Semiparametric Modeling of Health Care Cost and Resource Utilization

Frank E Harrell Jr
Division of Biostatistics and Epidemiology
Department of Health Evaluation Sciences
University of Virginia School of Medicine
Box 800717 Charlottesville VA 22908 USA
fharrell@virginia.edu
hesweb1.med.virginia.edu/biostat
Slides, Data, S-Plus and R Code at /presentations

Twenty-Fourth Annual
Midwest Biopharmaceutical Statistics Workshop
Muncie, Indiana
21-23 May 2001

Outline

Slide 1

1. Choosing the model
2. Difficulties with traditional parametric models
3. Advantages of making estimation of Y–transformation an explicit part of modeling
4. Cox model
5. AVAS transform–both–sides generalized additive model
6. Smearing estimator
7. Bootstrap for CLs of effects
8. S-Plus and R functions
9. Examples: Prediction of hospital costs

Slide 2

- To satisfy distributional assumptions so that CLs, P-values will be accurate
- To minimize lack of fit, make predictions more accurate
- Could choose a model to minimize complexity (especially interactions)
- Can't compare models on the basis of what was used to optimize one of them (R^2, SSE)
- Hard to compare R^2 from models for cost and log(cost)
- Rank correlations and robust error measures can be useful
- In upcoming example, Spearman ρ for cost model is 0.66, and is 0.67 for log(cost) model
- Median absolute difference between predicted and observed costs is

$19,300 and $8,000 respectively

Slide 3

Difficulties with Parametric Models

- $\hat{\beta}$ sensitive to outliers
- Finding the best transformation of Y
- Estimating $E[Y|X]$ on original scale
- If derive CLs, P–values as if the Y–transformation is pre–specified, inference is overconfident (Faraway, 1992 [3])

Slide 4

Advantages of Estimating Y–Transform

- Make estimation of Y–transform g part of the process
- If a programmable algorithm, can use the bootstrap to account for uncertainty in g (re–estimate g at each re–sample)
- Results in honest coverage probabilities, P–values

Slide 5

Cox Model[a]

- Prob$[Y > c|X] = S(c|X) = S_0(c)^{\exp(X\beta)}$
- $S_0(\cdot)$ estimated from the data
- $\hat{\beta}$ invariant to transformations on Y, robust to outliers
- $\hat{S}(c|X) = \hat{S}_0(c)^{\exp(X\hat{\beta})}$
- Quantile q of $Y|X$: $\hat{S}^{-1}(q)$
- Estimate of mean: area under $\hat{S}$
- Can handle right–censored costs but need to take into account

Slide 6

[a] See [2], `hesweb1.med.virginia.edu/biostat/teaching/hpstat95.pdf`, `hesweb1.med.virginia.edu/biostat/presentations/dia.econ97.pdf`.

informative censoring when censoring is on time and not $ scale (no literature yet)

Slide 7

AVAS

Slide 8

- Tibshirani (1988) *additivity and variance stabilization* (AVAS) [8] [7, pp. 236-242]
  $g(Y|X = x) = f_1(x_1) + f_2(x_2) + \ldots + f_p(x_p) + \epsilon$
- Fitting criteria: maximize R^2 while forcing Y–transform (monotonic) to result in nearly constant variance of residuals.
- Transformations are nonparametric (Friedman's *super smoother* [4])
- Estimating transformations for a number of variables will inflate R^2; use Efron bootstrap optimism estimator [6] to correct

Smearing Estimator

- Duan 1983 [1]
- Estimated Y–transform $\hat{g}$
- Residuals on transformed scale: $e_1, e_2, \ldots, e_n$
- Predicted $g(Y) : a = X\hat{\beta}$
- Statistical parameter of interest: θ, e.g. $E[Y|X]$
- Function of a vector of data that estimates this parameter: W
- Smearing estimator for θ: $\hat{\theta} = W(\hat{g}^{-1}(a+e_1), \hat{g}^{-1}(a+e_2), \ldots, \hat{g}^{-1}(a+e_n))$
- For AVAS nonparametric transformation $\hat{g}$ use inverse linear interpolation to obtain $\hat{g}^{-1}(\cdot)$

