

Memory Issues with R

- R is limited to about 4 GB of memory in a 32 bit environment.
- Upgrading to a 64 bit environment with a 64 bit version of R is very helpful (have not tried this)
- Even if your file loads, you still may run into issues with models. R allocates memory strangely.
- You can also use a package ((ff, filehash, R.huge, or bigmemory)
- It is easy to sample and bootstrap models.

Sampling/Bootstrapping

- Sample data
 - #Load the "Forbes2000" data frame already contained within R
 - `data("Forbes2000", package = "HSAUR")`

 - #View the dimensions of the data
 - `dim(Forbes2000)`

 - #Take a sample of size 5 without replacement
 - `Forbes <- Forbes2000[sample(nrow(Forbes2000), size = 5, replace = F),]`
- Now you can build a model on Forbes and repeat the process

Resources: R Working Party Wiki

An R Working Party Wiki site has been set up. It includes this presentation as well as other resources for those interested in R. The site can be found at

<http://rworkingparty.wikidot.com/>

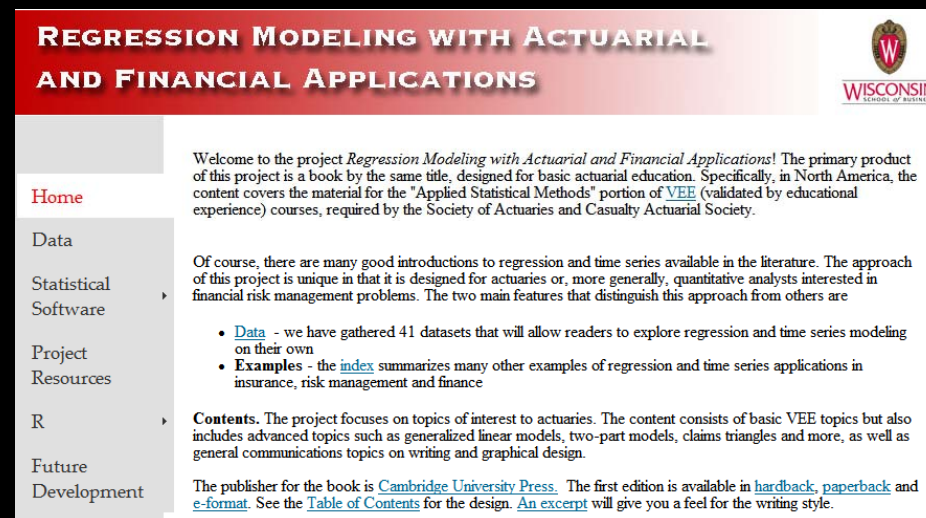
If you'd like to join this site, you may do so by creating an account with wikidot.com. The password for site membership is MembershipRcas. You can find these presentations on the 'Introduction to R' page on the wiki; they are downloadable under the 'Files' tab at the bottom. Contact caleb@kerper-bowron.com by next Wednesday if you have questions.

Resources: Regression Modeling


Many resources are available for further exploration in R.

Jed Frees' website provides an extensive introduction to R. He is an actuarial science professor at Wisconsin.

<http://research3.bus.wisc.edu/file.php/129/RTSABook/WebData/home.html>



REGRESSION MODELING WITH ACTUARIAL AND FINANCIAL APPLICATIONS



Home

Data

Statistical Software

Project Resources

R

Future Development

Welcome to the project *Regression Modeling with Actuarial and Financial Applications*! The primary product of this project is a book by the same title, designed for basic actuarial education. Specifically, in North America, the content covers the material for the "Applied Statistical Methods" portion of VEE (validated by educational experience) courses, required by the Society of Actuaries and Casualty Actuarial Society.

Of course, there are many good introductions to regression and time series available in the literature. The approach of this project is unique in that it is designed for actuaries or, more generally, quantitative analysts interested in financial risk management problems. The two main features that distinguish this approach from others are

- **Data** - we have gathered 41 datasets that will allow readers to explore regression and time series modeling on their own
- **Examples** - the [index](#) summarizes many other examples of regression and time series applications in insurance, risk management and finance

Contents. The project focuses on topics of interest to actuaries. The content consists of basic VEE topics but also includes advanced topics such as generalized linear models, two-part models, claims triangles and more, as well as general communications topics on writing and graphical design.

The publisher for the book is [Cambridge University Press](#). The first edition is available in [hardback](#), [paperback](#) and [e-format](#). See the [Table of Contents](#) for the design. [An excerpt](#) will give you a feel for the writing style.

[Home](#) > [Resources](#) > [Regression Modeling](#) >

Resources: FAViR: www.favir.net

Bayesian Claim Severity with Mixed Distributions
By Benedek Esposito

FAViR

This paper is produced automatically as part of FAViR. See <http://www.favir.net> for more information.

