VOLUME 19, NO 2, DECEMBER 2008

A joint newsletter of the Statistical Computing & Statistical Graphics Sections of the American Statistical Association

Statistical COMPUTING & GRAPHICS

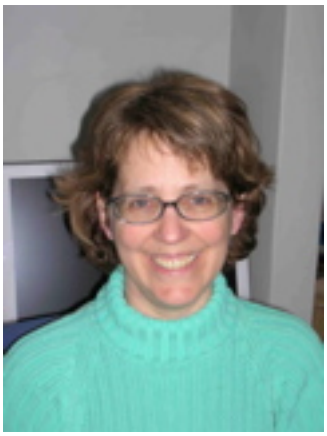
A Word from our 2008 Section Chairs

DAN ROPE
GRAPHICS

Perhaps I was hit in the head by too many flying monkeys at the mixer in August, but it seems as if statistics has gone mainstream. Sure, I have a lot of colleagues with a great sense of humor, but I imagine that the history of statistics is not rife with statisticians featured on hip political satire shows such as "The Colbert Report." Nonetheless, in addition to seeing the creator of www.fivethirtyeight.com on "Comedy Central," several people forwarded me this link of presidential prediction wizardry. I actually found myself reloading this web site more than any other on election night—mostly to check out how accurate the projections were. It was impressive. Take a look, and then look at the methodology.

Flipping to CNN that night was almost like watching a Steven Spielberg movie. Sure, there was that ghostly (um, freaky) hologram trick and the eerie floating capital building, but their toy that got my attention was the "Minority Report"—like interactive data analysis jumbo-tron touch screen. I have seen the demos of large touch screen technology in the past, but mostly applied to typical desktop-ish software.

Continues on Page 2.....



DEBORAH NOLAN
COMPUTING

Section members may be interested to hear about a new section of the ASA that has recently been approved by the Council of Sections Governing Board. In addition to the newly created Section on Statistical Learning and Data Mining, the Section for Statistical Programmers and Analysts will start in January 2009.

Continues on Page 2.....

Computing, Continues from Page 1....

The new section officers hail primarily from the pharmaceutical industry; they are: Steve Yao, Amgen, Chair, Wendy Zhang, Sanofi-Aventis, Chair-elect, Rayamjhi Jyoti, Eli Lilly, Program Chair, Jennifer Borkowsky, Genentech, Treasurer, Amanda Tweed, Millennium, Secretary, and Monica Johnston, UCSF, Section Representative to the Council of Sections.

According to its charter, the principle interests of the Section for Statistical Programmers and Analysts are to:

1. Encourage a broad discussion of programming concepts, theories, and techniques that are used for statistical analysis.

2. Encourage discussion of cross-functional activities and pre-analysis (e.g. data management) activities that affect statistical programming work.

3. Encourage discussions on career paths for members.

4. Establish and maintain liaison with other Association sections as appropriate.

Last August when the proposed section first came to our attention, there was concern over potential overlap with the mission of the Section on Statistical Computing. One reason for this was that, at the time, the proposed section planned to call itself the Section on Statistical Programmers. We met with the section officers to hear about their goals and to ask them to change their section name to avoid confusion with our section. Although we do not object to the start of this new section, our relationship with this section is yet undefined and we expect it will evolve over the next few years. Future possibilities might include joint membership, co-sponsorship of satellite meetings, and joint publications. This is new territory for us, and your ideas are more than welcome.

A second topic that I want to bring to your attention relates to the training of statisticians at all levels (BA, MA, and PhD). In the previous newsletter, I mentioned a series of workshops that I am organizing with Duncan Temple Lang, UC Davis, and Mark Hansen, UCLA. These workshops are funded by the National Science Foundation and the Consortium to Advance Undergraduate Statistics Education with the purpose of making significant changes to the computational training of statisticians. The basic premise is that statistics curricula need to be modernized to embrace computing as an essential building block of statistical creativity and practice. We held our second of three workshops this past summer, where over 30 faculty met to discuss ideas for changing their statistics programs in innovative ways to incorporate more computing. Out of the first two workshops, a Google group has formed and a wiki was created. The third workshop will be held this summer, and we expect to organize a session at the Joint Statistics Meeting in Washington D.C. to report on these activities. For more information on these efforts visit http://www.stat.berkeley.edu/~statcur. I encourage you to get involved in this effort and support significant change to the computational training of future statisticians.

- Deborah Nolan

Graphics, Continues from Page 1....

The statistical graphics were--of course, simple; but it was interesting to watch the hand gestures they used to interact with exit poll bar charts complete with live data filters. You know it was real because of the bugs—and because at one point a commentator was too short to reach the bar he wanted to discuss. So, a work in progress, but somehow it felt similar to the excitement of experimenting with CAVE technology back in grad school.

So, with statistics entering prime time—and graphics going along as the face, it is a great time to disseminate good visualization principles and techniques. Perhaps it is a small step, but the statistical graphics section will now begin contributing Webinars to the Committee on Outreach Education at ASA. This is a great opportunity to share your wealth of knowledge so both those within the statistical community and beyond can benefit. Naomi Robbins will host our first Webinar on avoiding common graphical mistakes. If you are interested in contributing you can contact either me—or the current chair of the graphics section.

Speaking of current events, Graham Wills has a new book titled "Visualizing Time" coming out around winter of 2009—which is sure to be as informative as

it is entertaining. Also, Lee Wilkinson, Guy Lebanon, George Michailidis, and other graphics community members are involved—and have received Department of Homeland Security funded NSF grants for—an interesting initiative called FODAVA (Foundations of Data and Visual Analytics) that seeks to develop mathematical underpinnings for visualization. Lee and Graham have a paper on this topic that will be published this month in Information Visualization--you can also read all about the FODAVA project at http://www.nsf.gov/pubs/2007/nsf07583/nsf07583.htm, http://nvac.pnl.gov/nsf.stm and

http://fodava.gatech.edu/node/5.

Finally, 2009 will be an "on" year for the Data Expo competition. It is a whopper of a dataset this time and the winner may even figure out how to never get stuck at Chicago's O'Hare airport. Now, that would certainly be worthy of a television appearance!

As usual, now is the time to start thinking about topic contributed sessions for 2009. These are a great way to gather a theme of talks together and they help gain future invited session allocations for our section.

Lastly, I'd just like to say that I really enjoyed my term as chair and I'd like to give a big thanks to everyone who contributed to a successful 2008 for statistical graphics. Antony Unwin will be taking over the chair position for graphics in 2009. So, we are in good hands and I'm looking forward to his airborne door prize nominations for 2009.

- Dan Rope

EDITORIAL NOTE
What's Inside this Issue

Michael O'Connell and Andreas Krause
moconnel@tibco.com
andreas.krause@actelion.com

In this issue we include two scientific articles; one on fitting the Weibull distribution by Stuart Randa and Robert Klare, and another on development and validation of logistic prognostic models using SAS macros by Rainer Muche, Christina Ring and Christoph Zeigler.

The issue includes a review of UseR 2008 by Uwe Ligges and a preview of the 2009 JSM graphics program by Steve MacEachern. In Tech Corner, we include some interesting photos of election night in New York, showing some great outdoor graphics at the Rockefeller Center and the NBC building...

Please contact us if you have any short articles, excerpts, software or graphs. This is a good forum for short articles on statistical computing or graphics. You can reference your newsletter article and still publish your work elsewhere.

We are seeking a new The Newsletter Editor for the Computing Section. Serving as Newsletter Editor is a great opportunity to get to know members of the statistical computing and graphics communities. It is also a valuable professional service function to both the ASA and the Statistical Computing and Graphics Sections.

Michael and Andreas will work closely with the new editor to manage the transition. Interested people should contact Michael (moconnel@tibco.com), Deb Nolan (Nolan@stat.berkeley.edu) or Andreas (andreas.krause@actelion.com).

Finally, we have the solution to the puzzle from the last issue - this was a popular puzzle with a couple of entries that were close to the solution.

Happy New Year !

Michael and Andreas

Scientific Article

AUTOWEIBULL

Stuart K. Randa
Randa Consulting
213 West Pembrey Drive
Wilmington, DE 19803 USA

Robert J. Klare
5003 South Mission Drive
St. Joseph, MO 64505 USA

1. Summary

AutoWeibull is used to analyze data for consistency and also to predict life phenomena. The three-parameter Weibull function fits a straight line to many different sets of data having broad ranges of distribution. These ranges include distributions from extensive positive skew (log-normal), to near normal distributions, and on to those of negative skew.