Slide 9

Bootstrap for CLs of Estimated Effects

- Nonparametric bootstrap to get pointwise CLs for transformations of each variable
- Effect of changing one predictor, holding others constant:
  Use ordinary bootstrap to estimate SD of difference in two smearing estimates (for two values of X), assuming normality of such differences

Slide 10

In Hmisc library [5]. Basic `avas` function by Tibshirani is built–in to S-PLUS, is in R `mva` package.

```
f ← areg.boot(Y ~ monotone(age) +
               sex + weight)
```

Slide 11

```
plot(f)        # show transformations, CLs
Function(f)    # generate S-PLUS/R functions
               # defining transformations
predict(f)     # get predictions,
               # smearing etimates
summary(f)     # compute CLs on effects of
               # each X
smearingEst() # generalized smearing
               # estimators
```

```
Mean(f)        # derive S-PLUS/R function to
               # compute smearing mean Y
Quantile(f)    # derive S-PLUS/R function to
               # compute smearing quantile
```

Slide 12

Example

Slide 13

- Prediction of hospital costs for 894 patients in SUPPORT (Study to Understand Prognoses Preferences Outcomes and Risks of Treatments) (`hesweb1.med.virginia.edu/biostat/s/data`)
- Predictors: age, SUPPORT coma score, disease group (8 levels), mean arterial blood pressure

```
f.areg ← areg.boot(totcst ~
  dzgroup + scoma + meanbp + age)
```

- Apparent R^2 = 0.43; bootstrap overfitting–corrected R^2 = 0.41

AVAS: Estimated Transformations

Slide 14

Slide 15

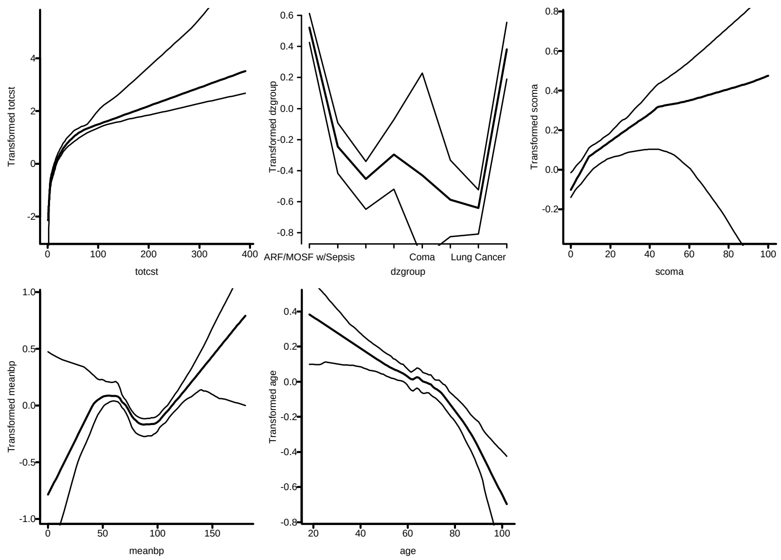

Figure 1: *Estimated transformations from AVAS, with pointwise 0.95 CLs computed by* `areg.boot`

Estimates of Effects of Predictors

```
summary(f.areg,
  values=list(scoma=c(0,44),
              meanbp=c(90,20,60,130)))
```

