Statistical Graphics for Exploring Data, Presenting Information, and Understanding Statistical Models

Frank E Harrell Jr
Department of Biostatistics
Vanderbilt University School of Medicine
f.harrell@vanderbilt.edu

biostat.mc.vanderbilt.edu at Jump: StatGraphCourse

October 20, 2005

BASS VII

8 November 2005 Savannah GA

Chapter 1

Principles of Graph Construction

The ability to construct clear and informative graphs is related to the ability to understand the data. There are many excellent texts on statistical graphics (many of which are listed at the end of this chapter). Some of the best are Cleveland's 1994 book *The Elements of Graphing Data* and the books by Tufte. The suggestions for making good statistical graphics outlined here are heavily influenced by Cleveland's books, and quotes below are from his 1994 book.

1.1 Graphical Perception

- Goals in communicating information: reader perception of data values and of data patterns. Both accuracy and speed are important.
- Pattern perception is done by

  detection : recognition of geometry encoding physical values

  assembly : grouping of detected symbol elements

  estimation : assessment of relative magnitudes of two physical values
- For estimation, many graphics involve discrimination, ranking, and estimation of ratios
- Humans are not good at estimating differences without directly seeing differences (especially for steep curves)
- Humans do not naturally order color hues
- Only a limited number of hues can be discriminated in one graphic
- Weber's law: The probability of a human detecting a difference in two lines is related to the ratio of the two line lengths
- This is why grid lines and frames improve perception and is related to the benefits of having multiple

graphs on a common scale.

  - eye can see ratios of filled or of unfilled areas, whichever is most extreme

- For categorical displays, sorting categories by order of values attached to categories can improve accuracy of perception. Watch out for over-interpretation of extremes though.
- The aspect ratio (height/width) does not have to be unity. Using an aspect ratio such that the average absolute curve angle is $45°$ results in better perception of shapes and differences (banking to $45°$).
- Optical illusions can be caused by:
  - hues, e.g., red is emotional. A red area may be perceived as larger.
  - shading; larger regions appear to be darker
  - orientation of pie chart with respect to the horizon
- Humans are bad at perceiving relative angles (the principal perception task used in a pie chart)
- Here is a hierarchy of human graphical perception abilities:

  1. Position along a common scale (most accurate task)
  2. Position along identical nonaligned scales
  3. Length

4. Angle and slope
5. Area
6. Volume
7. Color: hue (red, green, blue, etc.), saturation (pale/deep) and lightness
   - Hue can give good discrimination but poor ordering

1.2 General Suggestions

- Exclude unneeded dimensions (e.g. width, depth of bars)
- “Make the data stand out. Avoid Superfluity”; Decrease ink to information ratio
- “There are some who argue that a graph is a success only if the important information in the data can be seen in a few seconds. . . . Many useful graphs require careful, detailed study.”
- When actual data points need to be shown and they are too numerous, consider showing a random sample of the data.
- Omit “chartjunk”

- Keep continuous variables continuous; avoid grouping them into intervals. Grouping may be necessary for some tables but not for graphs.
- Beware of subsetting the data finer than the sample size can support; conditioning on many variables simultaneously (instead of multivariable modeling) can result in very imprecise estimates

1.3 Tufte on "Chartjunk"

> Chartjunk does not achieve the goals of its propagators. The overwhelming fact of data graphics is that they stand or fall on their content, gracefully displayed. Graphics do not become attractive and interesting through the addition of ornamental hatching and false perspective to a few bars. Chartjunk can turn bores into disasters, but it can never rescue a thin data set. The best designs . . . are *intriguing and curiosity-provoking*, drawing the viewer into the wonder of the data, sometimes by narrative power, sometimes by immense detail, and sometimes by elegant presentation of simple but interesting data. But no information, no sense of discovery, no wonder, no substance is generated by chartjunk.

— Tufte p. 121, 1983

1.4 Tufte's Views on Graphical Excellence

"Excellence in statistical graphics consists of complex ideas communicated with clarity, precision, and efficiency. Graphical displays should

- show the data
- induce the viewer to think about the substance rather than about methodology, graphic design, the technology of graphic production, or something else
- avoid distorting what the data have to say
- present many numbers in a small space
- make large data sets coherent
- encourage the eye to compare different pieces of data
- reveal the data at several levels of detail, from a broad overview to the fine structure
- serve a reasonably clear purpose: description, exploration, tabulation, or decoration
- be closely integrated with the statistical and verbal descriptions of a data set."

1.5 Formatting

- Tick Marks should point outward
- x- and y-axes should intersect to the left of the lowest x value and below the lowest y value, to keep values from being hidden by axes
- Minimize the use of remote legends. Curves can be labeled at points of maximum separation

1.6 Color, Symbols, and Line Styles

- Some symbols (especially letters and solids) can be hard to discern
- Use hues if needed to add another dimension of information, but try not to exceed 3 different hues. Instead, use different saturations in each of the three different hues.
- Make notations and symbols in the plots as consistent as possible with other parts, like tables and texts
- Different dashing patterns are hard to read especially when curves inter-twine or when step functions are being displayed
- An effective coding scheme for two lines is to use a thin black line and a thick gray scale line

1.7 Scaling

- Consider the inclusion of 0 in your axis. Many times it is essential to include 0 to tell the full story. Often the inclusion of zero is unnecessary.
- Use a log scale when it is important to understand percent change of multiplicative factors or to cure skewness toward large values
- Humans have difficulty judging steep slopes; bank to $45°$, i.e., choose the aspect ratio so that average absolute angle in curves is $45°$.

1.8 Displaying Estimates Stratified by Categories

- Perception of relative lengths is most accurate — areas of pie slices are difficult to discern
- Bar charts have many problems:
  - High ink to information ratio
  - Error bars cause perception errors
  - Can only show one-sided confidence intervals well
  - Thick bars reduce the number of categories that can be shown
  - Labels on vertical bar charts are difficult to read

- Dot plots are almost always better
- Consider multi-panel side-by-side displays for comparing several contrasting or similar cases. Make sure the scales in both x and y axes are the same across different panels.
- Consider ordering categories by values represented, for more accurate perception

1.9 Displaying Distribution Characteristics

- When only summary or representative values are shown, try to show their confidence bounds or distributional properties, e.g., error bars for confidence bounds or box plot
- It is better to show confidence limits than to show ± 1 standard error
- Often it is better still to show variability of *raw* values (quartiles as in a box plot so as to not assume normality, or S.D.)
- For a quick comparison of distributions of a continuous variable against many categories, try box plots.
- When comparing two or three groups, overlaid empirical distribution function plots may be best, as these

show all aspects of the distribution of a continuous variable.

1.10 Showing Differences

- Often the only way to perceive differences accurately is to actually compute differences; then plot them
- It is not a waste of space to show stratified estimates and differences between them on the same page using multiple panels
- This also addresses the problem that confidence limits for differences cannot be easily derived from intervals for individual estimates; differences can easily be significant even when individual confidence intervals overlap.
- Humans can't judge differences between steep curves; one needs to actually compute differences and plot them.

The plot in figure 1.1 shows confidence limits for individual means, using the nonparametric bootstrap percentile method, along with bootstrap confidence intervals for the difference in the two means.

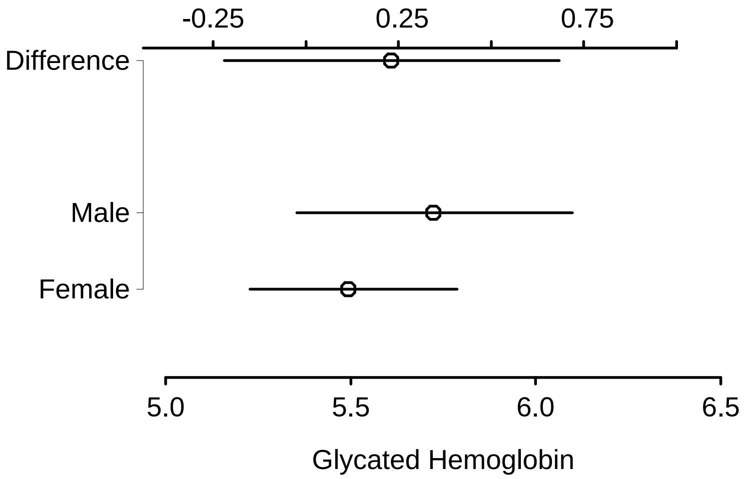


Figure 1.1: *Means and nonparametric bootstrap 0.95 confidence limits for glycated hemoglobin for males and females, and confidence limits for males - females. Lower and upper x-axis scales have same spacings but different centers. Confidence intervals for differences are generally wider than those for the individual constituent variables.*

1.11 Choosing the Best Graph Type

The recommendations that follow are good on the average, but be sure to think about alternatives for your particular data set. For nonparametric trend lines, it is advisable to add a “rug” plot to show the density of the data used to make the nonparametric regression estimate. Alternatively, use the bootstrap to derive nonparametric confidence bands for the nonparametric smoother.

