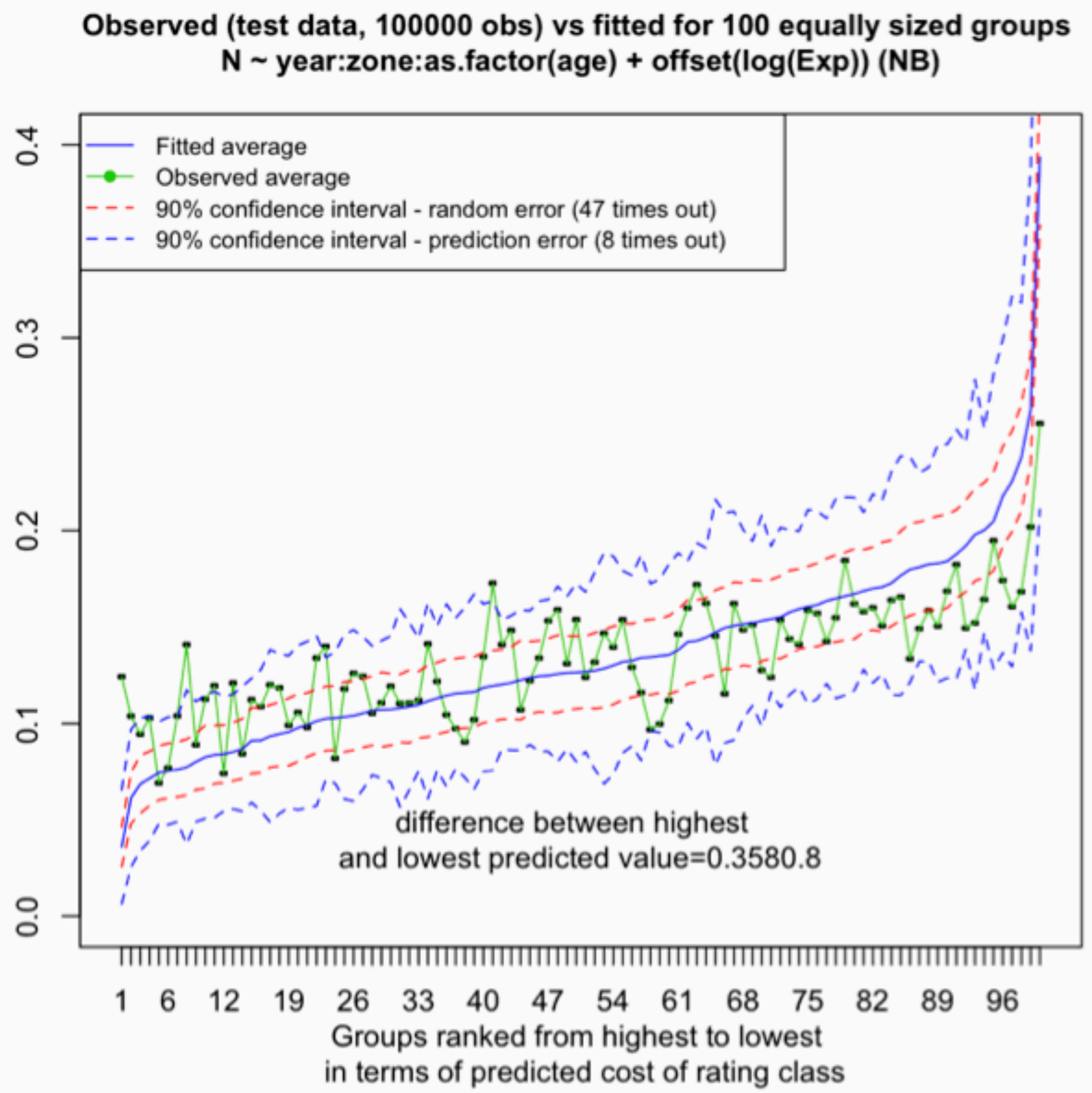


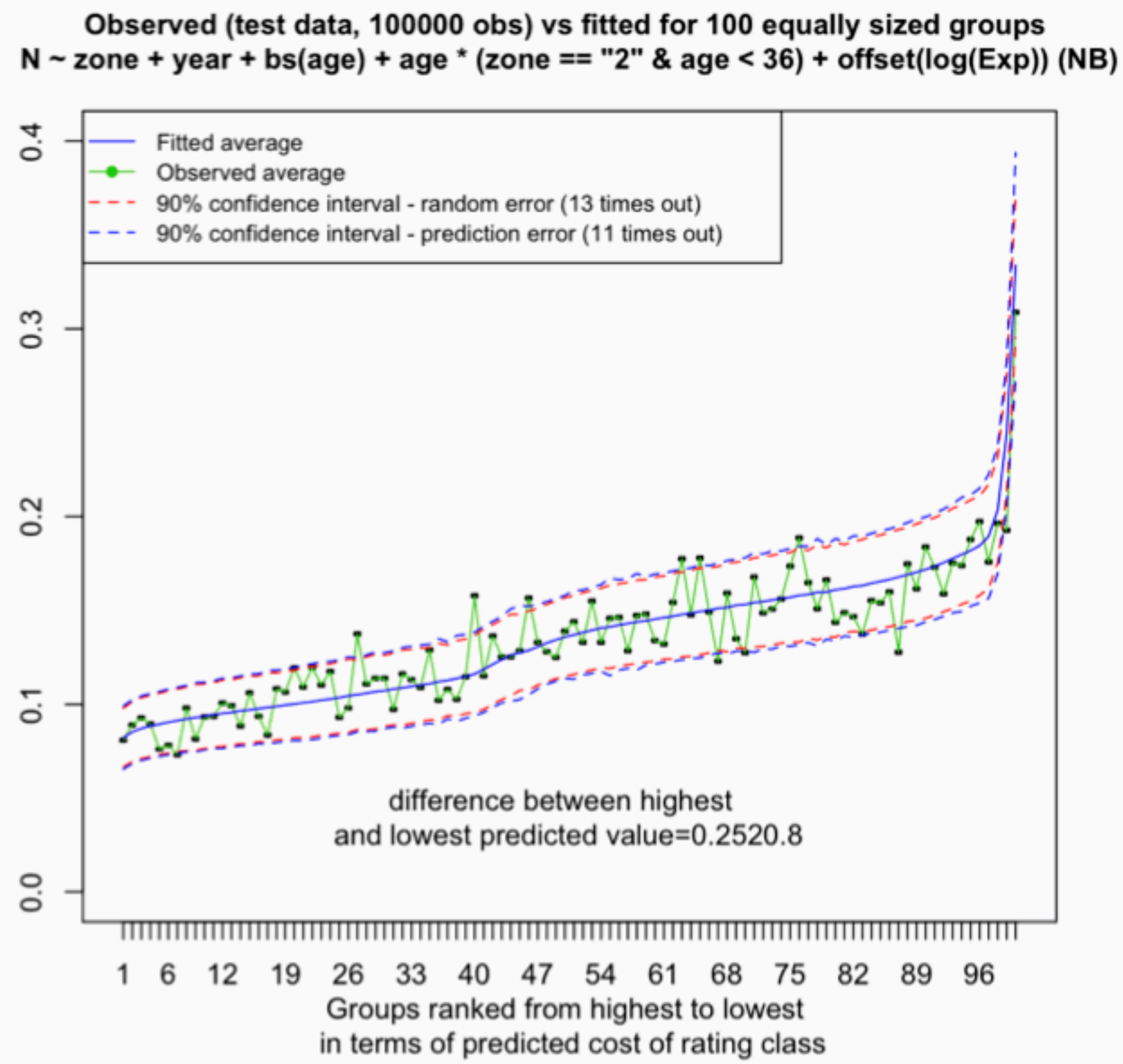
Draw trees to cross-check GLM findings and **detect local interactions**