Slide 16

```
Values to which predictors are set when estimating
effects of other predictors:

 totcst dzgroup scoma meanbp  age
   15.1     4.5    22     75 64.9

Estimates of differences of effects on Median Y (from first X value),
and bootstrap standard errors of these differences.
Settings for X are shown as row headings.

Predictor: dzgroup
                  Differences  S.E Lower 0.95 Upper 0.95     Z   Pr(|Z|)
ARF/MOSF w/Sepsis        0.00   NA         NA         NA    NA        NA
             COPD      -20.17 4.15      -28.3     -12.04 -4.87 1.14e-006
              CHF      -23.23 4.18      -31.4     -15.04 -5.56 2.73e-008
        Cirrhosis      -21.00 4.49      -29.8     -12.21 -4.68 2.87e-006
             Coma      -22.92 5.54      -33.8     -12.06 -4.14 3.54e-005
     Colon Cancer      -24.73 4.63      -33.8     -15.65 -5.34 9.29e-008
      Lung Cancer      -25.23 4.34      -33.7     -16.72 -5.81 6.33e-009
     MOSF w/Malig       -4.75 3.48      -11.6       2.07 -1.37 1.72e-001
```

```
Predictor: scoma
   Differences S.E Lower 0.95 Upper 0.95    Z Pr(|Z|)
 0        0.00  NA         NA         NA   NA      NA
44        5.27 1.8       1.75       8.79 2.93 0.00334

Predictor: meanbp
    Differences  S.E Lower 0.95 Upper 0.95      Z Pr(|Z|)
 90        0.00   NA         NA         NA     NA      NA
 20       -2.78 6.56    -15.647       10.1 -0.424 0.67187
 60        5.21 2.74     -0.158       10.6  1.902 0.05713
130        8.29 2.96      2.477       14.1  2.796 0.00518
```

Slide 17

```
Predictor: age
      Differences   S.E Lower 0.95 Upper 0.95     Z Pr(|Z|)
52.10       0.000    NA         NA         NA    NA      NA
64.90      -0.898 0.671      -2.21      0.417 -1.34  0.1806
74.66      -2.034 0.938      -3.87     -0.195 -2.17  0.0302
```

AVAS Residuals

Slide 18

Slide 19

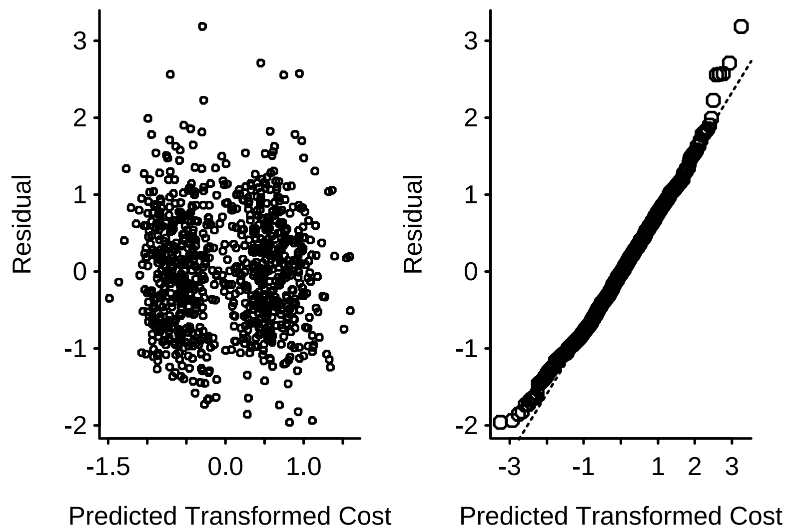

Figure 2: *Left panel: residuals from AVAS fit against predicted transformed Y. Right panel: q–q plot of residuals against the normal distribution.*

Slide 20

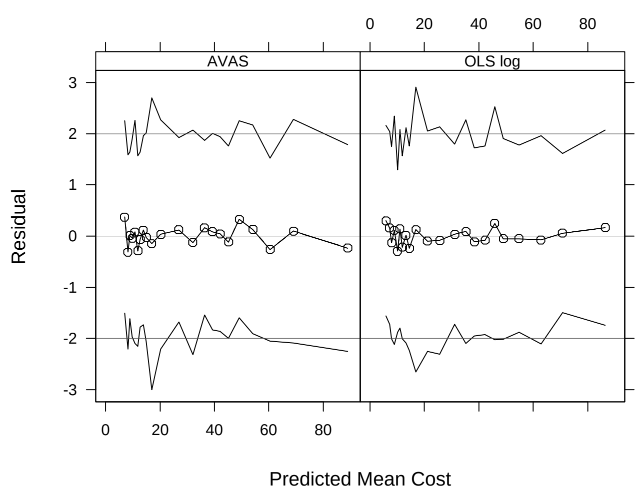

Figure 3: *Means and ± 2 standard deviations of residuals from AVAS (left panel) and OLS on log cost (right panel), after stratifying predicted mean costs into intervals containing an average of 80 patients. Residuals were first scaled to have overall standard deviations of 1.0 for both models. Both models appear to be equally variance stabilizing. There is a slight lack of fit of the log OLS model for very small predicted mean costs.*