1.11.1 Single Categorical Variable

Use a dot plot or horizontal bar chart to show the proportion corresponding to each category. Second choices for values are percentages and frequencies. The total sample size and number of missing values should be displayed somewhere on the page. If there are many categories and they are not naturally ordered, you may want to order them by the relative frequency to help the reader estimate values.

1.11.2 Single Continuous Numeric Variable

An empirical cumulative distribution function, optionally showing selected quantiles, conveys the most information and requires no grouping of the variable. A box plot will show selected quantiles effectively, and box plots are especially useful when stratifying by multiple categories of another variable. Histograms are also possible.

1.11.3 Categorical Response Variable vs. Categorical Ind. Var.

This is essentially a frequency table. It can also be depicted graphically

1.11.4 Categorical Response vs. a Continuous Ind. Var.

Choose one or more categories and use a nonparametric smoother to relate the independent variable to the proportion of subjects in the categories of interest. Show a rug plot on the x-axis.

1.11.5 Continuous Response Variable vs. Categorical Ind. Var.

If there are only two or three categories, superimposed empirical cumulative distribution plots with selected quantiles can be quite effective. Also consider box plots, or a dot plot with error bars, to depict the median and outer quartiles. Occasionally, a back-to-back histogram can be effective for two groups

1.11.6 Continuous Response vs. Continuous Ind. Var.

A nonparametric smoother is often ideal. You can add rug plots for the x- and y-axes, and if the sample size is not too large, plot the raw data. If you don't trust nonparametric smoothers, group the x-variable into intervals having a given number of observations, and for each x-interval plot characteristics (3 quartiles or mean $\pm$ 2 SD, for example) vs. the mean x in the interval.

1.12 Conditioning Variables

You can condition (stratify) on one or more variables by making separate pages by strata, by making separate panels within a page, and by superposing groups of points (using different symbols or colors) or curves within a panel. The actual method of stratifying on the conditional variable(s) depends on the type of variables.

Categorical variable(s) : The only choice to make in conditioning (stratifying) on categorical variables is whether to combine any low-frequency categories.

Continuous numeric variable(s) : Unfortunately, to condition on a continuous variable without the use of a parametric statistical model, one must split the variable into intervals. The first choice is whether the intervals of the numeric variable should be overlapping or non-overlapping.

Bibliography

[1] C. F. Alzola and F. E. Harrell. *An Introduction to* S *and the* *Hmisc* *and* *Design* *Libraries*. Available from `http://biostat.mc.vanderbilt.edu/s/Hmisc`.

[2] F. J. Anscombe. Graphs in statistical analysis. *American Statistician*, 27:17–21, 1973.

[3] J. Bertin. *Graphics and Graphic Information-Processing*. de Gruyter, Berlin, 1981.

[4] D. B. Carr and S. M. Nusser. Converting tables to plots: A challenge from Iowa State. *Statistical Computing and Graphics Newsletter, ASA*, December 1995.

[5] W. S. Cleveland. Graphs in scientific publications (c/r: 85v39 p238-239). *American Statistician*, 38:261–269, 1984.

[6] W. S. Cleveland. *Visualizing Data*. Hobart Press, Summit, NJ, 1993.

[7] W. S. Cleveland. *The Elements of Graphing Data*. Hobart Press, Summit, NJ, 1994.

[8] W. S. Cleveland and R. McGill. A color-caused optical illusion on a statistical graph. *American Statistician*, 37:101–105, 1983.

[9] W. S. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79:531–554, 1984.

[10] A. Gelman, C. Pasarica, and R. Dodhia. Let's practice what we preach: Turning tables into graphs. *The American Statistician*, 56:121–130, 2002.

[11] X. Li, J. Buechner, P. Tarwater, and A. Muñoz. A diamond-shaped equiponderant graphical display of the effects of two categorical predictors on continuous outcomes. *The American Statistician*, 57:193–199, 2003.

[12] F. E. Harrell. *Regression Modeling Strategies*. New York: Springer, 2001.

[13] G. T. Henry. *Graphing Data*. Sage, Newbury Park, CA, 1995.

[14] D. McNeil. On graphing paired data. *American Statistician*, 46:307–311, 1992.

[15] S. M. Powsner and E. R. Tufte. Graphical summary of patient status. *Lancet*, 344:386–389, 1994.

[16] P. R. Rosenbaum. Exploratory plots for paired data. *American Statistician*, 43:108–109, 1989.

[17] P. D. Sasieni and P. Royston. Dotplots. *Applied Statistics*, 45:219–234, 1996.

[18] P. A. Singer and A. R. Feinstein. Graphical display of categorical data. *Journal of Clinical Epidemiology*, 46:231–236, 1993.

[19] E. R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, Connecticut, 1983.

[20] E. R. Tufte. *Envisioning Information*. Graphics Press, Cheshire, Connecticut, 1990.

[21] E. R. Tufte. *Visual Explanations*. Graphics Press, Cheshire, CT, 1997.

[22] H. Wainer. How to display data badly. *American Statistician*, 38:137, 1984.

[23] H. Wainer. Three graphic memorials. *Chance*, 7:52–55, 1994.

[24] H. Wainer. Depicting error. *American Statistician*, 50:101–111, 1996.

[25] A. Wallgren, B. Wallgren, R. Persson, U. Jorner, and J. Haaland. *Graphing Statistics & Data*. Sage Publications, Thousand Oaks, 1996.

[26] C. Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann, San Francisco, 2004.

[27] L. Wilkinson. *The Grammar of Graphics*. Springer, New York, 1999.

Chapter 2

Examples

2.1 General Examples

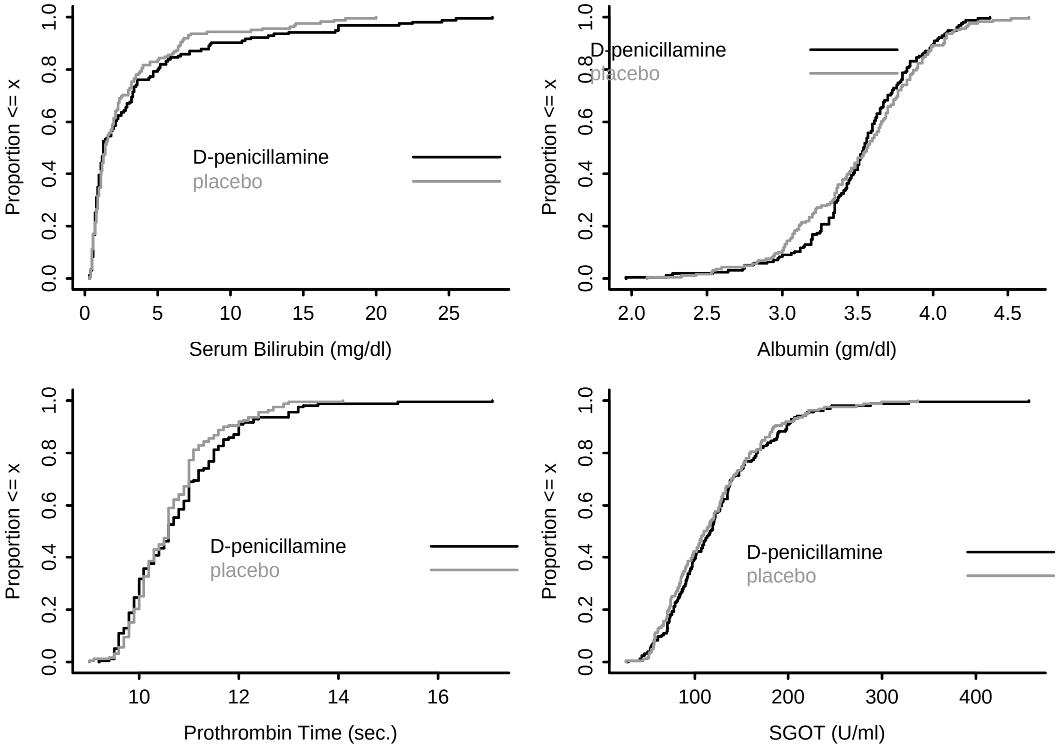


Figure 2.1: *Empirical cumulative distribution function for four continuous variables in the `pbc` dataset, stratified by randomized treatment*