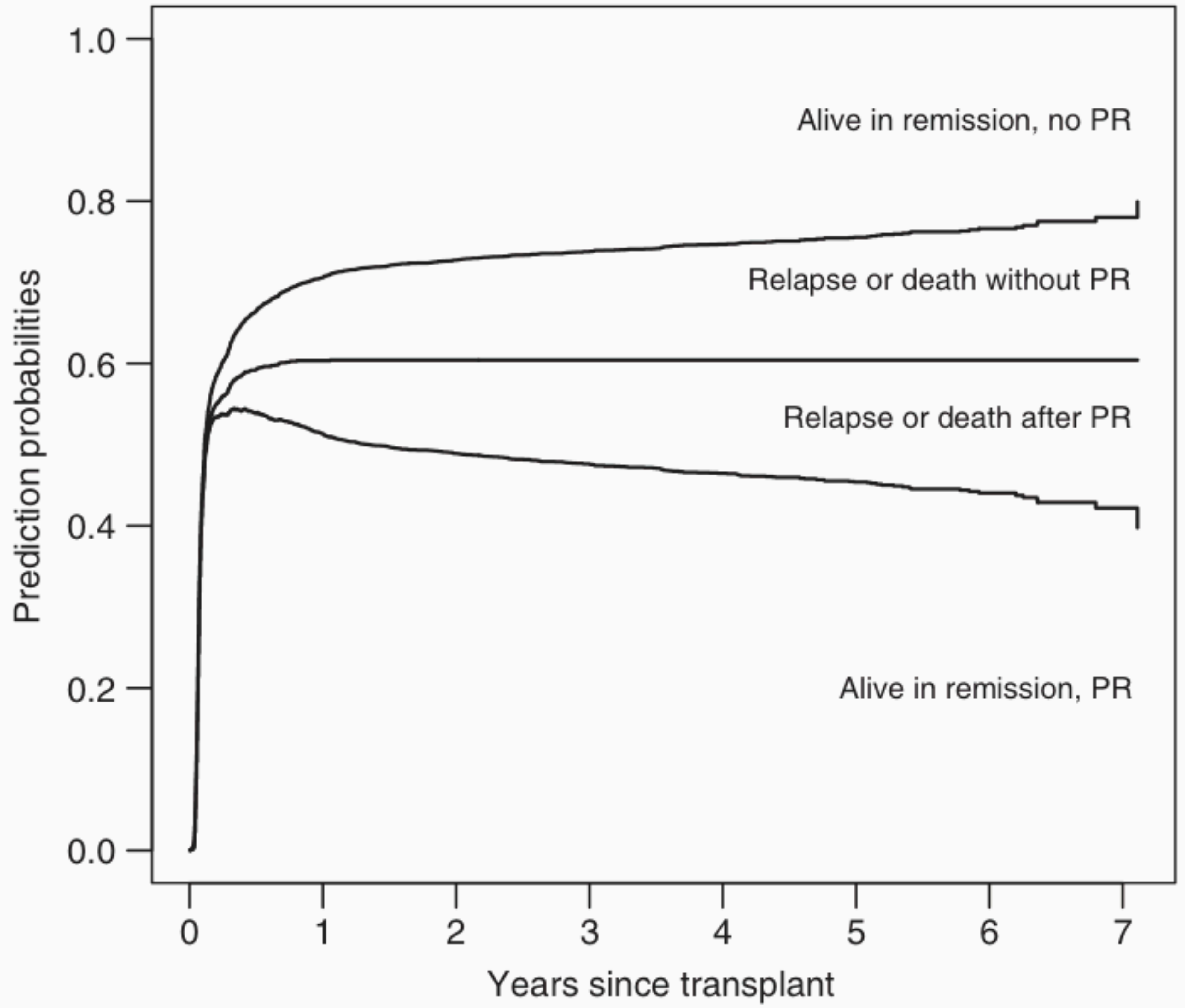
Quantify **prediction errors**
(random and parameters errors)

Based on a simulation: train data =50,000 obs to fit model and quantify errors / test data=100,000 obs