Slide 21

Slide 22

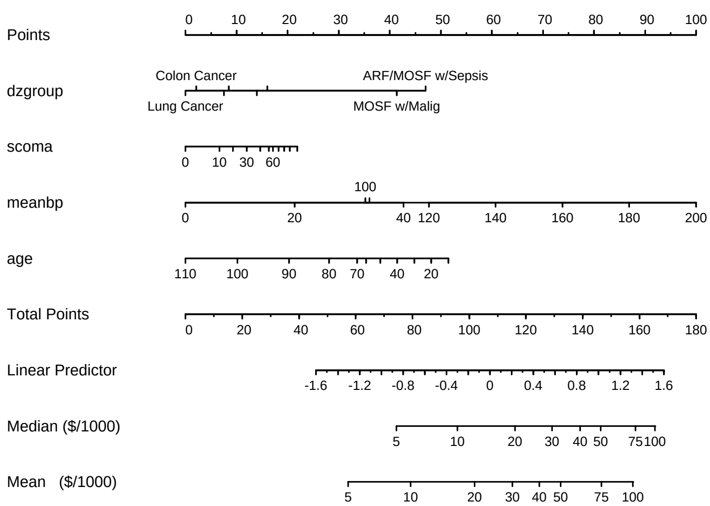

Figure 4: *Nomogram for predicting median and mean hospital cost for an individual patient.*

Slide 23

Slide 24

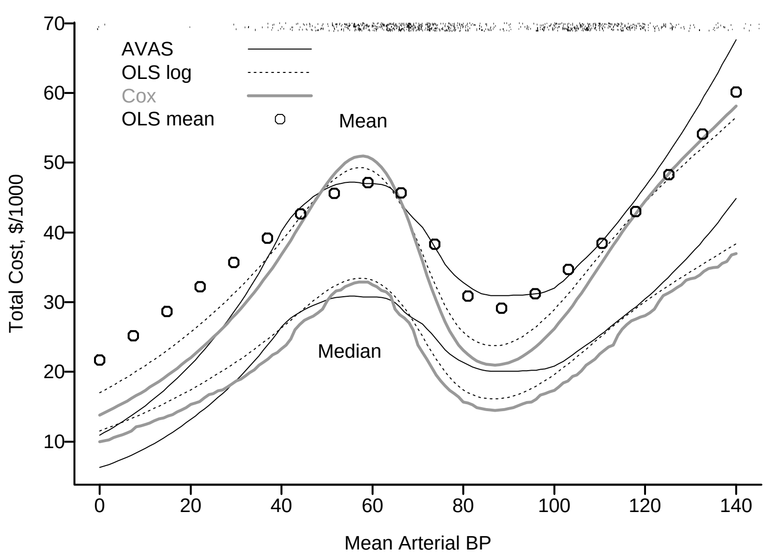

Figure 5: *Comparison of methods in predicting mean and median cost as a function of mean arterial blood pressure, for AVAS, ordinary least squares (OLS) on log cost, and the Cox model on the original cost scale. The rug plot at the top of the graph shows the data distribution for mean blood pressure. For AVAS, smearing estimators are used. For OLS based on log cost, the log–normal distribution is used so that the MLE of the estimated mean cost is* $\exp(\hat{\mu} + \frac{1}{2}\hat{\sigma}^2)$. *In addition,* $\hat{Y}$ *is presented on the original scale for an OLS model using that scale (shown as dots; note the scant data for b.p.* < 35*). This is a direct but non–robust estimate of the mean assuming no interactions.*