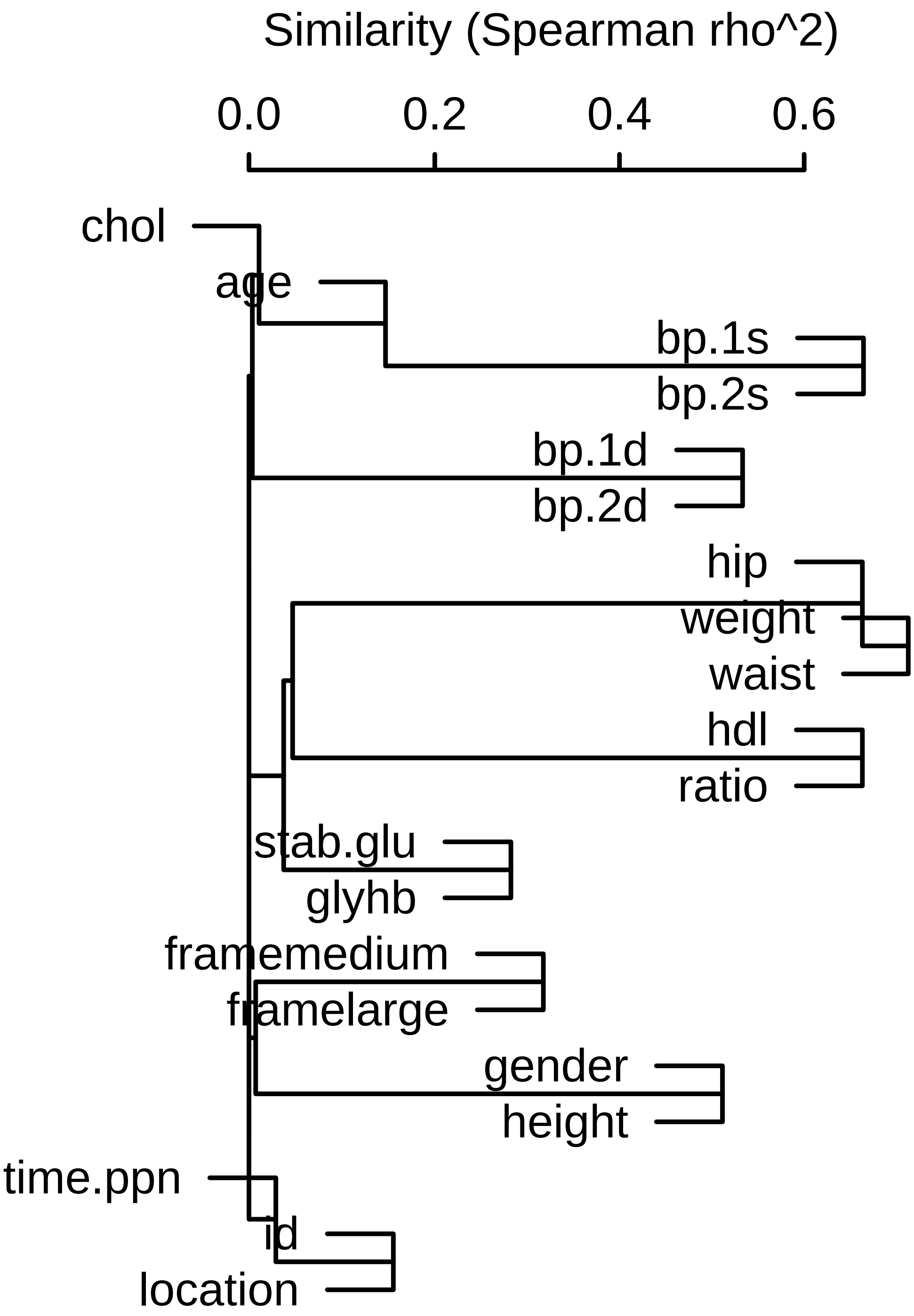


Figure 2.2: *Clusters of variables from `diabetes` using pairwise Spearman ρ^2 as the similarity measure*

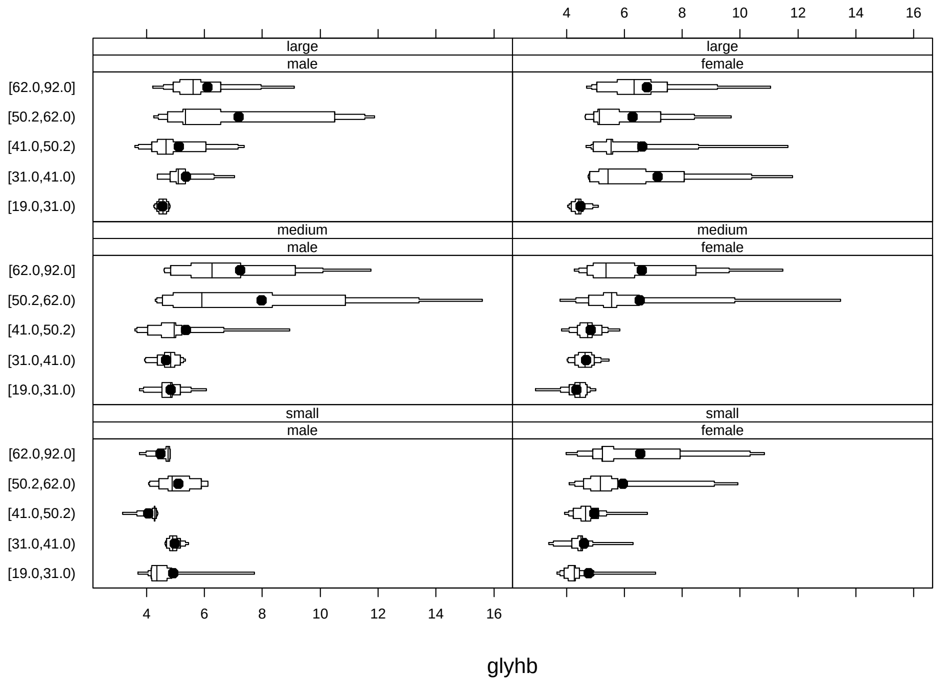


Figure 2.3: *Extended box plot displaying quantiles such that 0.25, 0.5, 0.75, and 0.9 of the data are contained within each pair of quantiles. The median is shown with a line, and the mean with a dot.*

