Presented to:

RAM IX Training Summit

Data Modeling & Analysis with R

TECHNOLOGY DRIVEN. WARFIGHTER FOCUSED.

Presented by:

Seth E. Farrington

Mechanical & Reliability Engineer

U.S. Army Aviation and Missile Research, Development, and Engineering Center

2 November 2016

About Me

- BSE & MSE in Mechanical Engineering from UAH
- Worked as a Reliability Engineer since 2009
  - Mathematical modeling and analysis on everything from materials to complex global networks
  - 6 technical publications on RAM topics
- Used R since 2010
- Currently working for the U.S. Army AMRDEC ED RAM Division building stochastic reliability models for Army Aviation Hardware

Why you should stay for my tutorial?

- R is an extremely powerful and accessible analysis tool
- R has become extremely popular across a variety of industries in recent years
- R does not currently have a large following in the RAM world, but I believe that it should.

Outline

- What is R?
- History
- Capabilities
- Resources
- Notable Add-On Packages For RAM
- Case Studies
  - Google
  - Facebook
  - ANZ Bank
  - Survival Analysis Simulation

What is R?

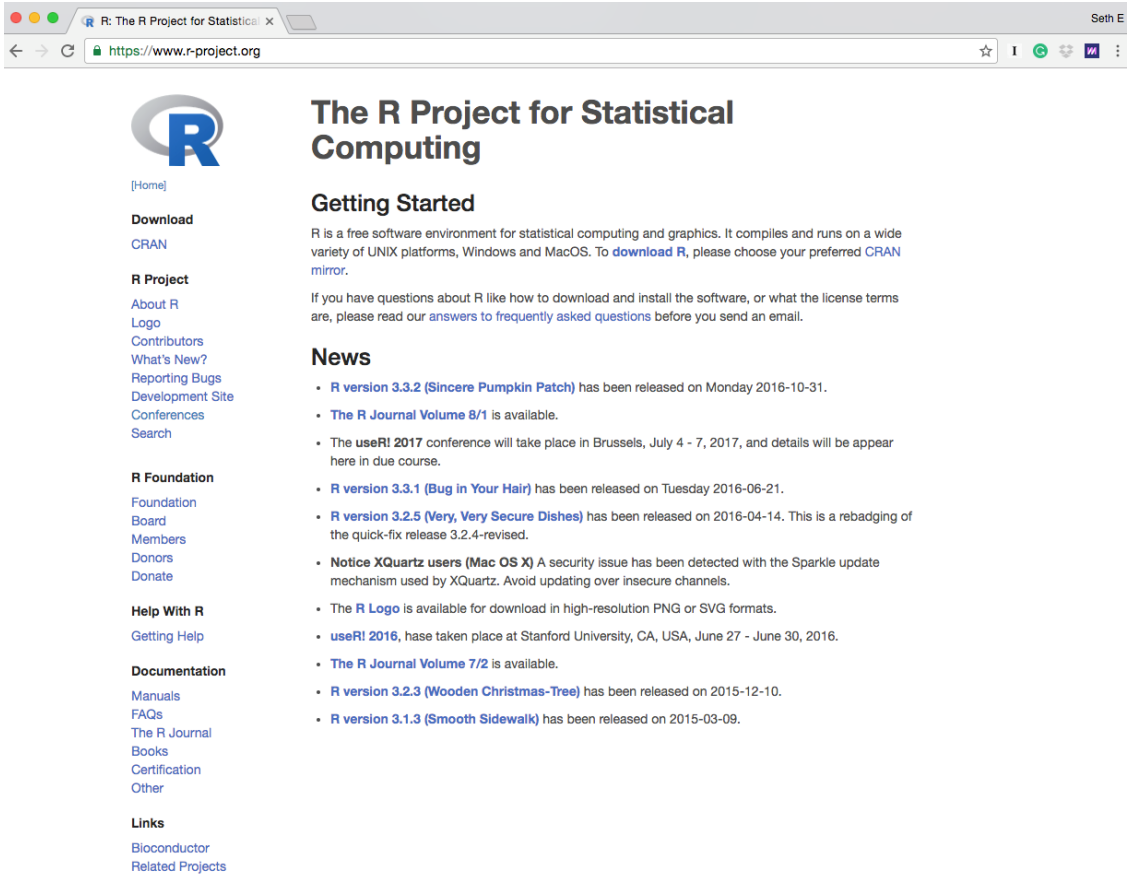
- "R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R." -
- R is similar to MATLAB and SAS in interface and capabilities, but free and open source