The Weibull analysis provides a minimum value, and often, a maximum value. Normal distribution and logarithmic analyses lack this broad distribution versatility.

The Weibull analysis also shows whether spurious variables are influencing the data. This influencing aspect is indicated by the presence of multiple distribution curves, or a very wide skew in a single distribution

Statisticians using the normal and logarithmic techniques obtain the median with fairly good accuracy. However, data at each of the ends of the span are often inconsistent with the straight graph lines created. This variance occurs because these extreme data points do not correspond to the graph line transversing to plus or minus infinite, or even to zero. In realm of the real world, no earthly phenomenon traverses to such far limits.

Weibull, on the other hand, creates a straight line that conforms to all the data. Weibull is a very appropriate analysis because most phenomena have a minimum positive value. Furthermore, a maximum value may exist, or at least a right tail of the distribution curve that is rapidly reaching a termination point.

The book entitled ***The Weibull Distribution Function*** and the ***AutoWeibull*** computer program are available from Randa Consulting LLC. (www.AutoWeibull.com).

2. Weibull Graph:

Dr. Weibull's mathematical derivation (reference 1) conforms to the straight-line equation below:

$Y = mz + b$, where;

$Y = \ln \ln [1 / (1 - F(x)]$

$F(x)$ = the ordinate median rank

$z = \ln [x - x_o]$

x = a value,

x_o = the lowest positive value for x

$b = - (m) \ln (\text{Theta})$

b = intercept, m = slope, Theta = characteristic value

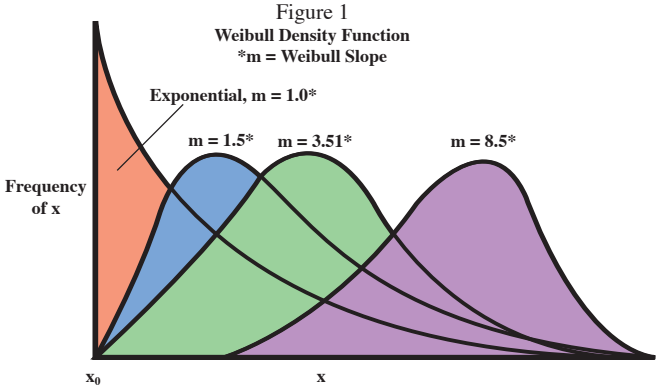
The ordinate value for Y is calculated as follows for a value like, 63.2 %:

$Y = \ln \ln [1 / 1 - 0.632] = 0$

In constructing the Weibull graph, this Y value of zero is positioned on the right hand ordinate and 63.2 % on the left hand ordinate (Figure 2). The entire Weibull ordinate is calculated and plotted in this fashion. Theta, at 63.2 %, is termed the characteristic value because of position at zero.

In two-parameter Weibull, the variable x is positioned on the logarithmic abscissa. In three-parameter Weibull, the variable $(x - x_o)$ is plotted on the abscissa in relation to median ranks on the ordinate scale. Three-parameter Weibull is unique and thus, more valuable, as it describes an entire range of the smooth curve distributions (Figure 1). Note a perfect normal distribution (slope 3.5000) cannot be placed on the graph, but one having a near normal distribution slope of 3.51 can be placed there.

Figure 1
Weibull Density Function
*m = Weibull Slope



AutoWeibull Program

The AutoWeibull computer program is used with Microsoft Excel. Ten megabyes of memory are needed to install the AutoWeibull program from the diskette into the computer. AutoWeibull handles up to 140 data points. Many Weibull computer programs available today consider only 2-parameter Weibull. AutoWeibull was created to handle both 2 and 3-parameter Weibull. In addition, use of the technique requires very little statistical background information.

The computer program obtains modified data points by subtracting the third parameter, x_o from all the data points $(x - x_o)$. After just seconds, the value of x_o selected is the positive number that yields the straight graph line as measured by the "least squares" technique. The x_o value of zero may be manually selected also.

Example One: More than One Distribution:

A laboratory test is routinely used to determine how well a powdered plastic resin fuses while melting. To determine the precision associated with this test, five technicians, over several days, measured the performance of one standard lot of material. The data gather are: 30, 32, 33, 35, 39, 40, 40, 41, 46, 46, 47, 49, 51, 52, 52,53, 53, 60, 60, and 65. The results of the analysis yields a median of 46 and the three-sigma range is plus 35 and minus 23. The Weibull slope shows a skewed distribution of slope 2.6. It is the combination of this skew and the wide three-sigma range of 58 that is concerning.

This indicates an unknown variable may be influencing the data. Consequently, further studies needed to be undertaken to identify the unknown variable. To continue testing, the next study involved only one technician and nine repeat tests on the same lot of plastic resin. The results yielded a slope of 2.5, a median value of 24 with three-sigma values of plus 36 and minus 22. There is no improvement in test precision or shifting of the Weibull slope to near normal (a slope near 3.5). In addition, the data appear to be into separate distributions due to the occurrence of two parallel graph lines.

In this case, the appearance of two distributions is taken to mean that there is a definite unknown variable acting on the data. So, an added investigation was required. The character of the powder resin could be the problem in filling the mold uniformly. So, powder screening directly into the mold was tested. This testing involved one technician and two consecutive days of testing. The results of this study yielded the following results:

AutoWeibull Data

Resin Fuse-ability with Resin 8-Mesh Screening Employed

Lower 3 Sigma	Lower 2 Sigma	Median Value	Upper 2 Sigma	Upper 3 Sigma	Weibull Slope
First Day Results					
25.4	27.7	32.5	36.3	37.8	5.5
Second Day Results					
23.3	23.5	25.3	29.3	32.0	1.6

The daily three-sigma precision has improved so the screening has helped. However, another variable is still influencing the data as the median shifts considerably in the two separate distributions. In addition, there is a large unacceptable shift in skew ranging from negative slope (5.5) to positive slope (1.6). So, evidence continues that there is another variable influencing the data.

The molding pressure on the plastic resin is applied manually. It is speculated that the sequence of pressure treatment in molding varies from day to day. To study this, an automatic controller for application of pressure was installed to achieve complete molding pressure uniformity. The technician conducted this

testing over a period of four days. Both resin screening and the automatic control of molding pressure were employed. The results obtained are listed below:

AutoWeibull Data

Resin Fuse-ability with Resin Screening and Automatic Pressure Control

Lower 3 Sigma	Median Value	Upper 3 Sigma
First Day Results		
16.5	20.5	25.7
Second Day Results		
17.0	19.5	21.5
Third Day Results		
18.0	22.0	25.5
Fourth Day Results		
19.5	21.5	23.6

The average of these numbers is a median of 21 with a three-sigma span of plus 3.1 and minus 3.2. These data represent a very near normal data span. The total three-sigma span is approximately six units. This span is an almost a ten-fold improvement from the initial three-sigma span of 56 noted.

Example Two: Four Distributions:

A lot of nylon, pigmented black, molded to cable Ty-wraps had an exceptional high level of brittle breaks. To understand this situation, 98 tensile bars were injection molded from the remaining resin. The bars were tested for tensile elongation at room temperature. This was done since none of the broken Ty-wraps were available to examine. The AutoWeibull graph analyses of the subsequent tensile elongation values are shown in Figure 2 below:

Figure 2

Room Temperature Tensile Elongation Data

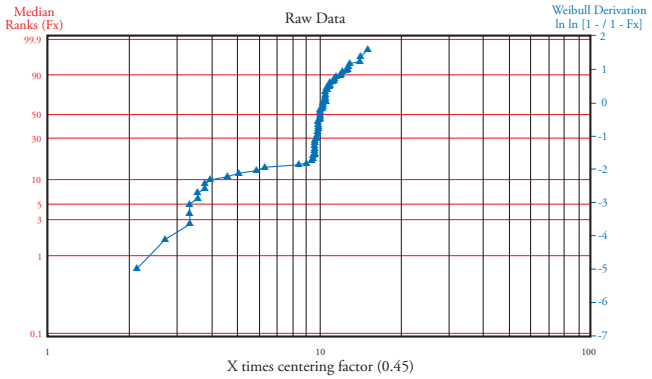


Figure 2 AutoWeibull Graph of Raw Data

In comparison to the previous example, the B graph line (second from bottom in the figure) is a new feature. According to Weibull procedures, the lines, A and B, kinked in this manner, should be combined to a single distribution. However, since the new line is so striking different in comparison to other studies with black nylon, the data are analyzed separately. This indicates that two additional variables are involved.