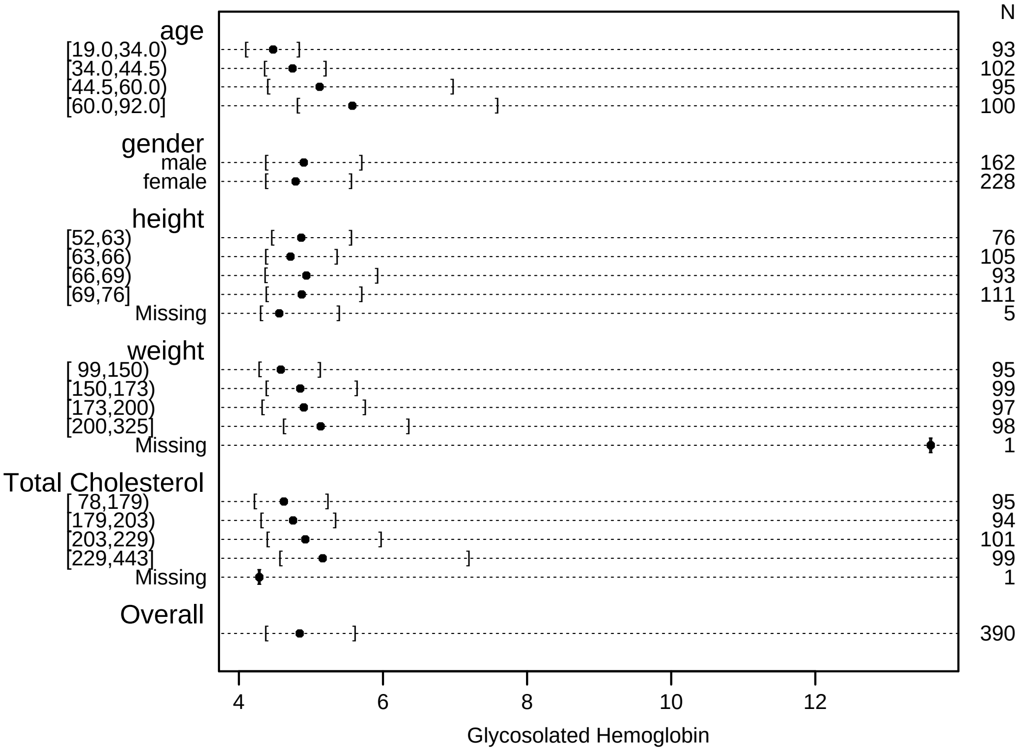


Figure 2.4: *Quartiles of `glyhb` stratified separately by several variables*

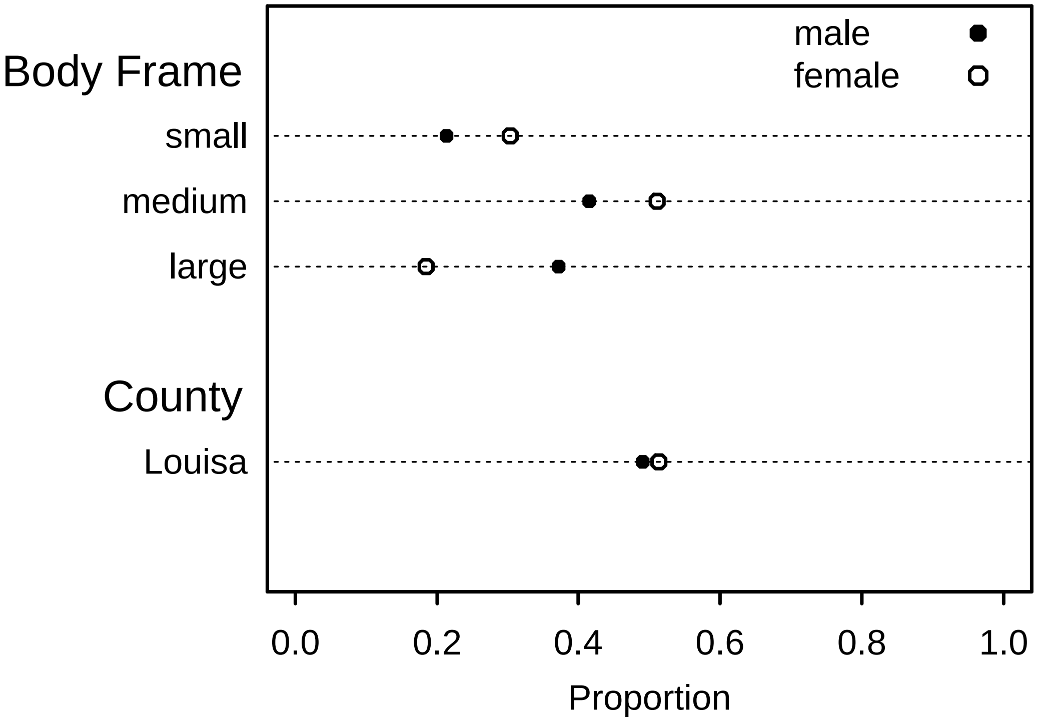


Figure 2.5: *Distribution of categorical varibles stratified by `gender`*

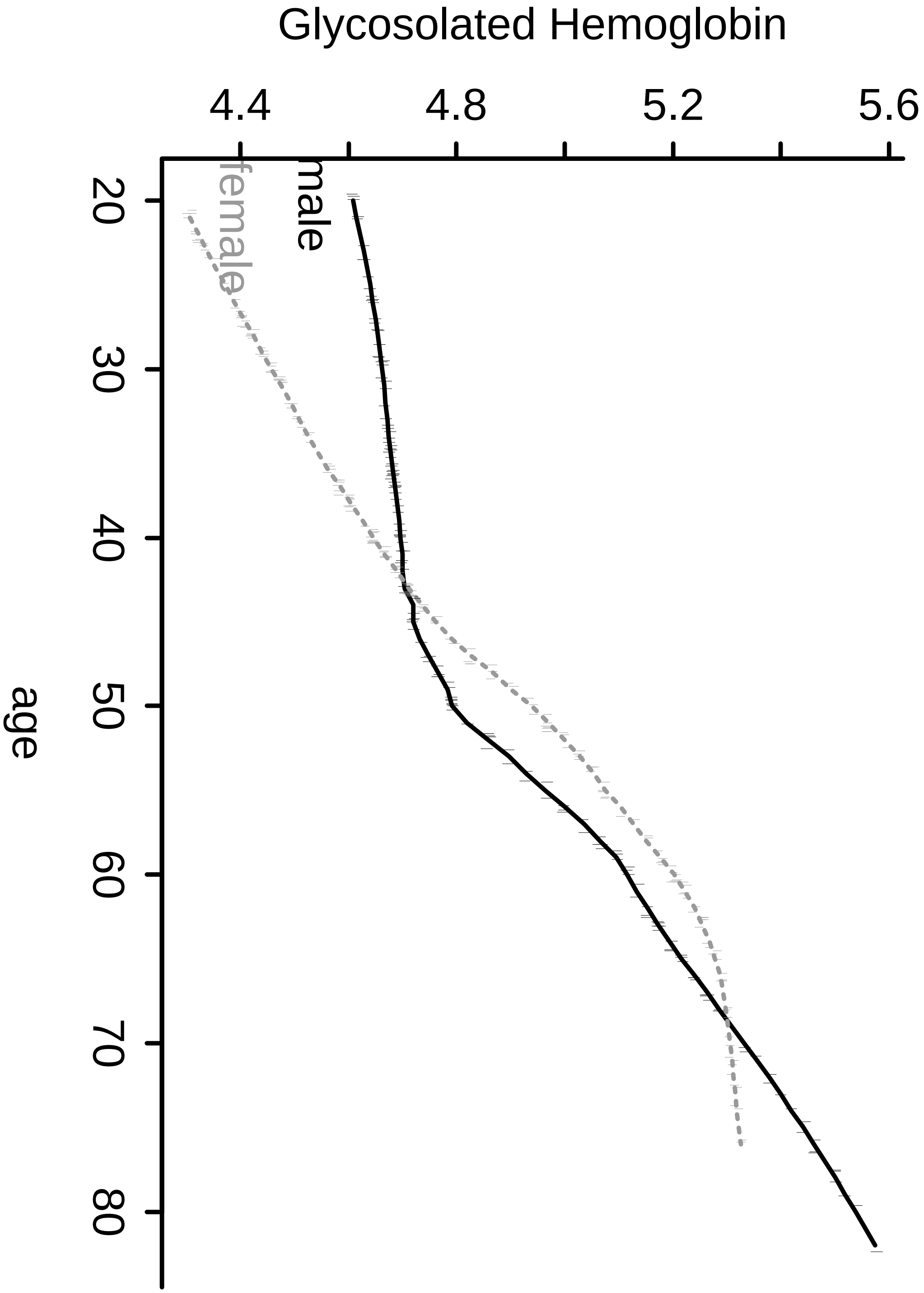


Figure 2.6: *Lowess nonparametric regression of* `glyhb` *vs.* `age`*, stratified by* `gender`