Abstract

Suppose the default claim severity distribution is a finite mixture of single severity distributions. For instance, many ISO distributions are mixed exponential. The technique in this paper can be used to adjust the weights of the mixture in a principled way to partially-visible observed claim severities.

In Bayesian terminology, this paper assumes a Dirichlet distribution over initial mixture weights. The posterior distribution, conditional on one or more observed claim severities, is computed using either a custom Gibbs sampler or the rstanj package.

1 Introduction

Many applications require the position of modeling claim severities based on limited historical data. One example is the pricing of excess of loss reinsurance layer. The technique in this paper is intended to help cover the awkward middle ground between having no data and having lots of data. If no claim data is available, some default claim severity distribution may be available. For instance, ISO publishes mixed exponential severity distributions based on aggregate data (see Palmer for a basic description of ISO's methodology). When many data points are available, maximum-likelihood curve fitting outside the ISO's method work well.

Instead, it seems the expected claim severity distributions should emerge from the default distribution into some fitted distribution as more and more claims are observed. The most principled way of doing this is to use Bayesian statistics. This paper models the situation under these assumptions:

1. The default severity distribution is a mixed exponential (such as those supplied by ISO). Other mixed distributions would probably work with minor modifications.
2. Posterior uncertainty is modeled using a Dirichlet distribution over the mixture weights. This requires one additional parameter, interpreted as the confidence in the default distribution.

1

4 GRAPHICAL DIAGNOSTICS

Figure 6: Comparison of Model Fits by Development Period

4 Graphical Diagnostics

Graphs are popular for evaluating the appropriateness of a stochastic reserving model (see Barnett and Gilmer for Barnett, Straton, and Venter for more information on their use as reserving diagnostic).

Figure 6 plots incremental loss ratio vs starting loss ratio for each development period. If the standard model worked perfectly, the incremental loss ratio would be proportional to the starting loss ratio; all the points would fall on a line going through the origin. If the Runquist-Ferguson method worked perfectly, the incremental loss ratio would be independent of starting loss ratio; all the points would fall on a horizontal line. The graph in Figure 6 is a simple way to visually judge which, if either, method is working. Another way is another way to judge the appropriateness of a model. A model's residual is an actual observed value minus the model's predicted value. Figure 7 shows residuals by

4

3.2 Variable Written Rate 3 METHOD OUTPUT

Figure 5: Parallelogram with Variable Premium Rate

3.2 Variable Written Rate

If the rate at which premium is written changes, this should affect the on-level factor used. This implementation of the parallelogram method allows the written rate function to be specified as a step function. Using the assumptions of section 2, the inferred graph changes as shown in Figure 5. In that plot, the dotted line represents specified rate of premium written, while the numbers labels still represent the on-level factor for each rating period. The on-level earned premium and earned premium factors that reflect the variable premium rate are shown in Figure 6.

Period	Start	Period End	Earned Premium	On-Level Earned Premium	On-Level Factor
2001	2002	30	1,304	30.1	
2002	2003	36	1,253	47.0	
2003	2004	60	1,225	78.5	
2004	2005	52	1,216	63.9	
2005	2006	42	1,080	48.1	

Figure 6: Variable Premium Results

Because the premium earned over any time period is just the area under the inferred premium curve divided by the time length, the traditional parallelogram method will work, although it is more complicated to calculate all the areas correctly. To obtain the variable

4

4 DETAILED MODEL

Figure 4: Graph of Results

4 Detailed model

Formally, the Bayesian probabilistic model used is defined by three equations:

$$p(\theta) = \frac{e^{-\theta}}{\Gamma(\theta)} \quad (1)$$

$$p(\theta | w_1, \dots, w_n) = \frac{e^{-\theta}}{\Gamma(\theta)} \prod_{i=1}^n w_i^{\theta-1} \quad (2)$$

$$p(w_1, \dots, w_n) = \frac{\Gamma(\theta)}{\Gamma(\theta_1) \dots \Gamma(\theta_n)} \prod_{i=1}^n w_i^{\theta-1} \quad \text{with } w_n = 1 - \sum_{i=1}^{n-1} w_i \quad (3)$$

or in other words,

$$\theta | i \sim \text{Exponential}(1, w_i)$$

$$w_1, \dots, w_n \sim \text{Dirichlet}(w_1, \dots, w_n)$$

$$w_1, \dots, w_n \sim \text{Dirichlet}(\theta_1, \dots, \theta_n)$$

where θ is an individual claim severity and θ_i is the expected value of exponential distribution i . θ and basket selection i are assumed independent given the basket weights (w_1, \dots, w_n) .