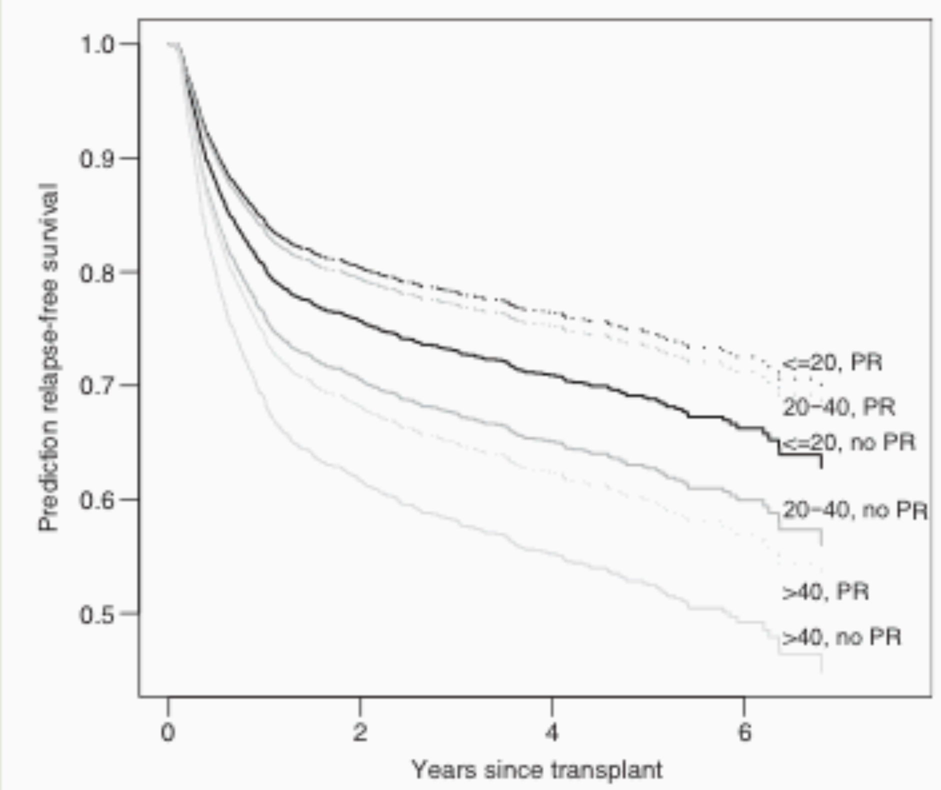
Fit survival models



Highly used in Biostatistics

But also very useful for insurance to model retention, cross-selling, report lag, claims process…