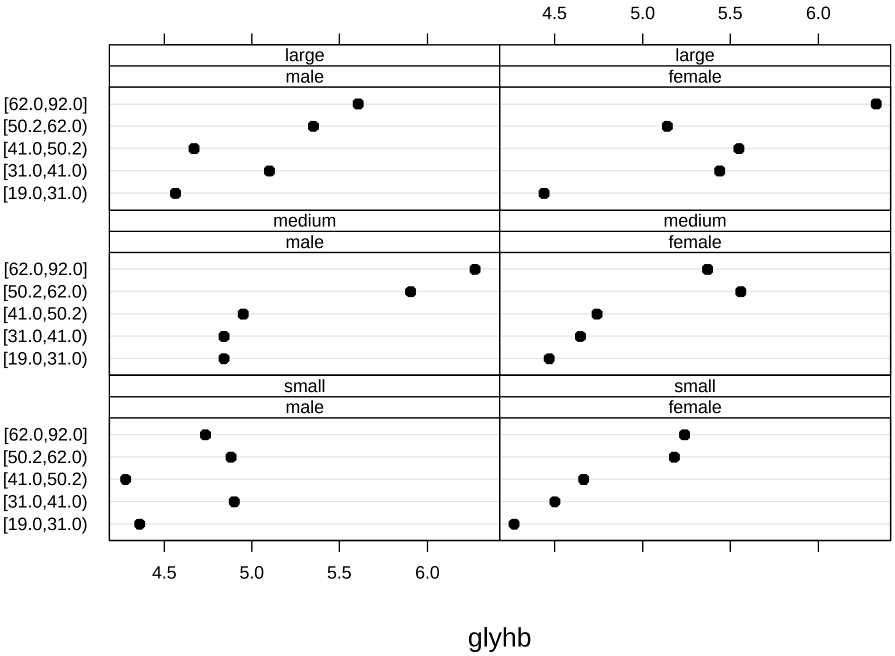


Figure 2.7: *Median `glyhb` stratified by `gender`, `frame`, and quintiles of `age`*

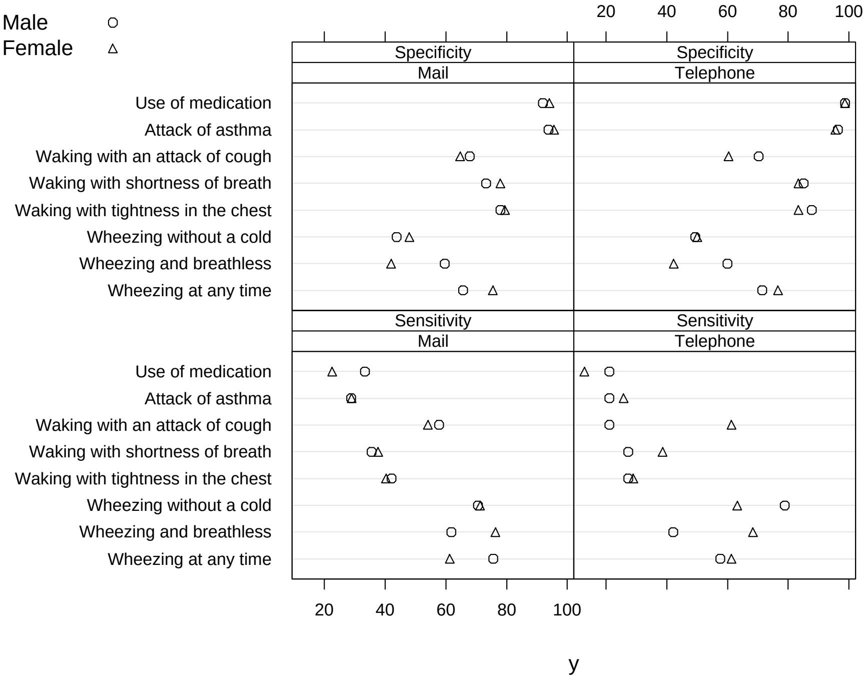


Figure 2.8: *Dot plot depicting sensitivity and specificity stratified by `method` and `sex`, with the two sexes superposed. Data are from Galobardes,* et al., J Clin Epi *51:875-881, 1998.*

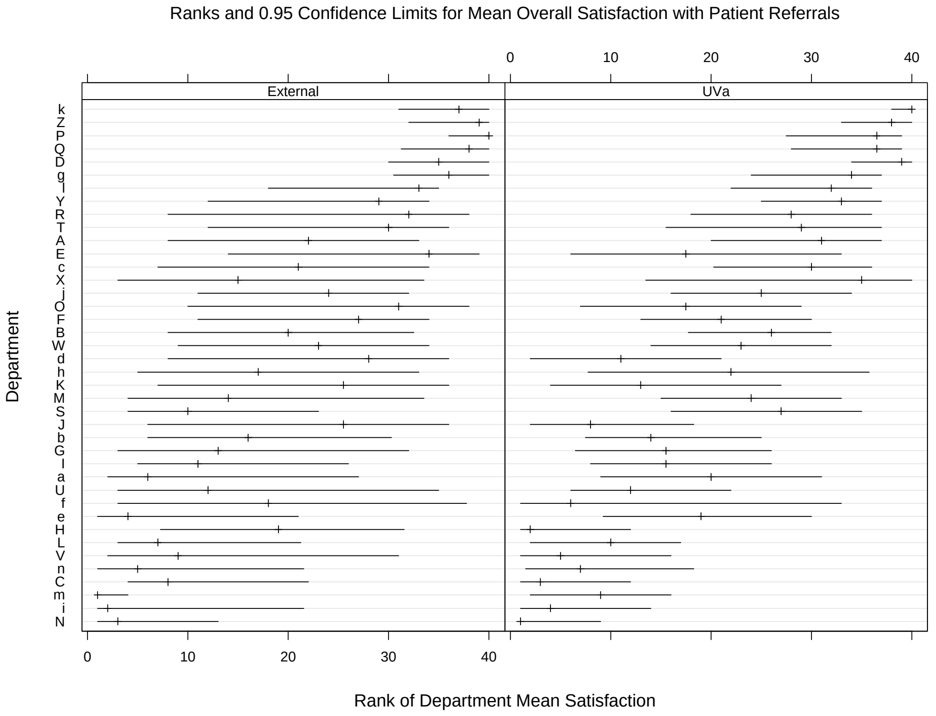


Figure 2.9: *Dot plot showing rank and bootstrap 0.95 confidence limits for the rank of mean satisfaction with service, stratified by UVa vs. outside referring physicians. Dots are sorted by descending order of the* ***mean*** *satisfaction across the two strata.*