History of R

- R is an implementation of the S language which was developed by John Chambers while at Bell Labs
- R was developed by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand
- The project was conceived in 1992 and the initial release in 1995, with the first stable beta version in 2000
- The R foundation was formed by the core development team to manage the continued development of R

Official Resources

- **Official Website:**
  - www.r-project.org
- **Official repository for r packages and materials cran.r-project.org**



Manuals and Books

- **“An Introduction to R” the official introductory manual**
- **“R Data Import/Export” official guide**
- **The R Journal -** The R Journal is the open access, refereed journal of the R project for statistical computing. It features short to medium length articles covering topics that might be of interest to users or developers of R, including
  - Add-on packages: short introductions to R extension packages.
  - Programmer's Niche: hints for programming in R.
  - Help Desk: hints for newcomers explaining aspects of R that might not be so obvious from reading the manuals and FAQs.
  - Applications: demonstrating how a new or existing technique can be applied in an area of current interest using R, providing a fresh view of such analyses in R that is of benefit beyond the specific application.

Conferences

- **useR! -** This is the main meeting of the R user and developer community, its program consisting of both invited and user-contributed presentations:
  - The invited keynote lectures cover a broad spectrum of topics ranging from technical and R-related computing issues to general statistical topics of current interest.
  - The user-contributed presentations are submitted as abstracts prior to the conference and may be related to (virtually) any R-related topic. The presentations are typically organized in sessions of either broad or special interest, which also comprise a “free” discussion format. Such a discussion format not only provides a forum for software demonstrations and detailed discussions but also supports the self-organization of the respective communities
  - Last held June 2016

Conferences

- **R/finance –** The annual R/Finance conference for applied finance using R, The two-day conference will cover topics including portfolio management, time series analysis, advanced risk tools, high-performance computing, market microstructure, and econometrics. All will be discussed within the context of using R as a primary tool for financial risk management, portfolio construction, and trading. Over the past seven years, R/Finance has included attendees from around the world. It featured presentations from prominent academics and practitioners. Last held May 2016
- **DSC – Directions in Statistical Computing –** DSC is a conference for the developers of statistical software and researchers in statistical computing which is somewhat focused on but not exclusively devoted to R. It aims at providing a platform for exchanging ideas about developments in statistical computing (rather than `only' the usage of statistical software for applications). As the associated papers are often technical and difficult to publish even in computational statistical journals, the DSC publishes post-conference proceedings of the papers that were accepted for publication. Last held July 2016

Courses

- **Coursera – Johns Hopkins University Data Science Specialization Program**
  - This Specialization covers the concepts and tools you'll need throughout the entire data science pipeline, from asking the right kinds of questions to making inferences and publishing results. In the final Capstone Project, you'll apply the skills learned by building a data product using real-world data. At completion, students will have a portfolio demonstrating their mastery of the material.
  - Topics include: Intro to R programming, Getting and Cleaning Data, Exploratory Analysis, Statistical Inference, Regression Models, Machine Learning
  - www.Coursera.org

Courses

- **Coursera – Duke – Statistics with R**
  - In this Specialization, you will learn to analyze and visualize data in R and created reproducible data analysis reports, demonstrate a conceptual understanding of the unified nature of statistical inference, perform frequentist and Bayesian statistical inference and modeling to understand natural phenomena and make data-based decisions, communicate statistical results correctly, effectively, and in context without relying on statistical jargon, critique data-based claims and evaluated data-based decisions, and wrangle and visualize data with R packages for data analysis.
  - Topics Include: Inferential Statistics, Linear Regression and Modeling, Bayesian Statistics
  - www.Coursera.org