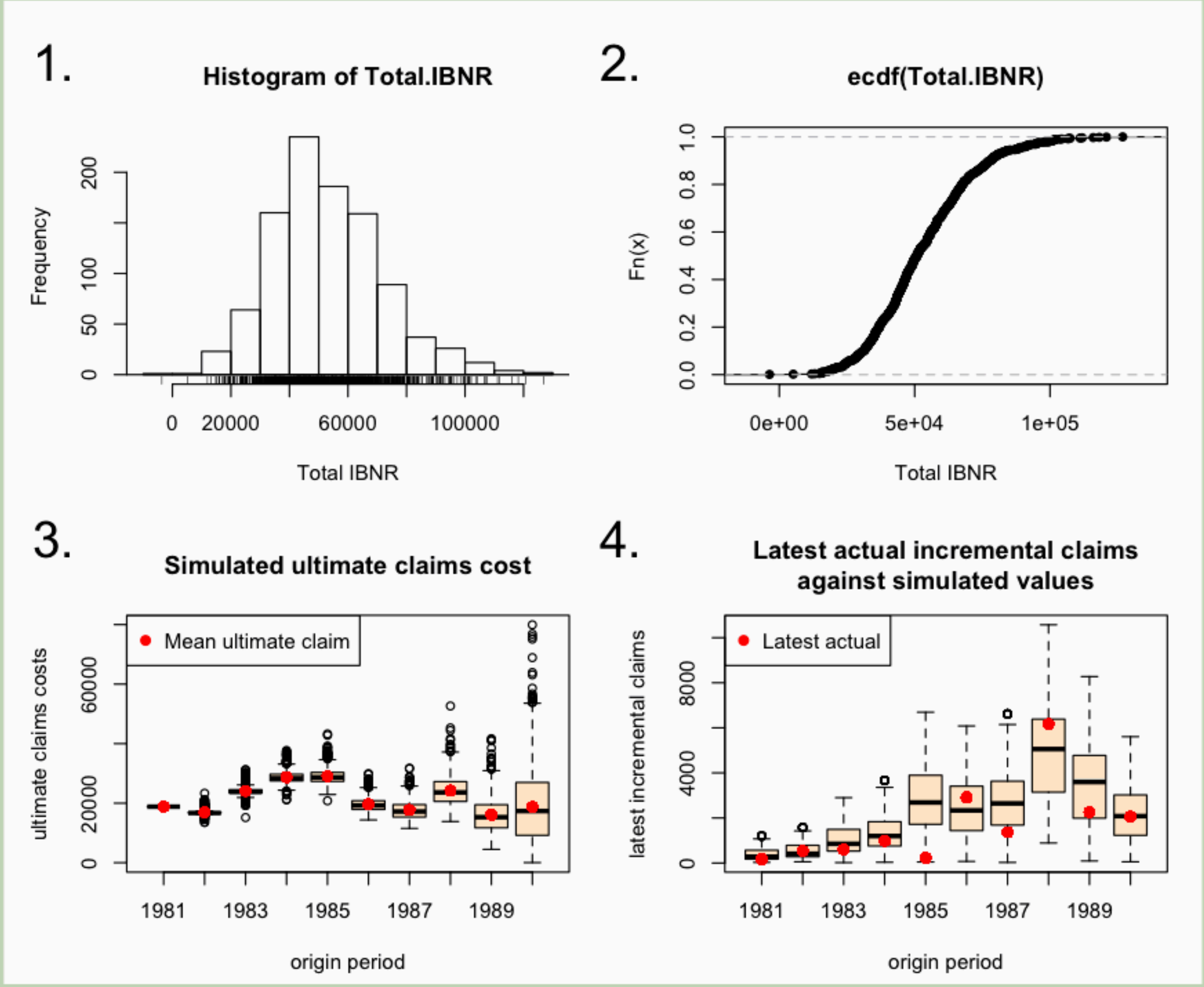
Identify **Who** is the most likely, **for What, Why** and **When**.



