Nonparametric Heteroscedastic Transformation Regression Models for Skewed Data, with an Application to Health Care Costs

Xiao-Hua Zhou, Huazhen Lin, Eric Johnson *Journal of Royal Statistical Society Series B* (2008)

Scott Coggeshall

April 25, 2013

Motivation: Health Care Cost Data

- Key component of risk assessment models used in insurance, health care industries
- Requires prediction of a patient's health care cost on the original scale
- Let Y be a patient's health care cost and $\mathbf{X}$ be a vector of patient characteristics and previous health states.
- Goal: Given a patient's covariate vector $\mathbf{x}$, can we accurately predict $\mu(\mathbf{x}) = E[Y|\mathbf{X} = \mathbf{x}]$?

Health Care Cost Data

Problems with health care cost data

- Skewed Distribution
- Heteroscedasticity
- "Spike" at Zero

Previous Approaches

Generalized Linear Models

- Prior research suggests estimates can be imprecise in this setting

Transformation Models

- Retransformation bias is a problem
- Usually require specification of transformation function

Proposed Model

$$H(Y) = \mathbf{X}'\boldsymbol{\beta} + \sigma\left(\mathbf{X}'\gamma\right)\epsilon$$

- $H(\cdot)$ is unknown, increasing function with $H(y_0) = 0$ for some finite y_0
- $\sigma(\cdot)$ is known variance function
- $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are vectors of unknown parameters
- ϵ is error term with $E[\epsilon] = 0$, $Var[\epsilon] = 1$, and unknown distribution function F

- How do we go from this model to a prediction $\hat{\mu}(\mathbf{x})$?

Smearing Estimator (Duan 1983)

- Let $Y_1, \ldots, Y_n$ denote the untransformed response and $\nu_i = H(Y_i)$ denote the transformed response for some known function H
- Fit linear model on transformed scale:

$$\nu_i = x_i\beta + \epsilon_i$$

where $\epsilon_i \sim F$ are i.i.d. error terms with mean 0 and variance σ^2.

- To avoid retransformation bias, estimate $E[Y_0|X = x_0]$ with the smearing estimator based on the empirical CDF:

$$\begin{aligned}\hat{E}[Y_0|X = x_0] &= \int H^{-1}(x_0\hat{\beta} + \epsilon)\mathrm{d}\hat{F}(\epsilon) \\ &= \frac{1}{n}\sum_{i=1}^{n} H^{-1}(x_0\hat{\beta} + \hat{\epsilon}_i)\end{aligned}$$

Smearing Estimator cont.

- Issues with Duan's smearing estimator
  - Transformation H must be specified
  - Assumes homoscedasticity
- Project paper extends smearing estimator to case with unknown transformation and heteroscedasticity

Deriving Our Estimator: Some Notation

Recall the proposed model:

$$H(Y) = \mathbf{X}'\boldsymbol{\beta} + \sigma\left(\mathbf{X}'\gamma\right)\epsilon$$

- Let $(Y_i, \mathbf{X}_i), i = 1, \ldots, n$ be a random sample that satisfies this model
- Let $Z_1 = \mathbf{X}'\boldsymbol{\beta}$, $Z_2 = \mathbf{X}'\gamma$, $Z_{1i} = \mathbf{X_i'}\boldsymbol{\beta}$, and $Z_{2i} = \mathbf{X_i'}\boldsymbol{\gamma}$
- Let $G(\cdot|z_1, z_2)$ be the CDF of $Y|Z_1 = z_1, Z_2 = z_2$ and $p(\cdot, \cdot)$ be the PDF of (Z_1, Z_2)
- Assume H, F, and G are differentiable with
  - $h(y) = dH(y)/dy$
  - $f(y) = dF(y)/dy$
  - $p(y|z_1, z_2) = dG(y|z_1, z_2)/dy$
  - $g_j(y|z_1, z_2) = dG(y|z_1, z_2)/dz_j, j = 1, 2$

Deriving Our Estimator

Some calculations (details omitted...) give us an expression for H(y):

$$H(y) = -\int_{y_0}^{y} \frac{\sum_{i=1}^{n} p(u|Z_{1i}, Z_{2i})p(Z_{1i}, Z_{2i})}{\sum_{i=1}^{n} g_1(u|Z_{1i}, Z_{2i})p(Z_{1i}, Z_{2i})} du$$

- To estimate H we need estimates of $p(z_1, z_2)$, $G(y|z_1, z_2)$, and derivatives of $G(y|z_1, z_2)$
- Estimate $G(y|z_1, z_2)$ with kernel estimator:

$$\begin{aligned} G_n(y|z_1, z_2) &= \frac{1}{nh_1 h_2 p_n(z_1, z_2)} \sum_{i=1}^{n} I(Y_i \leq y) K_1(\frac{Z_{1i} - z_1}{h1}) \\ &\quad \times K_2(\frac{Z_{2i} - z_2}{h_2}) \end{aligned}$$

- Estimate $p(z_1, z_2)$ with kernel density estimator:

$$p_n(z_1, z_2) = \frac{1}{nh_1 h_2} \sum_{i=1}^{n} K_1(\frac{Z_{1i} - z_1}{h_1}) K_2(\frac{Z_{2i} - z_2}{h_2})$$

Deriving Our Estimator, cont.

- Estimate $p(y|z_1, z_2)$ with kernel density estimator:

$$p_n(y|z_1, z_2) = \frac{1}{nh_1 h_2 h_0 p_n(z_1, z_2)} \sum_{i=1}^{n} K_0(\frac{Y_i - y}{h_0}) K_1(\frac{Z_{1i} - z_1}{h_1}) \times K_2(\frac{Z_{2i} - z_2}{h_2})$$

- Given H, we estimate $\boldsymbol{\beta}$ and γ simultaneously with the estimating equations:

$$\sum_{i=1}^{n} \frac{(H(Y_i) - \mathbf{X}_i'\beta)\mathbf{X}_i}{\sigma^2(\mathbf{X}_i'\gamma)} = 0$$

and

$$\sum_{i=1}^{n} \{(H(Y_i) - \mathbf{X}_i'\beta)^2 - \sigma^2(\mathbf{X}_i'\gamma)\}\mathbf{X}_i = 0$$

- How do we arrive at final estimates for $H, \boldsymbol{\beta}$, and γ?

A Familiar Algorithm...

1. Select initial values of H and $\boldsymbol{\beta}$
2. Estimate γ
3. Re-estimate H given current $\boldsymbol{\beta}$ and γ
4. Re-estimate $\boldsymbol{\beta}$ and γ given current H
5. Repeat Steps 3 and 4 until convergence

The Final Product

Given final estimates of $H, \boldsymbol{\beta}$, and γ, our prediction is given by:

$$\hat{\mu}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \hat{H}^{-1}\left(\mathbf{x}'\hat{\boldsymbol{\beta}} + \sigma(\mathbf{x}'\hat{\gamma})\frac{\hat{H}(Y_i) - \mathbf{X}_i'\hat{\boldsymbol{\beta}}}{\sigma(\mathbf{X}_i'\hat{\gamma})}\right)$$

What's Coming Next

- A closer look at implementation
- We've avoided assumptions to gain robustness. Have we sacrificed efficiency?
  - What happens as $n \to \infty$?
- The variance function $\sigma(\cdot)$ had to be specified beforehand
  - Simulations can help assess what happens when it is misspecified