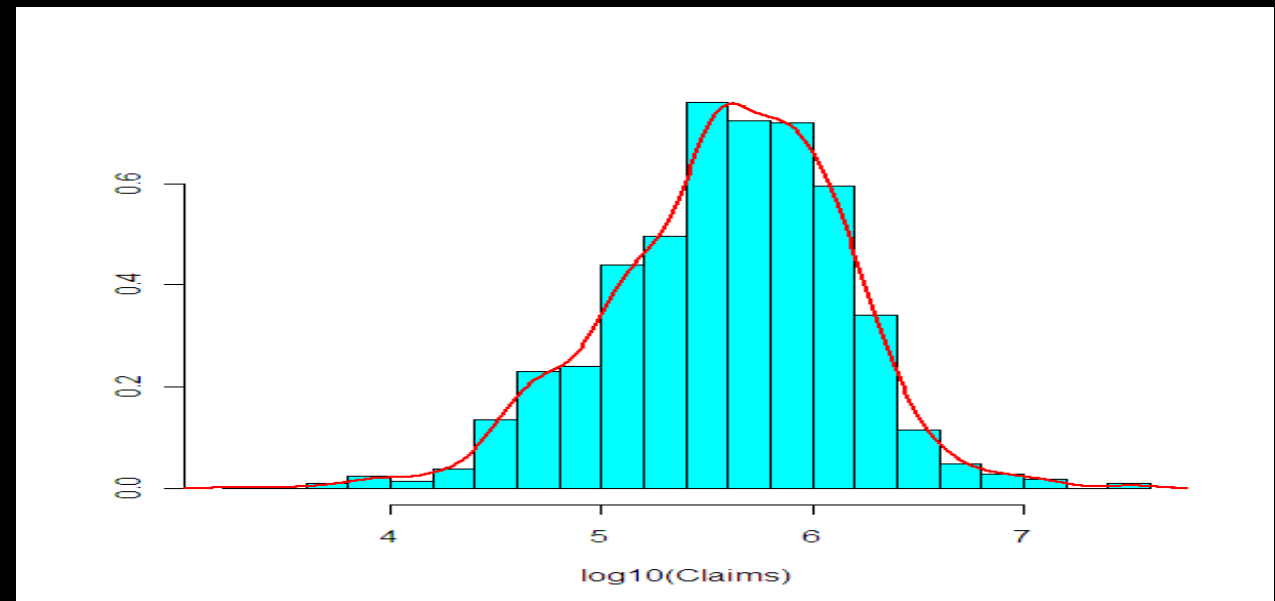
4

- FAViR is a new project to help actuaries with R. It's a combination of
 - A peer reviewed journal where all papers are automatically produced with R
 - An R package (on CRAN as "favir") which allows authors to easily create standardized, polished papers such as those shown above.
- Project Goals:
 - Make it easy for actuaries to use and present advanced and modern techniques
 - Establish trustworthy repository of peer-reviewed actuarial R code

Resources: Introduction to R for Actuaries

Introduction to R! for Actuaries: Histograms

Available on the R Wiki



Avraham Adler, FCAS, MAAA

GUY CARPENTER



MARSH MERCER KROLL
GUY CARPENTER OLIVER WYMAN

Resources: Text Mining

Text Mining

Available Soon on the R
Wiki



Matthew J. Flynn

Travelers

Resources: Additional Resources

Bayesian Loss Development Model

Chris Laws and Frank Schmid, NCCI

- 2009 Annual Meeting Presentation

casact.org/education/annual/2009/handouts/c1-laws_schmid.pdf

- lossDev Project

lossdev.r-forge.r-project.org/

Resources: Additional Resources

RSeek – dedicated search for R topics

rseek.org

Sweave (embed R code in LaTeX)

stat.uni-muenchen.de/~leisch/Sweave/

Maps in R

geography.uoregon.edu/GeogR/topics/maps.htm

R Cheat Sheets

devcheatsheet.com/tag/r/

R Colors

research.stowers-institute.org/efg/R/Color/Chart/index.htm

R Working Party Members/Contributors

Co-Chairpersons

Lee M. Bowron

Thomas R. Kolde

Members

Avraham Adler

James M. Boland

Kevin S. Burke

Alan Chalk

Donald L. Closter

Kiera Elizabeth Doster

Benedict M. Escoto

Sholom Feldblum

Matthew J. Flynn

Edward W. Frees

James C. Guszczka

Todd W. Lehmann

John J. Lewandowski

Stephen L. Lienhard

Peter James Mulquiney

Scott G. Sobel

Tony A. Van Berkel

Yi Zhang

Support

Cheri Widowski (staff liaison)

Caleb Moxley (project intern)

Any Questions?