Slide 25

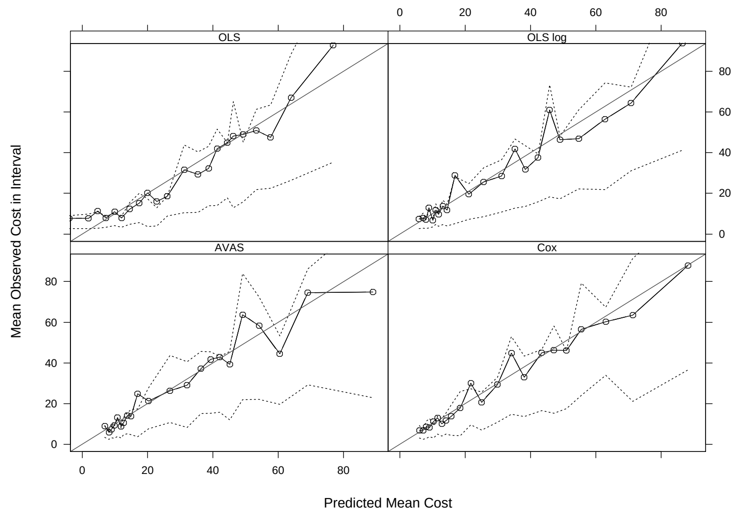

Figure 6: *Plots of means of predicted mean costs vs. mean observed costs, by intervals of predicted costs containing an average of 40 subjects. The line of identity is shown. Dotted lines depict outer quartiles of observed costs within the intervals. Note the systematic error for low predicted mean costs for OLS done on the original cost scale. The other three methods appear equally good. The minimum Spearman ρ between any two predicted means was 0.97.*

- Spearman ρ for predicted mean cost vs. actual cost in individual patients is 0.66, 0.69, 0.67, 0.68 for OLS, log OLS, AVAS, and Cox, respectively.

Slide 26

Abstract

Slide 27

Cost and other health resource utilization measures such length of hospital stay have strongly skewed distributions that make robust estimation of patient and provider effects difficult. The robust semi-parametric Cox proportional hazards model has been shown to have advantages for modeling hospital cost (Dudley et al. 1993 [2]). Ordinary least squares, a commonly used older approach, can perform well if the response variable has been suitably transformed and if appropriate non-linear and non-additive effects are allowed for the predictors. However, if one has to do exploratory analyses to determine the response transformation, variances of parameter estimates are no longer appropriate (Faraway 1992 [3]). This argues for making the determination of the transformation of the response to be an explicit part of the modeling process so that the bootstrap can be used to estimate variances correctly. Tibshirani's AVAS method [8] is a kind of generalized additive model in which the predictors are nonparametrically transformed to optimize R^2 and the response is nonparametrically transformed to stabilize variances of residuals. This talk will show how the AVAS approach can be extended to allow estimation of mean (using Duan's smearing estimator [1] and median cost given predictors, and how the bootstrap can be used to obtain confidence intervals for effects, taking all modeling steps into account. A fitted AVAS model will be compared to a Cox model for health care costs.

Slide 28

References

[1] N. Duan. Smearing estimate: A nonparametric retransformation method. *Journal of the American Statistical Association*, 78:605–610, 1983.

[2] R. Dudley, F. E. Harrell, L. Smith, D. B. Mark, R. M. Califf, D. B. Pryor, D. Glower, J. Lipscomb, and M. Hlatky. Comparison of analytic models for estimating the effect of clinical factors on the cost of coronary artery bypass graft surgery. *Journal of Clinical Epidemiology*, 46:261–271, 1993.

[3] J. J. Faraway. The cost of data analysis. *Journal of Computational and Graphical Statistics*, 1:213–229, 1992.

[4] J. H. Friedman. A variable span smoother. Technical Report 5, Laboratory for Computational Statistics, Department of Statistics, Stanford University, 1984.

[5] F. E. Harrell. `Hmisc`: A library of miscellaneous S-PLUS functions. Available from `lib.stat.cmu.edu` or `hesweb1.med.virginia.edu/s/Hmisc.html`, 2000.

Slide 29

[6] F. E. Harrell, K. L. Lee, and D. B. Mark. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15:361–387, 1996.

[7] MathSoft. *S-PLUS 2000 Guide to Statistics*, volume 1. MathSoft Data Analysis Products Division, Seattle, WA, 1999.

[8] R. Tibshirani. Estimating transformations for regression via additivity and variance stabilization. *Journal of the American Statistical Association*, 83:394–405, 1988.

Slide 30