A previous study showed another variable exists in the large distribution possessing a large positive skew with a 1.5 Weibull slope and a variable influencing the data. However, in this specific new case, data from lines C and D exhibit two distinct distributions. This is because a Weibull plot of the C and D data shows only a 79% least squares fit. So, data for lines C and D are split into two groups. The total AutoWeibull analyses of all of these data are shown below:

AutoWeibull Analysis

% Tensile Elongation @ 2 Inches/minute

Distribution	x_o	Sigma -three	Weibull Mean	Sigma +three	Line Fit %	Slope
A 8%	0	2.3	7.7	11.0	88	5.4
B 6% *	9.4	9.5	14.7	43.0	97	1.0
C 78%	19.9	20.0	22.7	27.4	88	2.2
D 8%	26.3	26.3	29.3	48.3	96	1.0

* Limited number of data points in this data span.

In distribution A, the low elongation values relate to carbon agglomerates.

The fracture area of the tensile bars exhibited agglomerates ranging from 4 to 6 mils in diameter. These agglomerates act as focal points for premature tensile failure.

This nylon resin also possessed a slightly depressed melting point. Another lot of black nylon not shown in this text and used for Ty-wraps lacked this depressed melting point and the graph line B. It is believed the depressed melting point leads to the formation of line B and not just accompanies it. The presence of extremely small carbon black particles, along with degraded nylon, in the neat nylon, relates to this depression. These extremely small diameter carbon black particles nucleate a finer crystalline structure from the melt as it solidifies. It is this fine structure that melts first. The overall high crystallinity of large and smaller crystalline structures enhances resin embrittlement and the low elongation values.

The degraded nylon would be created in the manufacturing conditions for the nylon. Badly worn internal metal parts are the usual cause. This is because the worn localities provide cavities for molten resin hold-up and subsequent resin degradation from prolonged thermal exposure.

In distribution C, the elongation values represent the general character of carbon filled nylon. The carbon black is of optimum size for coloring the nylon. Large particles that would be premature tensile focal points for failure should be absent. Also, there shall be no particles so small that they nucleate the finer crystallites.

In Distribution D, it is evident that some of the localities have the lower percentage of carbon. These localities tend to possess elongation values that tend to be associated with neat nylon's elongation range of 75 to 95 %. The occurrence of high elongation numbers is desirable as long as the nylon meets all specifications.

The overall quality of carbon filled nylon will vary around these extremes. To summarize, the high quality black nylon must be produced so no carbon particle agglomerates exist. In addition, exceptionally fine carbon particles should be absent and no degraded nylon present.

Example Three: Different Distribution Skew and x_o values:

For a wire construction to be listed in the National Electric Code (NEC) certain testing must be completed satisfactorily. For wet-applications, a 36-week water immersion test at 60° and 90° C is used. A 600-volt A.C. stress is continuously applied to the wire constructions. The three wire constructions are made of the same crystalline plastic material, but have slightly different amounts of a stabilizing additive. The constructions are removed from exposure for periodic insulation resistance testing. A specific amount of insulation resistance must be maintained throughout the exposure to pass the test.

After the testing at both temperatures, one type of wire construction failed catastrophically, another failed slowly and the third passed the tests beautifully. Retainer samples of each wire construction and resin were evaluated by dielectric break down testing, corona ignition testing and by visual tests. This testing was undertaken to solve the mystery as to each construction's performance in the long-term test.

The AutoWeibull results for dielectric results are shown below in Figure 3 for the sample at 60° C that failed catastrophically. Two detrimental factors are:

The x_o value is below the 600-volt test stress

The distribution skew is negative

Both factors show there is a continuous electrical corona at this 600-volt stress level within this plastic composition that is destroying it internally.

Figure 3
AutoWeibull Summary of Cable Failing Catastrophically at 60 C

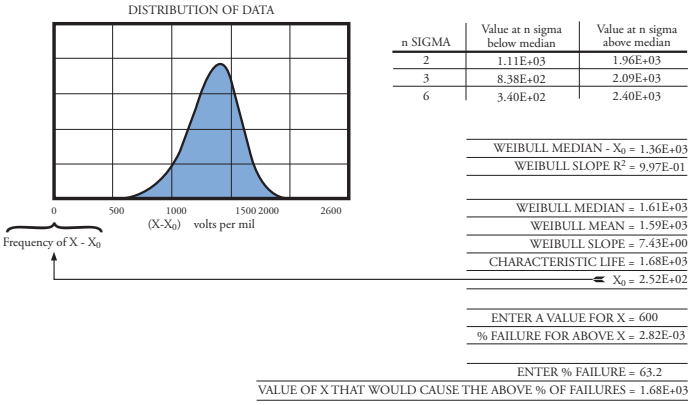


Figure 3 Summary AutoWeibull Analysis

The summary of AutoWeibull data relating to all three-wire constructions is detailed in the Table below.

AutoWeibull Results

(Fifteen samples, for each temperature test (60° & 90° C), each ten foot long)

Sample	Failure	Additive Amount	Lowest Xo volts/min	Range of Weibull Slope		Skew
A	Catastrophic	High	252	4.00	7.00	Negative
B	Slow Failure	Low	1100	3.45	3.55	Near Normal
C	Passes Test	Moderate	1200	1.30	2.44	Positive

The sample having the highest x_o and the positive distribution skew passed this crucial NEC test. More testing would, no doubt, reconfirm this trend with perhaps, slight shifts in x_o and slope.

In preparation for corona testing, molded plaques of resin C were found to be transparent. These plaques had a corona ignition voltage of 101 volts/mil. Similar plaques of resin A were translucent and possessed a corona ignition voltage of only 61 volts/mil. These corona tests relate to when 200 pico-columbs of current flows. Micro voids having a diameter greater than a half-wave length light apparently are the cause of the translucency and the low corona ignition voltage.

Therefore, wire constructions having good performance in this NEC test are accompanied by:

The proper additive concentration

A high x_o voltage value (volts per mil)

A data distribution curve of dielectric strength having positive skew

A high corona ignition voltage

In molded form, good light transparency

With this mystery of premature failure solved, other resins can be so tailored to pass this stringent NEC test requirement.

Example Four: Life Phenomena (Truck Tire Warranty)

The Esex Tire Company makes recap tires for semi-trailers. For a new re-capped tire design, they wished to determine the warranty that could be used to promote tire sales. The following study was designed considering a 3 % return rate on tire failure as acceptable:

Equip 15 trailers with eight tires each.

Determine the mileage of the first tire failure on each the fifteen trailers.

Continue the study to determine the mileage associated with all tire failures.

The results of the testing are shown in the table below:

Table 1
Mileage for Truck Tire Failure

									Minimum
Trailer 1	100,850	25,234	60,854	120,366	68,184	50,028	41,472	78,178	25,234
Trailer 2	47,324	98,690	45,176	132,300	32,050	57,142	87,296	101,650	32,050
Trailer 3	48,944	131,862	103,124	37,550	46,636	101,318	60,270	93,696	37,550
Trailer 4	118,272	101,562	41,350	75,964	57,850	83,894	111,046	87,942	41,350
Trailer 5	43,536	123,648	57,670	55,108	68,876	102,824	106,324	114,540	43,536
Trailer 6	75,316	105,560	102,392	115,582	94,888	47,150	67,318	65,330	47,150
Trailer 7	48,350	66,568	82,754	95,536	79,332	129,554	77,000	106,388	48,350
Trailer 8	93,084	136,342	76,290	70,664	93,960	100,170	106,150	51,728	51,728
Trailer 9	136,712	54,522	119,218	53,410	72,054	103,610	69,814	91,460	53,410
Trailer 10	57,504	129,994	79,656	79,484	62,324	121,218	112,134	104,010	57,504
Trailer 11	136,978	58,680	91,070	67,316	113,692	75,932	106,840	86,506	58,680
Trailer 12	77,362	126,040	125,336	62,982	116,454	101,436	81,040	95,468	62,982
Trailer 13	79,556	109,572	74,864	117,002	139,230	99,614	66,550	136,504	66,550
Trailer 14	108,806	125,266	59,460	122,382	81,448	86,150	68,916	131,996	59,460
Trailer 15	120,404	111,030	129,634	113,646	74,872	138,018	129,554	111,722	74,872

The AutoWeibull computer program was used in analyzing the data. This was to determine the mileage guarantee that can be considered for the 3 % tire failure. The likelihood of failures with these tires, occurring shortly after mounting on the trailers, is not anticipated. It's remotely possible, but certainly not normally expected. This is because the re-capped tires are well designed, carefully manufactured and have already passed through a quality control evaluation. The first failures are expected to be at some finite time. So, the analysis started in the following sequence:

Enter the 15 minimum values from each truck from the data chart.