2.2 Examples from Regression Modeling Strategies, NY: Springer 2001

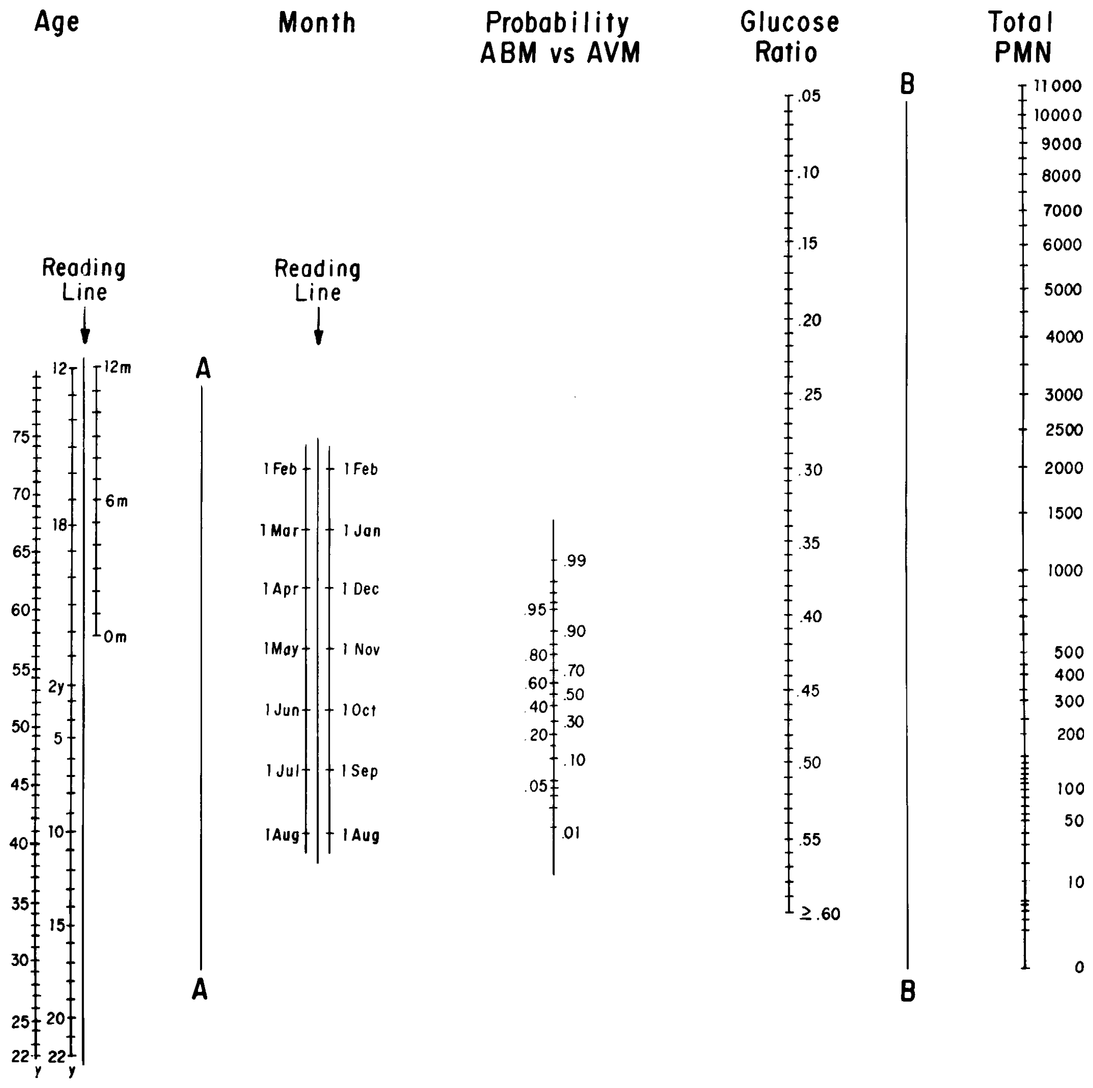


Figure 2.10: *Nomogram for estimating probability of bacterial (ABM) versus viral (AVM) meningitis. Step 1, place ruler on reading lines for patient's age and month of presentation and mark intersection with line A; step 2, place ruler on values for glucose ratio and total polymorphonuclear leukocyte (PMN) count in cerebrospinal fluid and mark intersection with line B; step 3, use ruler to join marks on lines A and B, then read off the probability of ABM versus AVM. Copyright 1989, American Medical Association. Reprinted by permission.*

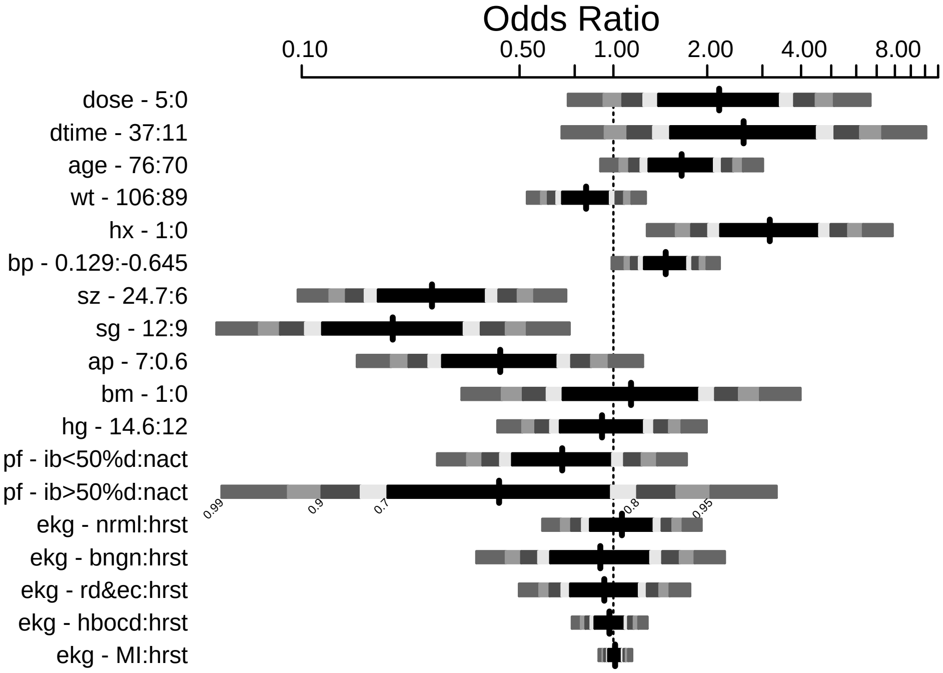


Figure 2.11: *Interquartile-range odds ratios for continuous predictors and simple odds ratios for categorical predictors. Numbers at left are upper quartile : lower quartile or current group : reference group. The shaded bars represent* $0.7, 0.8, 0.9, 0.95, 0.99$ *confidence limits. The intervals are drawn on the log odds ratio scale and labeled on the odds ratio scale. Ranges are on the original scale, even for transformed variables.*

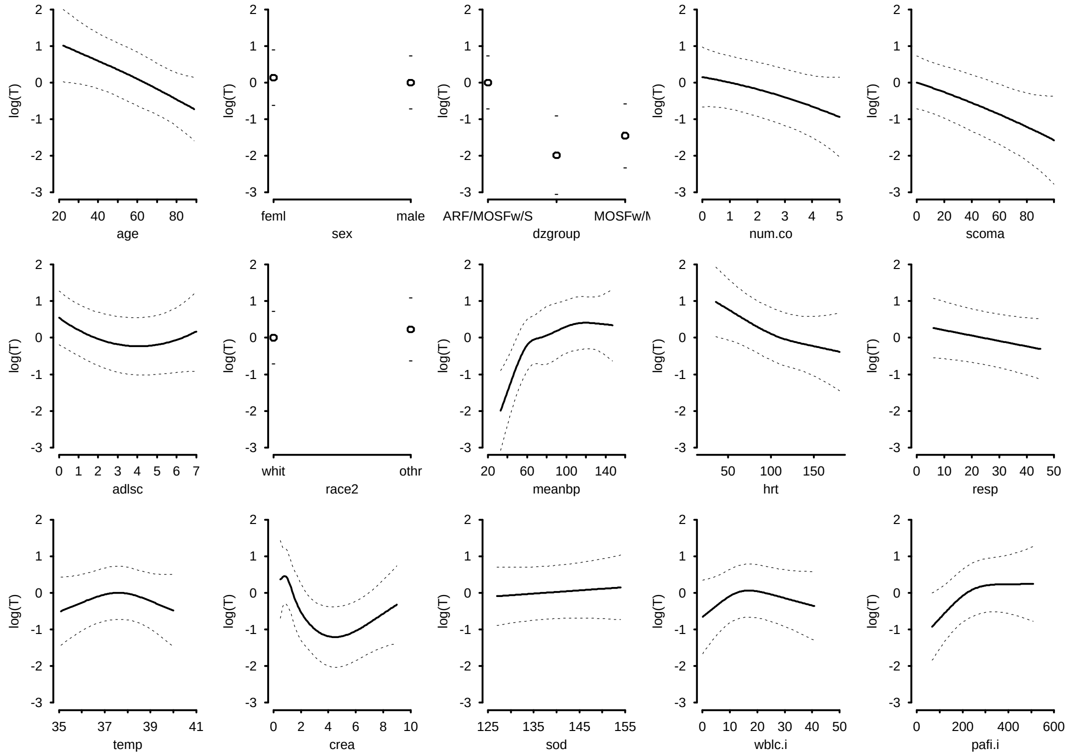


Figure 2.12: *Effect of each predictor on log survival time. Predicted values have been centered so that predictions at predictor reference values are zero. Pointwise 0.95 confidence bands are also shown. As all Y-axes have the same scale, it is easy to see which predictors are strongest.*

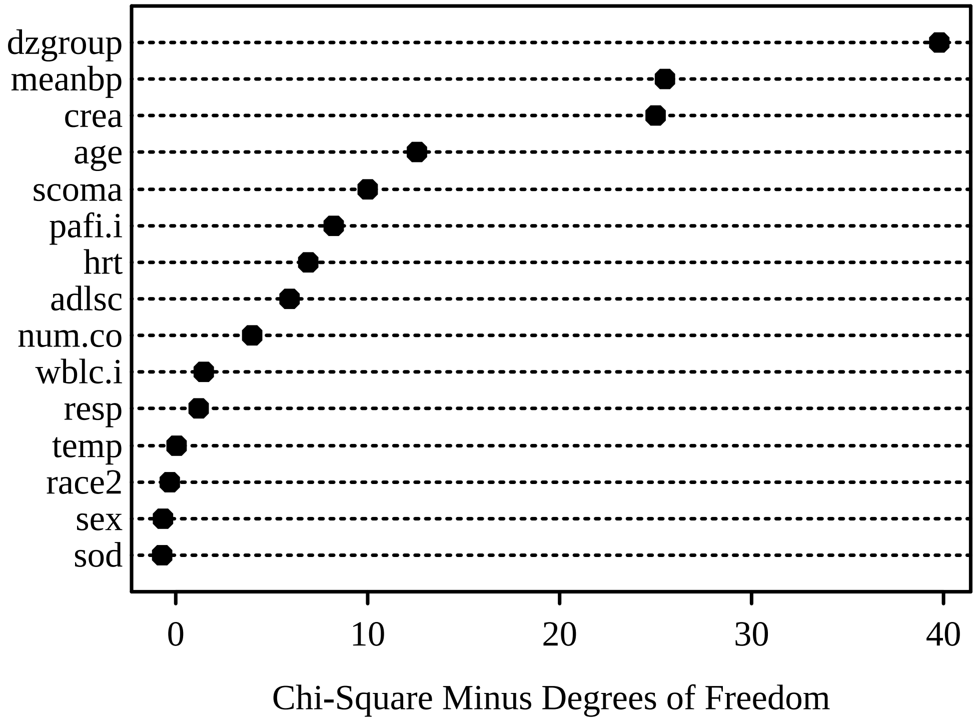


Figure 2.13: *Contribution of variables in predicting survival time in log-normal model.*