Courses

- **Udacity – Data Analysis with R – By Facebook**
  - Exploratory data analysis is an approach for summarizing and visualizing the important characteristics of a data set. Promoted by John Tukey, exploratory data analysis focuses on exploring data to understand the data's underlying structure and variables, to develop intuition about the data set, to consider how that data set came into existence, and to decide how it can be investigated with more formal statistical methods.
  - www.Udacity.com
- **Kaggle** was founded as a platform for predictive modelling and analytics competitions on which companies and researchers post their data and statisticians and data miners from all over the world compete to produce the best models. In addition to competitions, Kaggle has excellent educational resources developed by the community. R is one of the most popular environments for the competitions.
  - www.kaggle.com

What Sets R Apart?

- R is a highly capable command line interface statistical package, but it really shines because it is free, open source, and widely used

- That combination of free, open source, and popular has resulted in an active community of users developing an enormous array of add-on packages extending R's capabilities to just about every use case imaginable

Notable Add-On Packages

- Data connection and import
  - RODBC – R package for the industry standard ODBC database interface
  - dplyr – efficient data manipulation
- Analysis
  - Survival – Survival analysis
  - PhaseType – R package for working with Phase-Type distributions in R
  - ReliabilityTheory – A toolkit for structural reliability theory in R
  - mlmc – an implementation of Multi-Level Monte Carlo in R
  - FaultTree – package for fault tree analysis
  - EventTree – package for event tree analysis
  - Abrem – an R implementation of the methods described by Dr. Robert B. Abernathy in "The New Weibull Handbook"
- Output
  - Shiny – Interactive web apps in R
  - Knitr – elegant flexible and fast dynamic report generation in R

Survival Analysis

- **Survival Analysis Package**
- **Developed by Terry Therneau at the Mayo Clinic**
- Contains the core survival analysis routines, including definition of Surv objects, Kaplan-Meier and Aalen-Johansen (multi-state) curves, Cox models, and parametric accelerated failure time
- Extremely powerful survival analysis toolbox, distribution fitting with 2 lines of code

```
## Create survival object. A survival object is the data
## structure used by the survival package
y = Surv(data$Hours, data$Censored)

## Fit survival data, non-parametric, Kaplan-Meier Method
ys = survfit(y ~ 1, type="kaplan-meier")

## Fit Weibull distribution
yw = survreg(y ~ 1, dist="weibull")
```

System Lifetime Modeling

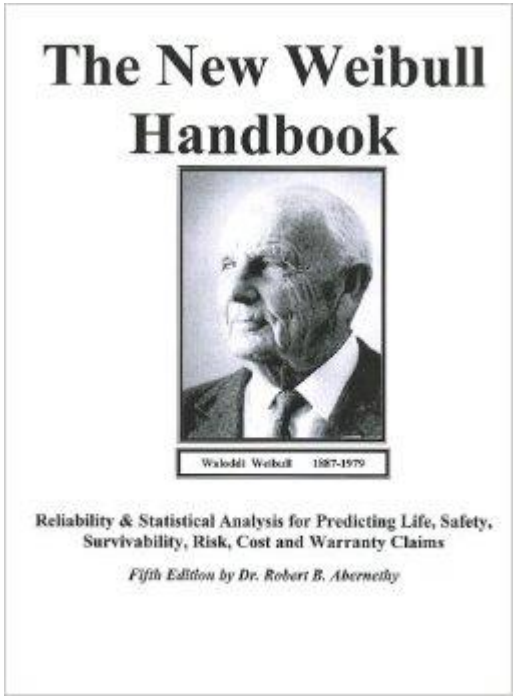
The following packages were developed and are maintained by Dr. Louis Aslett

- **PhaseType –** An early release of an R package for working with Phase-type distributions. At present, there are two functions for Bayesian inference on Phase-type models, which are high-speed C implementations of MCMCalgorithms. The package is available on CRAN.
- **ReliabilityTheory –** A toolkit for structural reliability theory. This includes methods of system reliability analysis based on structure functions, system signature and survival signatures. The package is available on CRAN.
- **mlmc –** An implementation of Multi-level Monte Carlo for R. This package builds on the original GPL-2 Matlab and C++ implementations by Mike Giles (see http://people.maths.ox.ac.uk/~gilesm/mlmc/) to provide a full MLMC driver and example level samplers. Multi-core parallel sampling of levels is provided built-in. The package is available on CRAN.