Set x sub zero (x_o) to be calculated by clicking the "Calculate x_o" icon button.

Activate the Run" button.

"Accelerated Sudden Death" graph appears. The sequence continues as follows.

On the yellow line, the number of replicates is entered (eight tires per trailer).

On the blue line, the value of three percent expected failures is entered.

Automatically, the number of tires failing at 3 % appears on the line below the blue line.

This entire sequence is continued again with x_o set to be zero to consider an alternate set of conditions.

These conditions involve the unlikely state where there is a failure of a tire occurring moments after tire mounting. In this case, following the same above sequence but alternately, the "x_o equals zero" icon button is depressed. The analysis is continued through to the "Sudden Death" sequence to obtain another 3 % value. The results obtained from both analyses are shown below:

3 % Failure	Mileage	Weibull Slope	Least Squares Fit (Percent)	x_o value (Mileage)
3 % failure (x_o calculated)				
	33,400	3.4	99.3	5,597
3 % failure (x_o equals zero)				
	39,200	3.9	99.6	zero

To formulate what might be a more balanced warranty mileage, the average of the two 3 % values (33,400 and 39,200) is obtained as 36,300 miles. So, a warranty of 36,000 miles is proposed. Full tire value will be given for tires failing at less than 10,000 miles and a mileage pro-rated value for tires failing between 10,000 miles and 36,000 miles.

Further procedure involved continual testing of the truck tires until all have failed. Several years later, the AutoWeibull program was used to analyze the 120 data points (eight tires on 15 trailers) when all the data became available. This was done to double check the earlier "Sudden Death Analysis". The calculation proceeded by calculating x_o and also, letting x_o be zero. The results are shown below:

3 % Failure	Mileage	Weibull Slope	Least Squares Fit (Percent)	x_o value (Mileage)
3 % failure (x_o calculated)				
	36,800	3.1	98.8	7,085
3 % failure (x_o equals zero)				
	36,100	3.4	98.5	zero

Note that the x_o value of 7,085 miles is a higher value and fairly close to the x_o value of 5, 597 miles initially calculated. This x_o analysis here probably is the most accurate as the least squares fit is numerically 0.3 % better then the other. In addition, this analysis shows the initial warranty of 36,000 miles was a very good selection. So, based upon these confirming data, the warranty is continued. Further analysis shows the number of tires that failed at mileages less than 10,000 is a small number. Full replacement was promised for those tires in the warranty. These numbers are as follows:

Number of Tires failing at less Than 10,000 miles (x_o calculated)	Percentage Failed before 10,000 miles
24 tires per 1,000,000 tires sold (x_o equals zero)	0.00238
371 tires per 1,000,000 tires sold	0.0371

It is believed the warranty initially proposed would continue to be successful.

Article Synopsis:

This article reviewed the ***AutoWeibull*** computer program and its use of three-parameter Weibull. Several technical dilemmas have been solved through the use of AutoWeibull in testing for data consistency. The technique is easy to use.

Two parameter Weibull, the alternate technique, is used for one example and is used extensively for life phenomenon predictions in industry. The book, ***The Weibull Distribution Function®***, covers that aspect as well as more details on three-parameter Weibull.

Reference:

1) "*A Statistical Distribution Function of Wide Applicability*," by Waloddi Weibull. Journal of Applied Mechanics, September 1951.

Scientific Article

DEVELOPMENT AND VALIDATION OF LOGISTIC PROGNOSTIC MODELS USING PREDEFINED SAS MACROS

Rainer Muche[1], Christina Ring[1] and Christoph Ziegler[2]

[1] Institute of Biometrics, University of Ulm, Germany
[2] Biostatistics Department, F. Hoffmann-La Roche Ltd, Switzerland

E-mail: rainer.muche@uni-ulm.de

Abstract

Medical therapies or diagnostic procedures are influenced by either prognosis or by magnitude of a disease. In addition to the subjective assessment by a clinician, mathematical models are used for providing a prognosis in clinical situations. Such models are frequently multivariate regression models. In the case of a dichotomous outcome the multiple logistic regression model is applied as the standard model. In this paper we describe SAS macros for the development of such a model. Additionally, macros for the examination of its prognostic performance and model validation are presented.

This set of 14 SAS macros serves as a tool for setting up the whole process of deriving a prognostic model based on the logistic regression model. In particular, the provision of model validation, which is rarely used in practice, will help getting much better prognoses in future applications.

Keywords: prognostic model, logistic regression, model validation, SAS macro

1 Introduction

"Prognosis is a prediction of the future course of a disease from its conception"[7]. In clinical situations, the prognosis of the course of the disease should be carried out as accurately as possible for assisting therapeutic decision making. In addition to the subjective assessment of the clinicians, mathematical models are used for providing such prognoses. In most situations, multivariate regression models are used as prognosis models. In the case of a dichotomous outcome the multiple logistic regression model is frequently applied as the standard model [11].

This paper presents a practical, ready to use analysis strategy and SAS macros for the computation of prognosis modeling using the logistic regression model. The modeling and verification of the prognostic performance is carried out in 3 steps: (1) model development, (2) assessment of prognostic performance, and (3) model validation.

2 Logistic Regression

The logistic regression model has long been the standard procedure for analyzing binary outcomes [11]. The occurrence of an event (Y=1) given the covariates $X_1, X_2, \ldots, X_k$ is modeled as follows:

$$P\left(Y = 1 \middle| X_j = x_j\right) = \frac{1}{1 + \exp\left(-\left(\alpha + \sum_j \beta_j x_j\right)\right)} \qquad j = 1,K,k$$

The regression coefficients β_i are usually determined by the maximum likelihood approach. The logistic regression can be carried out with several SAS procedures: PROC LOGISTIC, PROC CATMOD, PROC GENMOD, PROC PROBIT. We use the procedure PROC LOGISTIC in our macros.

3 Prognostic Modeling Implementation

We suggest proceeding towards prognostic modeling in 3 steps: (1) model development, (2) prognostic performance assessment, and (3) model validation, as shown in Figure 1.

In the first step, the model prerequisites and assumptions are verified by looking at descriptive measures, collinearity and influential observations. Sometimes, a missing value imputation is required to fulfill the requirements. In the next step, the modeling by univariate and/or multiple logistic regression can be carried out. An assessment of the goodness-of-fit of the resulting model completes the model development.

The performance of the model is assessed by ROC analysis using a reclassification of the observations by the model. Because of an overestimation of the performance measures in the ROC analysis, a model validation has to be performed afterwards. This is done in the third step, using one of 5 validation procedures provided in the macro package.

At the end of this procedure, a prognostic model and realistic prognosis measures are obtained for assessing whether the model is useful in clinical practice.

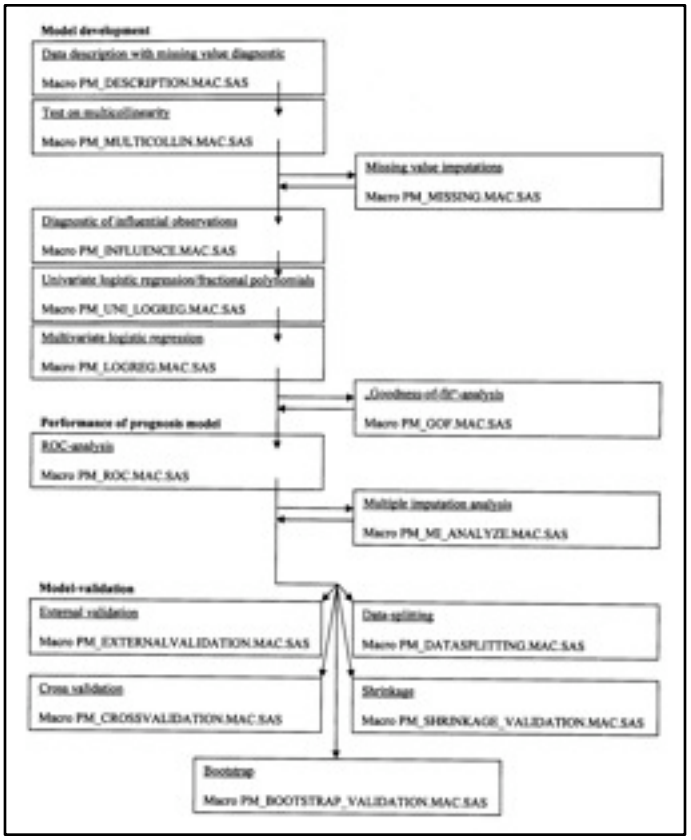


Figure 1: Prognostic and diagnostic modeling steps based on logistic regression methods, available SAS macros, and dependencies [13]

4 Procedures within the SAS Macros

This chapter provides a brief description of the macros. An extended and elaborated description is provided in [13] and [19], respectively (in German). The prognosis modeling is divided into 3 main parts as shown in the previous chapter:
model development, prognostic performance assessment, and model validation.