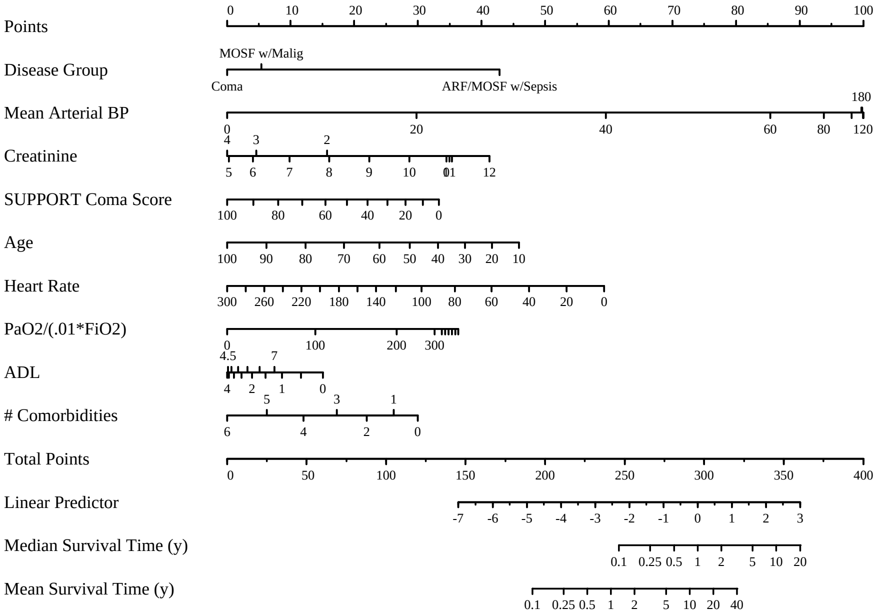
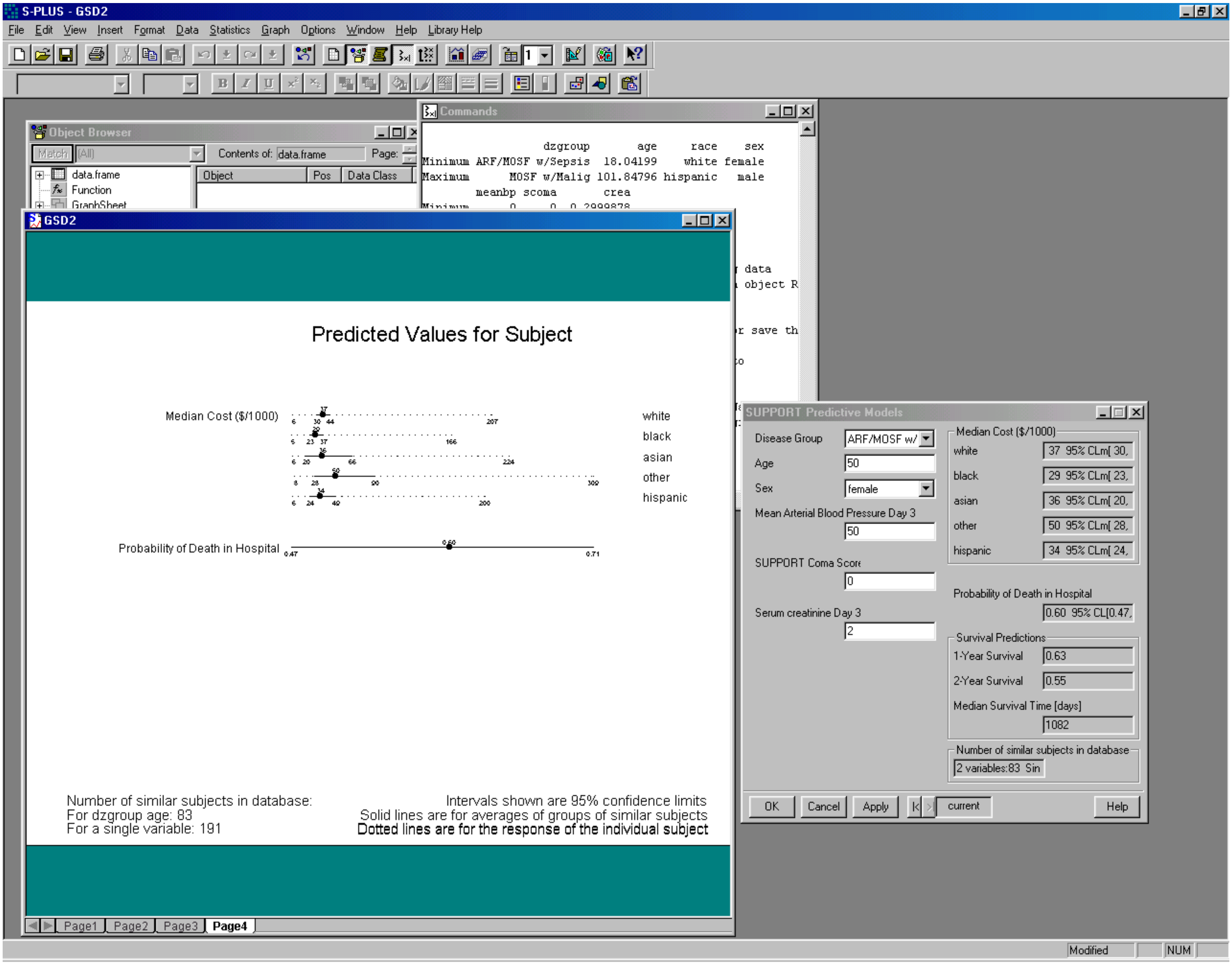


Figure 2.14: *Nomogram for predicting median and mean survival time, based on approximation of full model.*


S-PLUS - GSD2
File Edit View Insert Format Data Statistics Graph Options Window Help Library Help
Object Browser
Match (All) Contents of: data.frame Page:
data.frame
Function
GraphSheet
Object Pos Data Class
Commands
GSD2
Page1 Page2 Page3 Page4
Modified NUM

```
                 dzgroup       age     race    sex
Minimum ARF/MOSF w/Sepsis  18.04199    white female
Maximum      MOSF w/Malig 101.84796 hispanic   male
        meanbp scoma        crea
Minimum      0     0  0.2999878
```

```
data
object R
r save th
o
```

Predicted Values for Subject

Number of similar subjects in database:
For dzgroup age: 83
For a single variable: 191

Intervals shown are 95% confidence limits
Solid lines are for averages of groups of similar subjects
Dotted lines are for the response of the individual subject

SUPPORT Predictive Models

Disease Group: ARF/MOSF w/
Age: 50
Sex: female
Mean Arterial Blood Pressure Day 3: 50
SUPPORT Coma Score: 0
Serum creatinine Day 3: 2

Median Cost ($/1000)
white: 37 95% CLm[30,
black: 29 95% CLm[23,
asian: 36 95% CLm[20,
other: 50 95% CLm[28,
hispanic: 34 95% CLm[24,

Probability of Death in Hospital: 0.60 95% CL[0.47,

Survival Predictions
1-Year Survival: 0.63
2-Year Survival: 0.55
Median Survival Time [days]: 1082

Number of similar subjects in database: 2 variables:83 Sin

Figure 2.15: *Graphical user interface for entering predictors and obtaining predicted values from three models, created using the* S-PLUS *2000 Design library* Dialog *function.*