Fault Tree and Weibull Analysis

The following packages were developed by openreliability.org

- **FaultTree –** This R package is used to build a fault tree as a dataframe object. A tree is constructed by an initial ftree.make() call. Subsequent addition of add... functions build up the tree. The logic gates of a fault tree are calculated from bottom to top in a batch fashion.
- **EventTree –** This R package is used to build an event tree as a dataframe object. A tree is constructed by an initial etree.make() call. Subsequent addition of addControl and addOutcome functions build up the tree. Event tree calculations proceed during the tree construction. There is no GUI associated with this package, nor is one expected in the R environment. A user is expected to code scripts defining the tree as a final version.
- **Abrem –** an R implementation of the methods described by Dr. Robert B. Abernathy in “The New Weibull Handbook”



knitr

- **Knitr – Elegant, flexible and fast dynamic report generation with R**
- The design of **knitr** allows any input languages (e.g. R, Python and awk) and any output markup languages (e.g. LaTeX, HTML, Markdown, AsciiDoc, and reStructuredText)

A Minimal Demo of knitr

Yihui Xie

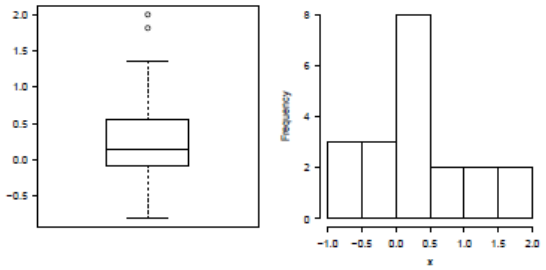
August 11, 2016

You can test if **knitr** works with this minimal demo. OK, let's get started with some boring random numbers:

```
set.seed(1121)
(x <- rnorm(20))

##  [1]  0.1449583  0.4383221  0.1531912  1.0849426  1.9995449 -0.8118832  0.1602680
##  [8]  0.5858923  0.3600880 -0.0253084  0.1508809  0.1100824  1.3596812 -0.3269946
## [15] -0.7163819  1.8097690  0.5084011 -0.5274603  0.1327188 -0.1559430

mean(x)

## [1] 0.3217385

var(x)

## [1] 0.5714534
```

The first element of x is 0.1449583. Boring boxplots and histograms recorded by the PDF device:

```
## two plots side by side (option fig.show='hold')
par(mar = c(4, 4, 0.1, 0.1), cex.lab = 0.95, cex.axis = 0.9, mgp = c(2, 0.7, 0), tcl = -0.3,
    las = 1)
boxplot(x)
hist(x, main = "")
```