4.1 Macros for model development

Different assessments of model prerequisites and assumptions based on the data are required before the model development by logistic regression is actually carried out. This step includes analyzing the variables univariately (descriptively), analyzing the relationship between the variables (multi-collinearity), as well as the analysis of the impact of individual observations (influential observations). Another problem in modeling is the effect of missing values that could influence the results. The following macros given in table 1 help in carrying out the evaluation before the data are adapted to the logistic regression model procedure.

Macro-Name	Short description
PM_DESCRIPTION.MAC.SAS	Description of outcome and influential variables
PM_MULTICOLLIN.MAC.SAS	Multi-collinearity analysis of influential variables
PM_MISSING.MAC.SAS	Testing for and substitution of missing values
PM_INFLUENCE.MAC.SAS	Identification of influential observations

Table 1: SAS macros for checking the assumptions for model development

Using the macro PM_DESCRIPTION.MAC.SAS enables all influential variables as well as outcomes to be evaluated by univariate descriptive methods. Depending upon whether discrete or continuous variables are analyzed, the SAS procedures PROC

UNIVARIATE or PROC FREQ are used for analysis. By using the parameter miss= the user can decide whether a complete case analysis or all observations are evaluated. For assessing the missing value situation in the data, not only the number of missing values per variable but also the number of missing values per observation is displayed.

Testing for multicollinearity in PM_MULTICOLLIN.MAC.SAS is performed by
- pairwise correlation (Spearman, PROC CORR)
- variance inflation factors (VIF, PROC REG)
- eigenvalue analysis ([3], PROC REG / COLLINOINT)
The evaluation of the principal component analysis in PROC REG is weighted and carried out by PROC LOGISTIC with estimated probabilities [1].

The missing values can be handled in the following ways:
- Complete-case-analysis (miss=0)
- Single Imputation (continuous: PROC STDIZE, discrete: an extra category MISSING)
- Multiple Imputation (Checking for missing pattern, PROC MI)

The macro PM_INFLUENCE.MAC.SAS identifies the most influential observations for the modeling. It mainly involves an assessment of the Pearson-statistic after one observation has been removed. Large changes suggest a high influence on the parameter estimation. Step by step, the most influential observations are identified (until a given fixed value is reached), however they are not automatically removed from the data.

Macro-Name	Short description
PM_UNI_LOGREG.MAC.SAS	Univariate logistic regression
PM_LOGREG.MAC.SAS	Multivariate logistic regression including variable selection
PM_GOF.MAC.SAS	Goodness-of-fit evaluation of the resulting model

Table 2: SAS macros for model development

In PM_UNI_LOGREG.MAC.SAS, an individual logistic regression model is fit for every variable and the corresponding p-value is calculated. One can refer to the complete-case-data record for this purpose (miss=0). Categorical variables are used as dummy variables in the model, and tests are performed on these dummy variables using the CLASS statement. Continuous variables are used in the model as linear terms. Additionally, the assumptions for modeling continuous variables are assessed using ”fractional polynomials“ up to degree 2 [15].

The multiple logistic regression analysis is carried out using the main macro PM_LOGREG.MAC.SAS. Using this macro a multiple logistic regression model is fit with stepwise variable selection. This macro provides special output files that can be used for further analysis (ROC, model validation). For assessing the influence of variables, the TEST-statement can be used that does not work in parallel with the CLASS statement in PROC LOGISTIC. If problems arise during parameter estimation, particularly by a quasi-complete separation, then a corrected estimation method is executed automatically by the FL-macro [9].

The macro PM_GOF.MAC.SAS provides tests for model adaptability. Along with parameters from PROC LOGISTIC, special tests for sparseness from the literature (few observations per parameter combination) are integrated (macros from [12], [14]), because in this situation the usual Hosmer-Lemeshow-test is not suitable.

4.2 Macro for testing the prognostic performance

Testing the prognostic performance requires answering the question**: ”How well can the outcome of the patient be predicted by the model in advance?“** The assessment of the prognostic performance is carried out by means of a reclassification. In doing so, the data of the patient is used in the model and hence the probability of an outcome for the patient is esti-

mated. A comparison with the observed values helps in assessing the conformity.

By choosing a cut point, these probabilities can be characterized as “big“ or ”small”. By correlating this with the observed values, parameters like sensitivity, specificity, predictive value, Youden-index etc. can be calculated. Additionally, global measures for prognostic performance are indicated, including AUC, Somer´s D, Emax, Brier Score etc. (all independent of a particular cut point). These parameters are derived within the scope of a ROC analysis.

Macro-Name	Short description
PM_ROC.MAC.SAS	ROC analysis (cut point-dependent and independent prognostic performance including graphics)
PM_MI_ANALYZE.MAC.SAS	Summary of prognostic performance for multiple imputation of missing values

Table 3: SAS macros for prognostic performance

The prognostic performance is executed using a ROC analysis with macro PM_ROC.MAC.SAS. All mentioned diagnostic measures (additionally the confidence interval for AUC) as well as some important graphs (like ROC-curves (incl. confidence bounds [10]), relationship between the Youden-Index and Cutpoint) are displayed.

Finally, the prognostic performance (in logistic regression and ROC-analysis) using multiple imputation for the missing values can be derived by the macro PM_MI_ANALYZE.MAC.SAS.

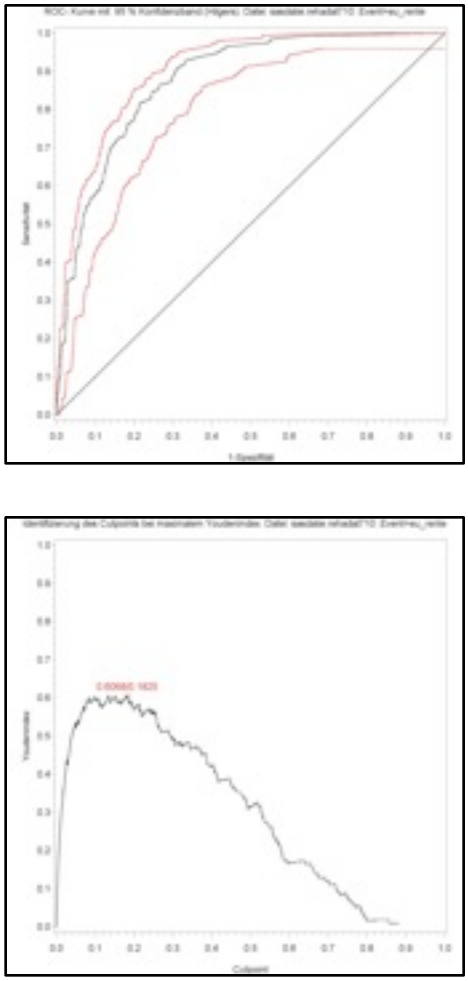


Figure 2: Example of graphical output from the macro PM_ROC.MAC.SAS

4.3 Macros for model validation

After the analysis for prognostic performance it can be decided whether the prognostic model is reliable, i.e. generates few errors in prognosis. The question **“How reliable is the prognostic performance for independent observations?“** has not yet been answered. The problem is that the prognostic performance after reclassification is determined using the same patient data used for estimation of the regression coefficients by the maximum likelihood approach. Thus, there is a possibility of a bias for a too optimistic prognostic performance after ROC analysis.

For analyzing this bias, a model validation should be performed. There are several procedures available that are implemented in the following five macros. In the literature, along with external validations the Bootstrap method is preferred [17]. Output would be the cutpoint -dependent and independent prognostic performance measures of ROC analysis before and after the validation as well as the absolute and relative changes.

Macro-Name	Short description
PM_EXTERNAL_VALIDATION.MAC.SAS	External validation using two data records
PM_DATASPLITTING.MAC.SAS	Splitting up the data records into several parts
PM_CROSSVALIDATION.MAC.SAS	Cross-validation
PM_BOOTSTRAP_VALIDATION.MAC.SAS	Bootstrap method for model validation
PM_SHRINKAGE.MAC.SAS	Coefficient correction using shrinkage

Table 4: SAS macros for model validation

During external validation the prognostic performance is determined using two independent data bases. This method should be preferred if a second data record exists. Unfortunately this is rarely the case and methods are assessed that are based on available data (internal validation methods).