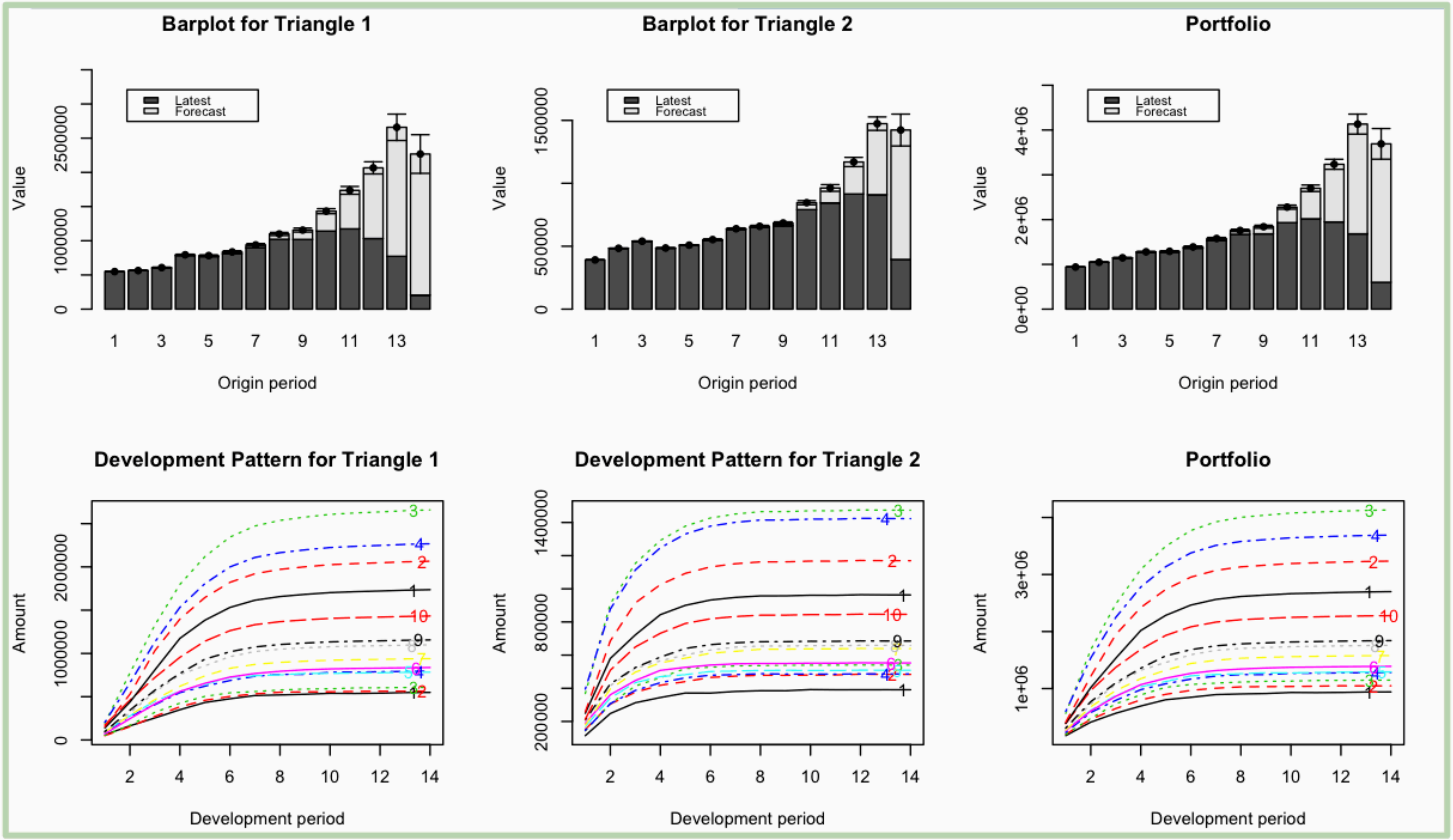
Source: Tutorial in biostatistics: Competing risks and multi-state models
H. Putter1, M. Fiocco1 and R. B. Geskus - *Statist. Med.* 2007; **26**:2389–2430

A dedicated package for Stochastic reserving (Mack, ODP bootstrap…)

BootChainLadder example



MultiChainLadder example



What else?

- **Much more can be done**. Thousands of packages are available:
  - **"official" R packages** that are created by the R Core Team
  - hundreds of packages that have been contributed by many people. Some of these packages represent **cutting-edge statistical research** as a lot of statistical research is first implemented in R.

- Some have been **developed by actuaries for actuaries**
  - Markus Gesmann and Wayne Zhang's **Chainladder** package which implements Chain Ladder based stochastic reserving methods
  - The **actuar** Package developed and maintained by Vincent Goulet which includes several functions of interest to actuaries
  - **lossDev**, by Christopher Laws and Frank Schmid which uses a Bayesian method of stochastic reserving

Other packages of interest to actuaries include

- MASS : many useful functions, data examples and negative binomial linear models
- gamlss : Generalized Additive Models for Location, Scale and Shape
- lme4, gamlss.mx, hglm : packages for Generalized Linear Mixed Models
- VGAM: Ordinal and nominal regressions
- rpart : Classification and Regression Trees
- survival : survival analysis including parametric accelerated failure models and Cox model
- splines : B-splines and natural cubic splines
- copula : commonly used copulas including elliptical (normal and t), Archimedean (Clayton, Gumbel, Frank, and Ali-Mikhail-Haq)
- POT, evir : functions related to the Extreme Value Theory
- rjags, r2jags, R2WinBugs : Markov Chain Monte Carlo methods
- …

Why R a risk management toolkit ?