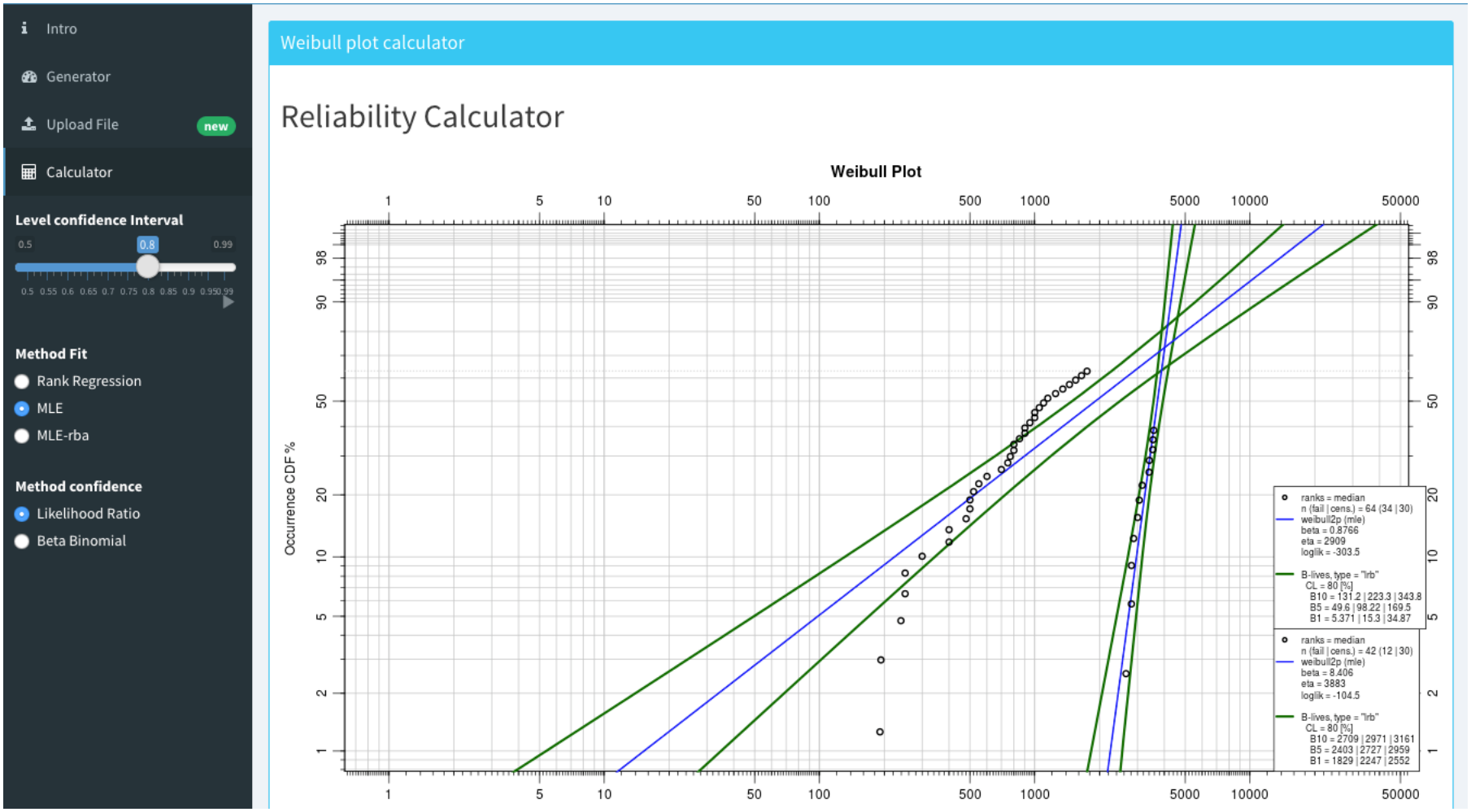


Do the above chunks work? You should be able to compile the TEX document and get a PDF file like this one: https://github.com/yihui/knitr/releases/download/doc/knitr-minimal.pdf. The Rnw source of this document is at https://github.com/yihui/knitr/blob/master/inst/examples/knitr-minimal.Rnw