Using data splitting the data record is divided into two parts. One part is used for model development and the other part is used for validation (see external validation). Thereby the model development and model validation is based on a significantly lower number of cases and hence this procedure is rarely used. The macro splits the data randomly based on given proportions. Finally, the macro PM_EXTERNAL_VALIDATION .MAC.SAS is invoked for validation.

Cross-validation has been the standard procedure for model validation for a long time. This procedure is based on sampling schemes without putting the drawn patient data back. The procedure can be explained as follows: The data record is divided in K parts; subsequently the model is developed in K-1 parts and validated in the K-th partition. The whole procedure would be repeated for all K partitions and possibly for several random splits into the K subsamples. The following methods are programmed in the macro: K-fold cross validation, adjusted cross validation [5], Jackknife-cross-validation.

The Bootstrap-validation [2] is also a resampling-procedure, based on sampling without replacement: databases of the same length as the underlying database are created from the available data. Using these new data, the modeling and/or validation of the models takes place. By appropriate merging of the individual results, the bias for prognostic performance can be estimated. The following proposals from Efron [6] are implemented: simple- / enhanced bootstrap as well as the approach of averaging the regression coefficients (Mean Model).

The shrinkage method correlates the estimated regression coefficient [18] such that the prognostic performance is determined using the correlated model. Three methods are implemented in the macro: heuristic Shrinkage, global Shrinkage [18], and parameter-related shrinkage factors [16].

5 Macro Invocation, Validation and Technical Prerequisites

The invocation is similar for all macros. In Figure 3, the macro invocation with the main parameters is displayed. Along with other parameters, the data record (data=), the outcome with event of interest (resp_var=, event=) and the variables (discrete: cvar=, continuous: xvar=) are given. With the parameter miss= , the observations with missing values can be excluded from the analysis for a complete case analysis.

The macros (for SAS version 8.2) were extensively validated. The validation report is available for

download from the internet page given in Chapter 6 (in German). After revising the macros for SAS version 9 because of some changes in the ODS report module in SAS, some of the validation programs were repeated showing the same results as for version 8.2 macros.

For effective usage of the macros, few hard- and software requirements have to be taken into account. The minimum requirements are:

- SAS version 8.2 or higher (SAS 9 macros available)
- SAS modules BASE, STAT, GRAPH, IML
- Hardware requirements for SAS 8.2, 9 and higher (recommended: RAM 512 Mb, Processor > 1 GHz)

The SAS macros use several external programs, among them many diagnostic programs. The total macro package consists of around 100 programs and files. Hence there are some prerequisites:

- the total macro package should be stored in one folder (invocation of macro_path=),
- the variables to be evaluated have to be numeric,
- the variables should be formatted numerically (preferably with no formats used),
- a special variable is required that uniquely identifies the observations (like patient number).

```
%MACRO macroname (  data        =,
                    resp_var    =,
                    event       =1,
                    xvar        =,
                    cvar        =,
                    ref         =,
                    miss        =0,
                    ...
                    weitere spezifische Parameter,
                    ...
                    macro_path  =,
                    );
%MEND macroname;
```

Figure 3: General invocation of the SAS macros

6 Summary

The important problems in modeling clinical prognosis are summarized by Feinstein [4] as follows: frequently no specified definition of the variables, multi-collinearity, non-consideration of influential observations, non-fulfilled model assumptions, non-linear correlation, over-conformity, unspecified variable selection, no interdependency checks as well as missing model validation.

The introduced strategy for model development and validation using our SAS macro package considers all these problems and ensures the adequacy of assumptions for future prognostic modeling based on logistic regression. Thus, the macros described here will help to ensure reliable prognosis in biometric and medical statistics to improve clinical decision making.

The macros are available for download from http://www.uni-ulm.de/med/med-biometrie/forschung/sas-makros-fuer-prognosemodellierung-mit-logistischer-regression.html. Usage and background information is available in [13] (in German).

Literature

1. Allison PD. Logistic Regression using the SAS System. Cary NC: SAS Institute Books By Users; 1999
2. Assfalg I. Die Bootstrap-Methode zur internen Validierung von Prognosemodellen. Diplomarbeit FH Ulm; 2003
3. Belsley DA. Conditioning diagnostics – Collinearity and weak data in regression. New York: John Wiley & Sons; 1991
4. Concato J, Feinstein AR, Holford TR. The risk of determining risk with multivariable models. Ann. Intern. Med. 1993;118: 201-210
5. Davison AC, Hinkley DV. Bootstrap methods and their application. Cambridge: Cambridge University Press; 1997
6. Efron B, Tibshirani RJ. An Introduction to the Bootstrap. New York: Chapman & Hall; 1993
7. Fletcher RM, Fletcher SW, Wagner EH. Klinische Epidemiologie. Wiesbaden: Ullstein Medical Verlag; 1999

8. Harrell FE Jr. Regression Modeling Strategies. New York: Springer Verlag; 2001

9. Heinze G, Schemper M. A solution to the problem of separation in logistic regression. Stat. Med. 2002;21:2409-19

10. Hilgers R. Distribution-free confidence bounds for ROC curves. Meth. Inform. Med. 1991;30:96-101

11. Hosmer DW, Lemeshow S. Applied Logistic Regression (2nd Edition). New York: John Wiley & Sons; 2000

12. Kuss O. Global goodness-of-fit-tests in logistic regression with sparse data. Stat. Med. 2002;21:3789-801

13. Muche R, Ring Ch, Ziegler Ch. Entwicklung und Validierung von Prognosemodellen auf Basis der logistischen Regression Aachen: Shaker Verlag; 2005

14. Pulkstenis E, Robinson TJ. Two goodness-of-fit tests for logistic regression with continuous covariates. Stat. Med. 2002;21:79-93

15. Royston P, Altman DG. Regression using fractional polynomials of continuous covariates. Appl. Statist. 1994;43:429-67

16. Sauerbrei W. The use of resampling methods to simplify regression models in medical statistics. Appl. Statist. 1999;48:313-29

17. Steyerberg EW, Harrell FEJr., Borsboom GJJM et al. Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. J. Clin. Epidemiol. 2001;54:774-81

18. van Houwelingen H, LeCessie S. Predictive value of statistical models. Stat. Med. 1990;9:1303-25

19. Ziegler Ch. Ein SAS-Makro-Paket zur Entwicklung und Validierung von Prognosemodellen auf Basis der logistischen Regression. Diplomarbeit FH Ulm; 2004

Technology and Commerce Corner

In the last issue we introduced a new section on Technology and Commerce. We hope to encourage all kinds of folks, especially entrepreneurs, who are working with current computational and graphical technologies, to tell us about their work.

Last issue, we included an interview with the New York Times graphics editor, Steve Duennes on our website.
http://blog.stat-computing.org/2008/05/talk-to-newsroom-interview-with-steve.html

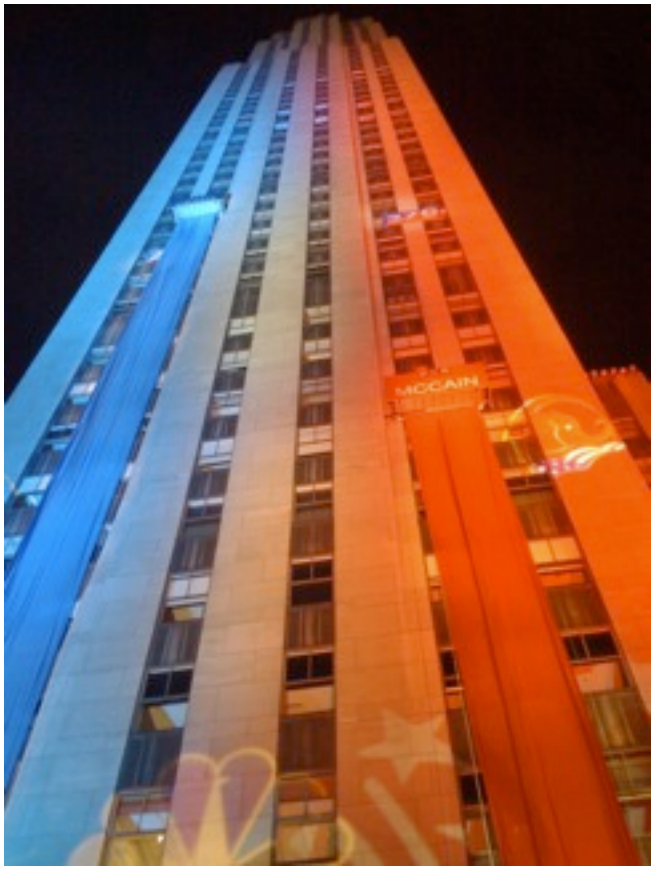
We didn't get any contributions for this issue from our readership. We encourage folks to send us contributions for next issue. Any tidbit can be interesting!