Chapter 3

Graphics for One or Two Variables

See

- www.math.montana.edu/~umsfjban/Courses/Stat438/Text/Comprehensive.toc.html

- http://exploringdata.cqu.edu.au

- http://davidmlane.com/hyperstat/desc_univ.html

- http://www.statsoft.com/textbook/stgraph.html

- http://www.itl.nist.gov/div898/handbook/eda/sectio
eda15.htm

3.1 One-Dimensional Scatterplot

- Rug plot; useful by itself or on curves or axes
- Shows all raw data values
- For large datasets, draw random thirds of vertical tick to avoid black blob
- Old-style *dot plots* are similar to rug plots
- Can use Cleveland's dot charts to show raw data

3.2 Histogram

- Used for estimating the *probability density function*

$$f(x) = \lim_{\delta \to 0} \mathrm{Prob}(x - \delta < X \leq x)/\delta \tag{3.1}$$

- Very dependent on how bins formed, and number of bins
- y-axis can be frequency or proportion

- No statistical estimates can be read directly off a histogram or density plot

3.3 Density Plot

- Smoothed histogram
- Smooth estimate of $f(x)$ above
- Depends on choice of a smoothing parameter

3.4 Empirical Cumulative Distribution Plot

- Population *cumulative distribution function* is

$$F(x) = \mathrm{Prob}(X \leq x) \tag{3.2}$$

- $F(b) - F(a) = \mathrm{Prob}(a < X \leq b)$ and is the area under the density function $f(x)$ from a to b
- Estimate of $F(x)$ is the *empirical cumulative distribution function*, which is the proportion of data values $\leq x$

- Cumulative histogram
- Works fine if histogram has one observation per bin
- ECDF requires no binning and is unique
- Excellent for showing differences in entire distributions between two or three overlaid groups
- Quantiles can be read directly off ECDF

3.5 Box Plot

- Most useful for comparing many groups
- Basically uses 3-number summary: 3 quartiles
- Easy to also show mean
- Can be extended to show other percentiles, especially farther out in the tails of the distribution

- Usually show lower and upper “adjacent” values (“whiskers” and “outside” values; some find these not to be useful

3.6 Scatter Plots

- Excellent for showing relationship between a semi-continuous X and a continuous Y
- Does not work well for huge n unless relationship is tight
- Can use transformed axes, or transformed data may be plotted
- Can show a limited number of classes of points through the use of different symbols

Chapter 4

Conditioning and Plotting Three or More Variables

4.1 Conditioning

- Choose one or two variables of principal interest
  - Typically one for histograms, ECDFs, density plots
  - Two for scatterplots
  - One or two for dot plots
- Can condition on (hold constant) effects of other variables using a statistical model (not covered in this course) or by subsetting data

- Subsets usually non-overlapping for categorical conditioning (stratification) variables

- May or may not be overlapping (shingles) intervals for continuous conditioning variables

- Conditioning may be shown in many ways
  - different symbols or colors for different groups on a scatterplot or dot plot

  - different line styles or colors on a lines plot showing multiple curves, or carefully labeled curves which use the same line styles

  - adjacent lines of dots on a dot plot

  - different vertical, horizontal (or both) panels

  - different pages, including layered transparencies

  - dynamically in real time using “brushing” and other interactive techniques

- Cleveland's principal of small multiples

See Section 1.12 of these lecture notes.

4.2 Dot Plots

- Ideal for showing how one or more categorical variables are related to a single continuous numeric response variable
- Continuous conditioning variables must be categorized
- This is usually done by creating intervals containing equal sample sizes
- Can show error bars and other superpositioning

4.3 Thermometer Plots

- Useful in problems that are similar to those handled by dot plots

- But thermometers may be positioned irregularly
- Ideal for geographical displays

4.4 Extensions of Scatterplots

4.4.1 Single Plots

- Vary symbols, colors—best for conditioning on categorical variables
- Bubble plots: can depict an addition continuous variable which may be a second *response* variable
- Radius of circles plotted is proportional to the third variable

4.4.2 Scatterplot Matrices

- Show all pairwise relationships from among 3 or more continuous variables

4.5 3-D Plots for Almost Smooth Surfaces

- Perspective plot: simulated 3-D surface
- Contour plot
- Image plot: 3rd variable categorized into, for example, 10 intervals;
  Shown using color (e.g., heat spectrum) or grayscale
  See main web page for image plot examples

4.5.1 Wireframe and Perspective Plots

- Works for smooth data or somewhat tight relationships
- Can interactively look at 3-D plot from different perspectives
- Or can automatically get a matrix of plots from varying perspectives

4.5.2 Brushing and Spinning

- Useful for examining relationships between multiple continuous variables when some of the relationships are somewhat tight (depending on the sample size)
- Brushing: highlight points in one 2-D scatterplot; shows corresponding points in other 2-D plots
- Spinning: use motion to simulation 3-D point clouds, rotating 3rd variable in and out of display

4.5.3 "Live" Graphics on Web Sites

Java Graphlets

- S-PLUS 6.x has a Java graphics device (used like postscript device but can specify underlying data)
- Allows drilling down to other pre-programmed results
- Simple to use on web sites

S-PLUS StatServer and R

- Can build web sites at which users click on options, S-PLUS is run on a server, non-pre-programmed graphics are created on the fly

- R can be freely used on web servers. Information about R may be found at `www.r-project.org`.

Chapter 5

Nonparametric Trend Lines

- Continuous X, continuous or binary Y

- Nonparametric smoother only assumes that the shape of the relationship between X and Y is smooth

- A smoother is like a moving average but better
  - Moving average is a moving flat line approximation

  - Moving averages have problems in the left and right tails

- Best all-purpose smoother: `loess`

- Is called a scatterplot smoother or moving weighted linear regression

- By having moving slope and intercept, with overlapping windows, the smooth curve is more accurate and has no problems in left and right tails

- `loess` can handle binary response variable if you turn off outlier rejection (i.e., tell the algorithm to do no extra iterations)

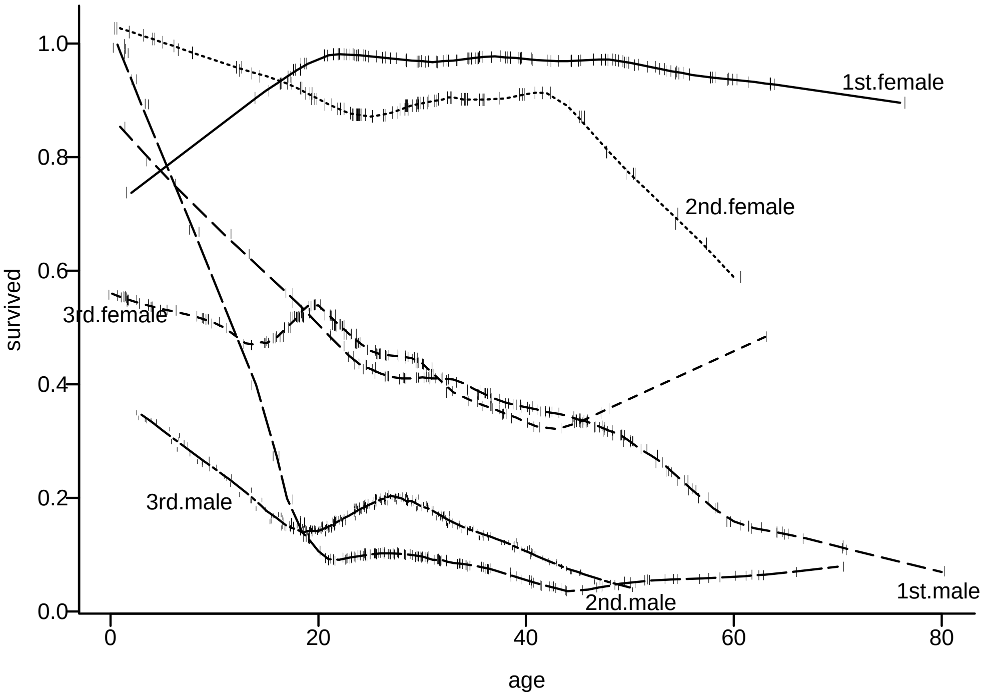


Figure 5.1: `loess` *smoothed estimates of the probability of surviving the* Titanic *as a function of passenger age, sex, and ticket class*