1

Shiny

- **Dynamic web-apps directly from R code**



Facebook Data Visualization

Facebook Data Visualization

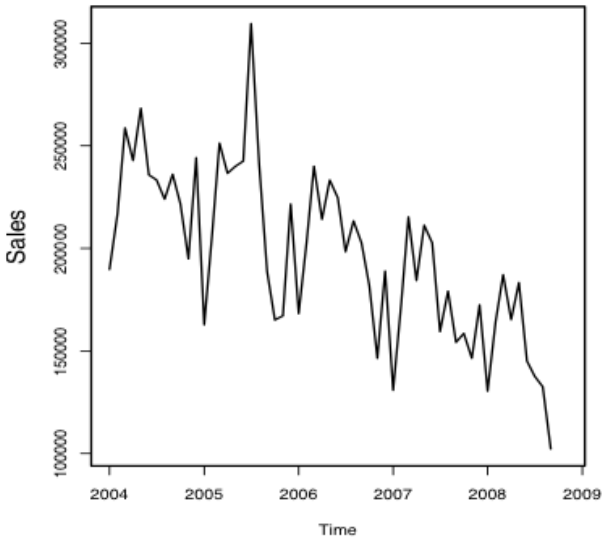
Facebook intern Paul Butler was interested in exploring the locality of friendship

> *"When the data is the social graph of 500 million people, there are a lot of lenses through which you can view it. One that piqued my curiosity was the locality of friendship. I was interested in seeing how geography and political borders affected where people lived relative to their friends. I wanted a visualization that would show which cities had a lot of friendships between them."*

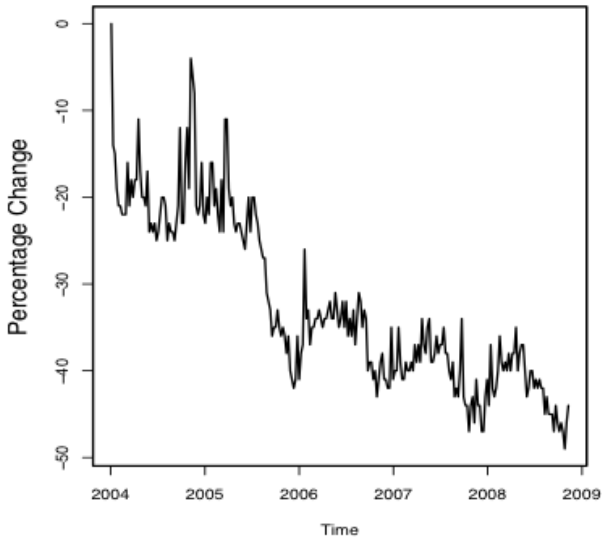
- Paul began by taking a sample of about 10 million pairs of friends from Facebook's Apache Hive data warehouse, combined with current city and summed the number of friends between each pair of cities. Then merged with the longitude and latitude of each city
- He defined weights for each pair of cities as a function of the Euclidean distance between them and the number of friends between them. Then plotted lines between the pairs by weight, so that pairs of cities with the most friendships between them were drawn on top of the others. He used a color ramp from black to blue to white, with each line's color depending on its weight. And also transformed some of the lines to wrap around the image, rather than spanning more than halfway around the world.
- https://www.facebook.com/notes/facebook-engineering/visualizing-friendships/469716398919

Google Uses R to Predict Economic Activity

- Google Chief Economist Hal Varian asks the question: can we use Google Insights data like this to predict the economic variables, even before they are reported?
- Government economic reports are released weeks after the fact, and even then only with preliminary data to be revised upwards or downwards later. The key insight is that the volume of Google searches for particular keywords is correlated with related economic indicators. For example, here's a comparison of searches for the keyword "Ford" with actual monthly sales of Ford vehicles:



(a) Ford Monthly Sales



(b) Ford Google Trends

Google Uses R to Predict Economic Activity

- Seems like you can. Varian used R to fit seasonal autoregressive models to retail sales, automotive sales, home sales, and passenger arrival data, and in each case made better predictions by including Google Trends data as a predictor than without. You can try this yourself by downloading R and using Google Insights for Search to download the trend data -- Varian has helpfully provided the R code for the Ford model in the paper. (Just be sure to log into Google Insights with your Google Account to enable downloading of trends data as a CSV file.)
- https://research.googleblog.com/2009/04/predicting-present-with-google-trends.html