We provide a link to the NY Times election results below:

http://elections.nytimes.com/2008/results/president/map.html

Fascinating stuff - look at the voting shifts view. It looks like the entire country except the appalachian states (Tennessee, W Virginia, Kentucky), Arkansas/ Oklahoma and some areas around the gulf of Mexico (Louisiana) voted more Democratic in this election compared to the last election.

I was in NY the night of the election. The outdoor 'graphics' at the Rockefeller Center - the NBC building and the ice rink - were quite something as well.



The NBC building was lit half in red and half in blue with outdoor elevators for Obama and McCain. As results came in the elevators moved up the building towards the 270 target electoral votes. The photo above shows Obama getting past the 270 target after the west coast states were called.

The Rockefeller Center ice rink had a map of the US drawn across its surface. Each time a state was called it was spray painted red or blue. The photo below was taken after the west coast states were called and shows the blue on each coast and across the top.

Please send us some contributions for the next newsletter. A link to a YouTube video, a link to an interesting graphic, anything that catches your eye. You'd be surprised at how little bits and pieces can lead to new ideas. YouTube itself was started by a couple guys who were frustrated in not being able to find a clip of the Janet Jackson superbowl wardrobe malfunction....

- Michael O'Connell

The Graphics program at JSM 2009

Steve MacEachern
Program Chair

In 2009, the Section on Statistical Graphics has a hard act to follow. The 2008 JSM was characterized by a fine set of well-attended invited sessions put together by program chair David Hunter. The few technical problems with the ASA's session management system were handled well by the session chairs and speakers. As always, the Monday evening mixer was a memorable event, featuring discussion of newly proposed sections for the ASA and howling, flying monkeys, in addition to the ever popular raffle.

The Graphics Section is organizing four sessions of invited talks at JSM 2009 in Washington, D.C. Thanks to Bill Cleveland, Stan Wasserman and Ann McCranie, Linda Pickle, Naomi Robbins and Rich Heiberger for their work on organizing the sessions and for largely writing the descriptions below. The sessions of talks are:

Data Display for Large Complex Datasets

Large, complex data sets are ubiquitous, the standard now rather than the exception. They present challenging problems for their visualization because of their size and the complexity of their data structures and patterns. Wholly new approaches, methods, and computational algorithms for data display are needed. These new methodological developments lie at the interface of Statistics and Computer Science, with healthy borrowings from other disciplines. This session will feature a suite of talks by experts working at the interface.

Making Statistics Go Viral: Visualizing Illness in Social Networks

Network science lends itself to striking and powerful images, but interpreting those graphics can be tricky business. For individuals who do policy-relevant research, clear specification of the meaning of visualization is key, as the images themselves can be quite evocative and can garner much attention. This session features talks by scholars who have significant experience with policy-relevant social network visualization, in particular as it relates to the paths of disease transmission. Attention will be given to discussion of how network data can be conveyed to the public and policy makers in clear and meaningful terms.

The Influence of Psychology, Cartography and Computer Science on the Design of Interactive Graphics for Spatial Statistical Data

Increasingly powerful computer hardware and software have led to the development of tools to display and explore spatial statistical data--data sets that heretofore were too large and complex for dynamic interactive display. Design principles for these new tools are influenced by research in disciplines other than Statistics. For example, cartographers have found that certain map designs are better able than others to convey quantitative data patterns to the reader. Computer scientists have experimented with various methods for efficiently filtering and drilling down to specific information with graphical user interface (GUI) tools. Psychologists have found that people vary in their visual, cognitive and analytic skills, but that there are cognitive strengths and weaknesses that most of us share. This session will include speakers who will address interactive graphical design for spatial statistical data from the points of view of Cartography/Geography, Computer Science, Cognitive Psychology and Statistics.

Professional Statistics and Graphics with Spreadsheets Using the R-Excel Interface

Many people use Excel, some even use it for their statistical analyses and production of their graphs. Providing superior tools to these users is of utmost importance. The RExcel interface provides a spreadsheet front end to the superior graphics and statistical analyses available in R. This session has talks that feature practical means by which R and Excel can be combined to produce a powerful, practical tool for statistical analysis and presentation.

In addition to the invited sessions, I hope that many of you will consider organizing a topics contributed session. These sessions get a little more visibility in the program, a little more time for speakers, and have the potential to increase the number of invited sessions allocated to the Graphics Section. If you have ideas for a topics contributed session, contact me at snm@stat.osu.edu.

Steve MacEachern

Photos from the 2008 JSM Statistical Computing and Graphics Mixer.

News

2009 CHAMBERS AWARD

JR Lockwood,
Awards Officer, 2009
Statistical Computing Section

The Statistical Computing Section of the American Statistical Association announces the competition for the John M. Chambers Statistical Software Award. In 1998 the Association for Computing Machinery presented its Software System Award to John Chambers for the design and development of S. Dr. Chambers generously donated his award to the Statistical Computing Section to endow an annual prize for statistical software written by an undergraduate or graduate student. The prize carries with it a cash award of $1000, plus a substantial allowance for travel to the annual Joint Statistical Meetings where the award will be presented.

Teams of up to 3 people can participate in the competition, with the cash award being split among

team members. The travel allowance will be given to just one individual in the team, who will be presented the award at JSM. To be eligible, the team must have designed and implemented a piece of statistical software. The individual within the team indicated to receive the travel allowance must have begun the development while a student, and must either currently be a student, or have completed all requirements for her/his last degree after January 1, 2007. To apply for the award, teams must provide the following materials:

Current CV's of all team members.

A letter from a faculty mentor at the academic institution of the individual indicated to receive the travel award. The letter should confirm that the individual had substantial participation in the development of the software, certify her/his student status when the software began to be developed (and either the current student status or the date of degree completion), and briefly discuss the importance of the software to statistical practice.

A brief, one to two page description of the software, summarizing what it does, how it does it, and why it is an important contribution. If the team member competing for the travel allowance has continued developing the software after finishing her/his studies, the description should indicate what was developed when the individual was a student and what has been added since.

Access to the software by the award committee for their use on inputs of their choosing. Access to the software can consist of an executable file, Web-based access, macro code, or other appropriate form. Access should be accompanied by enough information to allow the judges to effectively use and evaluate the software (including its design considerations.) This information can be provided in a variety of ways, including but not limited to a user manual (paper or electronic), a paper, a URL, online help to the system, and source code. In particular, the entrant must be prepared to provide complete source code for inspection by the committee if requested.

All materials must be in English. We prefer that electronic text be submitted in Postscript or PDF.

The entries will be judged on a variety of dimensions, including the importance and relevance for statistical practice of the tasks performed by the software, ease of use, clarity of description, elegance and availability for use by the statistical community. Preference will be given to those entries that are grounded in software design rather than calculation. The decision of the award committee is final.

All application materials must be received by 5:00pm EST, Monday, February 23, 2009 at the address below. The winner will be announced in May and the award will be given at the 2009 Joint Statistical Meetings.

Information on the competition can also be accessed on the website of the Statistical Computing Section (www.statcomputing.org or see the ASA website, www.amstat.org for a pointer), including the names and contributions of previous winners. Inquiries and application materials should be emailed or mailed to:

Chambers Software Award
c/o J.R. Lockwood
The RAND Corporation
4570 Fifth Avenue, Suite 600
Pittsburgh, PA 15213
lockwood@rand.org

Meeting Roundup

UseR

by Uwe Ligges

The international R user conference `useR! 2008' took place at the Technische Universitat Dortmund, Dortmund, Germany, August 12-14, 2008.

This world-wide meeting of the R user community focussed on
- R as the `lingua franca' of data analysis and statistical computing;

- providing a platform for R users to discuss and exchange ideas about how R can be used to do statistical computations, data analysis, visualization and exciting applications in various fields;
- giving an overview of the new features of the rapidly evolving R project.

The program comprised invited lectures, user-contributed sessions and pre-conference tutorials.

More than 400 participants from all over the world met in Dortmund and heard more than 170 talks. It was a pleasure for us to see that the lecture rooms were available after heavy weather and some serious flooding of the building two weeks before the conference.

People were carrying on discussions the whole time: before sessions, during the coffee and lunch breaks, and during the social events such as
- the reception in the Dortmund city hall by the Mayor (who talked about the history, present, and future of Dortmund), and
- the conference dinner in the `Signal Iduna Park' stadium, the famous home ground of the football (i.e. soccer) team `Borussia Dortmund'.