- With R or equivalent solutions, the actuary can
  - Identify opportunities and reduce adverse selection by **uncovering risk segments with inadequate pricing**
  - Reduce estimation risk by using **appropriate multivariate techniques**
  - Manage specification risk by selecting the **model whose assumptions best match with real life**
  - Mitigate user bias by **selecting the "best" model objectively**
  - Keep rating structure simple by **distinguishing between more significant and less significant factors** and avoiding over-fitting
  - Reduce risk of loss of business by **modelling demand side**
  - Reduce risk of insolvencies by **quantifying the magnitude of potential deviations from his "Best Estimates"**

Questions ?

- <u>Contact details</u> :

Xavier Conort

Gear Analytics

Email : xconort@gear-analytics.com

Mobile : + 65 9339 8636

R script for slide 11 (1/2)

```
par(list(mfrow=c(2,4),cex=0.5,mar=c(4.5,4.5,4.5,3), cex.main=2))

# Call Cars93 dataset from the MASS package
library(MASS);data(Cars93);str(Cars93);summary(Cars93)
# Draw pie chart and bar chart
library(grDevices) # for rainbow colors
pie(table(Cars93$Type),main="Pie chart",col=rainbow(6))
barplot(table(Cars93$Origin,Cars93$Type),main="Bar chart",legend.text=TRUE,ylab="Makes
count”)
# Fit distribution and plot density
dens<-density(Cars93$Horsepower)
truehist(Cars93$Horsepower,ymax=max(dens$y)*1.3,main="Density plot")
lines(dens,col=4)
fit.ln <- fitdistr(Cars93$Horsepower, "lognormal")
curve(dlnorm(x,meanlog=fit.ln$estimate[1],sdlog=fit.ln$estimate[2]), lwd=2, col=2,
add=TRUE)
legend("topright", lwd=c(1,2), col=c(4,2), legend=c("Kernel density", "Lognormal
density"))
# Draw QQ-plot
library(car)
qqPlot(Cars93$Horsepower, distribution="lnorm", meanlog=fit.ln$estimate[1],
sdlog=fit.ln$estimate[2])
title(main="QQ plot of empirical quantiles \n against lognormal quantiles")
```

R script for slide 11 (2/2)

```
 # Draw Box-plot and Violin-plot
boxplot(Cars93$Horsepower~Cars93$Type,main="Side-by-side
boxplots",ylab="Cars93$Horsepower",col="blue")
library(UsingR)
simple.violinplot(Horsepower~Type, data =Cars93, col = "orange")
title("Violin plots")

# Univariate analysis
pricegroup <- as.factor(paste("$",cut(Cars93$Price, breaks=c(0,10,20,30,70))))
plot.design(Cars93$Horsepower~Cars93$Origin+Cars93$Type+pricegroup, main="Univariate
Analysis",xaxt="n")

# Fit GLM and Draw Scatterplot
Cars93<-Cars93[order(Cars93$Price),]
library(splines)
fit<- glm(Horsepower~bs(Price),data=Cars93,family=Gamma(link="log"))
plot(Cars93$Price,predict(fit,data=Cars93, type="response"), col=4,type="l",
main="Scatterplot", ylab="Cars93$Horsepower")
legend("topleft",c(levels(Cars93$Origin),"curve fitting of horsepower as function of
price"),pch=c("x","x",NA),col=c(1,2,4),lty=c(NA,NA,1),cex=0.8)
text(Cars93$Price,Cars93$Horsepower, Cars93$Type, cex=Cars93$Weight/max
(Cars93$Weight), col=ifelse(Cars93$Origin=="USA",1,2))
text(230,10, "size as function of cars weight")
```