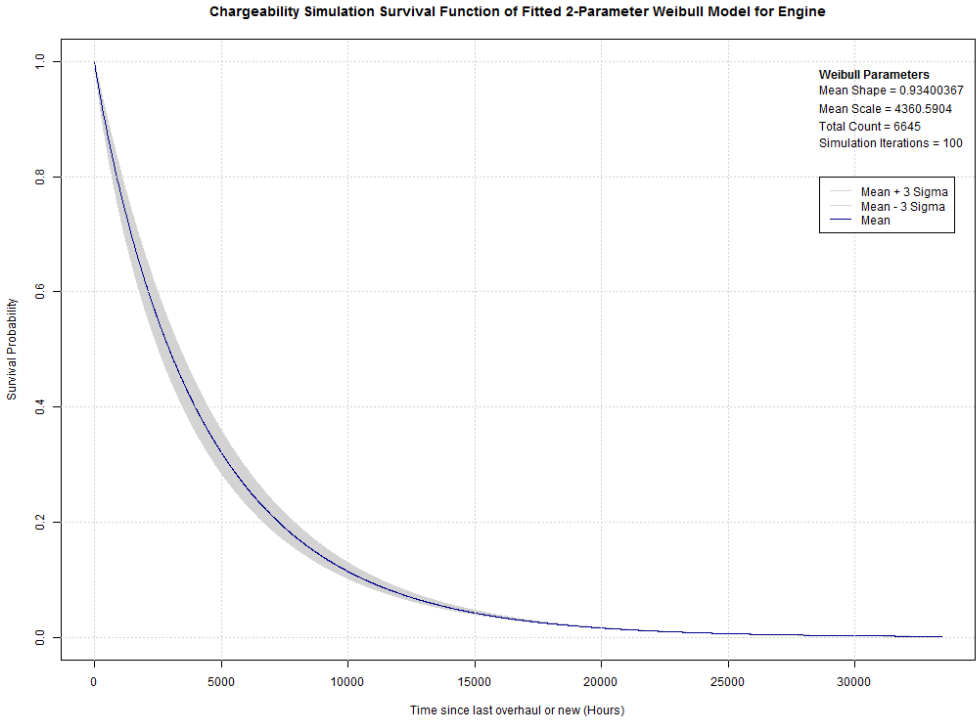
ANZ Bank

- In 2011 Hong Ooi from ANZ (Australia and New Zealand Banking Group) gave a presentation on "Experiences with using R in credit risk".
  - How R is used to fit models for mortgage loss at ANZ. A custom model is created to estimate probability of default for individual loans, with a heavy-tailed T distribution for volatility. (Slide 12 shows how the standard lm function for regression is adapted for a non-Gaussian error distribution -- one of the many benefits of having the source code available in R.)
  - A comparison between R and SAS for fitting such non-standard models. Ooi notes that SAS does have various options for modeling variance (e.g. SAS PROC MIXED, PRIC NLIN), but "none of these are as flexible or powerful as R". The key difference, Ooi says, is that R modeling functions return an *object* (as opposed to merely text output) which can be modified and manipulated by the R programmer to adapt to new modeling situations and generate predictions, summaries, etc.
  - How ANZ implemented a stress-testing simulation, made available to business users via an Excel interface. The main computation is done in R in just two minutes (compared to an original all-SAS version that "took ~4 hours to run, often crashed due to lack of disk space"). Since the data is stored in SAS, SAS code is still used to generate the source data ... although an R script (seen on slide 25) is used to automate the process of writing the SAS code to do so (neatly stepping around the flexibility limitations of SAS).

Simulation to Account for Inaccuracies

While running survival analysis questions arose about the accuracy of the chargeability of failures. This simulation was developed to characterize resulting the uncertainty.

- This is a Monte Carlo simulation where the chargeability of a failure is the random variable.
- Chargeability is used to determine if a record is censored
- For each iteration of the simulation the 2-parameter Weibull distribution is fit using MLE and the parameters are saved



The input data for this example is fictitious and was created for demonstration purposes only

AMRDEC Web Site
www.amrdec.army.mil

Facebook
www.facebook.com/rdecom.amrdec

YouTube
www.youtube.com/user/AMRDEC

Twitter
@usarmyamrdec

Public Affairs
AMRDEC-PAO@amrdec.army.mil