Buses brought the participants to the stadium where a virtual game was going on.

After the dinner, Marc Schwartz got the birthday cake he really deserves for his work in the R community.

It was a great pleasure to meet so many people from the whole R community in Dortmund, people who formerly knew each other only from `virtual' contact by email and had never met in person.

Pre-conference Tutorials

Before the start of the official program, half-day tutorials were offered on Monday, August 11.

In the morning:

- Douglas Bates: Mixed Effects Models
- Julie Josse, Francois Husson, Sebastien Li: Exploratory Data Analysis
- Martin Maechler, Elvezio Ronchetti: Introduction to Robust Statistics with R
- Jim Porzak: Using R for Customer Segmentation
- Stefan Roping, Michael Mock, and Dennis Wegener: Distributed Data Analysis Using R
- Jing Hua Zhao: Analysis of Complex Traits Using R: Case studies

In the afternoon:

- Karim Chine: Distributed R and Bioconductor for the Web
- Dirk Eddelbuettel: An Introduction to High-Performance R
- Andrea S. Foulkes: Analysis of Complex Traits Using R: Statistical Applications
- Virgilio Gomez-Rubio: Small Area Estimation with R
- Frank E. Harrell, Jr.: Regression Modelling Strategies
- Sebastien LÍ, Julie Josse, Francois Husson: Multiway Data Analysis
- Bernhard Pfaff: Analysis of Integrated and Co-integrated Time Series

At the end of the day, Torsten Hothorn and Fritz Leisch opened the Welcome mixer by talking about the close relationship among R, the useR! organizers, and Dortmund.

Invited Lectures

Those who attended had the opportunity to listen to several talks. Invited speakers talked about several hot topics, including:

- Peter Bohlmann: Computationally Tractable Methods for High-Dimensional Data
- John Fox and Kurt Hornik: The Past, Present, and Future of the R Project, a double-feature presentation including
  + Social Organization of the R Project (John Fox),
  + Development in the R Project (Kurt Hornik)
- Andrew Gelman: Bayesian Generalized Linear Models and an Appropriate Default Prior
- Gary King: The Dataverse Network
- Duncan Murdoch: Package Development in Windows
- Jean Thioulouse: Multivariate Data Analysis in Microbial Ecology - New Skin for the Old Ceremony
- Graham J. Williams: Deploying Data Mining in Government - Experiences With R/Rattle

User-contributed Sessions

User contributed sessions brought together R users, contributors, package maintainers and developers in the S spirit that `users are developers'. People from different fields showed us how they solve problems with R in fascinating applications. The scientific program was organized by members of the program committee, including

Micah Altman, Roger Bivand, Peter Dalgaard, Jan de Leeuw, Ramon Diaz-Uriarte, Spencer Graves, Leonhard Held, Torsten Hothorn, Francois Husson, Christian Kleiber, Friedrich Leisch, Andy Liaw, Martin Maechler, Kate Mullen, Eiji Nakama, Thomas Petzoldt, Martin Theus, and Heather Turner,

and covered topics such as
- Applied Statistics & Biostatistics
- Bayesian Statistics
- Bioinformatics
- Chemometrics and Computational Physics
- Data Mining
- Econometrics & Finance
- Environmetrics & Ecological Modeling
- High Performance Computing
- Machine Learning
- Marketing & Business Analytics
- Psychometrics
- Robust Statistics
- Sensometrics
- Spatial Statistics
- Statistics in the Social and Political Sciences

- Teaching
- Visualization & Graphics
- and many more

Talks in Kaleidoscope sessions presented interesting topics to a broader audience, while talks in Focus sessions led to intensive discussions on the more focussed topics.

As was the case at former useR! conferences, the participants listened to talks the whole time;
it turned out that the organizers underestimated the interest of the participants for some sessions
and chose lecture rooms that were too small to accommodate all who wanted to attend.

More Information

A web page offering more information on `useR! 2008' is available at http://www.R-project.org/useR-2008/.

That page now includes slides from many of the presentation as well as some photos of the event. Those who could not attend the conference should feel free to browse and see what they missed, as should those who attended but were not able to make it to all of the presentations that they wanted to see.

The next useR! 2009 will be hosted by Agrocampus Ouest, Rennes, France, July 08-10, 2009. Information is available at http://www.R-project.org/useR-2009/.

For the organizing committee,

Uwe Ligges

Puzzles

We had quite a bit of interest in puzzle #2 from the last issue of the newsletter. We had two entries that pretty much nailed the problem. They were from Tim Hesterberg and Mike Olsen. We also had a hand derivation of the solution to puzzle #1 from John Monahan.

If anyone has any more interesting puzzles, please send them in !!

Question 1
Given a square piece of property of unit side you wish to build fences so that it is impossible to see through the property, i.e. there is no sightline connecting two points outside the property and passing through the property that does not intersect a fence. The fences do not have to be connected and several fences can come together at a point.
The fences can be placed in the interior of the property, they aren't restricted to the boundary. What is the minimum total length of fencing required and how is it arranged. For example you could place fencing along all four sides. This would have total length 4 but is not the best possible.
Hint: You can do better than 2*sqrt(2)

Question 2
A banana plantation is located next to a desert. The plantation owner has 3000 bananas that he wants to transport to the market by camel, across a 1000 kilometer stretch of desert. The owner has only one camel, which carries a maximum of 1000 bananas at any moment in time, and the camel eats one banana every kilometer it travels.

What is the largest number of bananas that can be delivered at the market?

Hint: You can cache bananas in the desert.

Answer: Tim Hesterberg 533 1/3; Mike Olsen 532

Seeking New Awards Chair for ASA Statistical Computing and Graphics Sections

The Awards Chair is responsible for coordinating and overseeing the annual Student Paper Competition, the Chambers Award, and new this year, the JSM Best Contributed Poster Competition (this replaced the JSM Best Contributed Paper Competition). Responsibilities include securing participation of and coordinating with judges, preparing and distributing competition announcements, receiving entries from and coordinating with applicants, overseeing final decisions to ensure fair implementation, and coordinating with ASA officers to give the awards to winners including plaques, checks, and JSM travel reimbursement.

J.R. Lockwood, the current awards chair, is ending his three-year term this year, and we are seeking a new chair to begin a three-year term in September 2009. J.R. would work closely with the new chair to manage the transition and would be available throughout the year to help answer questions.

Serving as chair is a great opportunity to get to know prominent members of the statistical computing and graphics communities, as well as the students and other young researchers participating in the competitions. It is also a valuable professional service function to both the ASA and the sponsoring Sections.

Anyone interested in learning more should contact J.R. at lockwood@rand.org.

Seeking New Newsletter Editor for ASA Statistical Computing Section

The Newsletter Editor for the Computing Section works with the Newsletter Editor for the Graphics Section (Andreas Krause) to publish the Statistical Computing and Graphics newsletter twice per year. They are responsible for coordinating content from the Computing and Graphics Section chairs (currently Deb Nolan for Computing and Dan Rope for Graphics), the JSM program chairs (currently Steve MacEachern for Graphics and Rob McCulloch for Computing) and other content such as brief scientific articles and conference reports.

The newsletter has a template in the Pages software that makes it simple to lay out the content. Michael O'Connell, the current Computing Newsletter Editor, is ending his term this year, and we are seeking a new chair to begin as soon as possible. Michael will work closely with the new Editor to manage the transition and will be available on an ongoing basis to help answer questions.

Serving as Newsletter Editor is a great opportunity to get to know members of the statistical computing and graphics communities. It is also a valuable professional service function to both the ASA and the Statistical Computing and Graphics Sections.

Anyone interested in learning more should contact Michael (moconnel@tibco.com), Deb Nolan (Nolan@stat.berkeley.edu) or Andreas Krause (andreas.krause@actelion.com).

Statistical COMPUTING & GRAPHICS

The Statistical Computing & Statistical Graphics Newsletter is a publication of the Statistical Computing and Statistical Graphics Sections of the ASA. All communications regarding the publication should be addressed to:

Michael O'Connell, Editor,
Statistical Computing Section
moconnel@tibco.com

Andreas Krause, Editor
Statistical Graphics Section
andreas.krause@actelion.com
http://www.elmo.ch

All communications regarding ASA membership and the Statistical Computing and Statistical Graphics Section, including change of address, should be sent to American Statistical Association, 1429 Duke Street Alexandria, VA 22314-3402 USA
(703)684-1221, fax (703)684-2036
asainfo@amstat.org