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

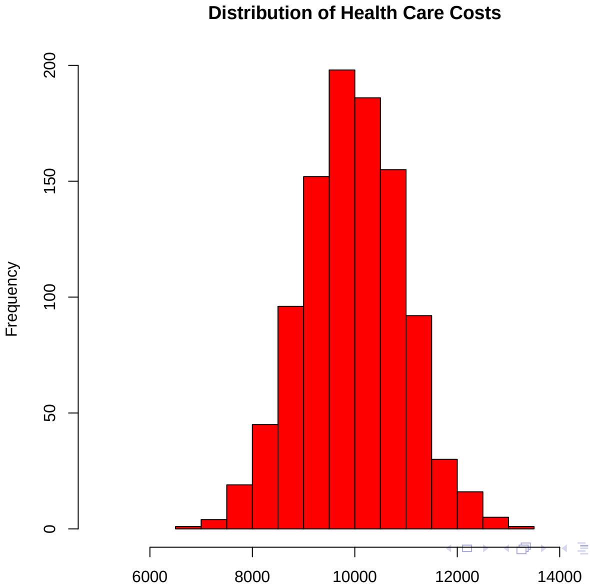
June 6, 2013

Background and 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 $\mu(\mathbf{x}) = E[Y|\mathbf{X} = \mathbf{x}]$?

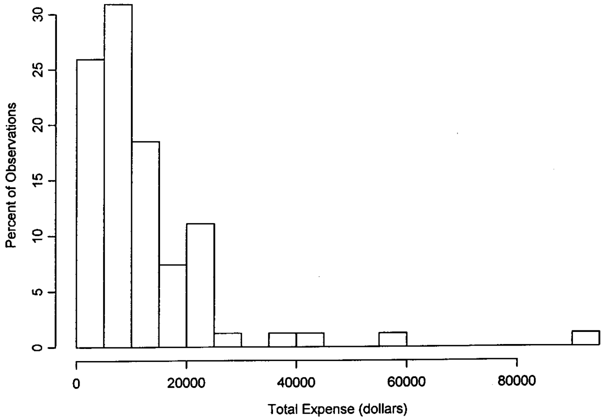
Health Care Cost Data

- What we'd **like** to have...



Health Care Cost Data

- And what we **actually** have...



Health Care Cost Data

- Skewed Distribution
- Heteroscedasticity
- Estimates of $\mu(\mathbf{x})$ can vary widely depending on how estimators handle these aspects of the data

Transformation: A Common Approach

- Suppose we observe a patient's health care cost Y and a vector of patient characteristics $\mathbf{X}$
- A common approach is to fit a linear model to a transformation of the data

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

- Is $H(Y)$ actually of interest?
- Does the model tell us anything about the data on the original scale?

Transformation Bias

- Suppose we fit a linear model on the transformed scale
- Bias is often introduced when retransforming
- In general,

$$E[H^{-1}(H(Y))|X] \neq H^{-1}E[H(Y)|X]$$

- How do we get unbiased estimate on original scale?

Duan's Smearing Estimator

- Assume H is known and data are homoscedastic
- Fit linear model on transformed scale to obtain parameter estimate $\hat{\boldsymbol{\beta}}$ and residuals $\hat{\boldsymbol{\epsilon}}$
- Unbiased estimate on original scale is guaranteed by taking expectation with respect to residuals:

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

- Suppose we want
  - Robustness to model misspecification?
  - Ability to handle heteroscedasticity?

Extending Duan's Smearing Estimator

- Proposed Model

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

- Knowns
  - $\sigma(\cdot)$
  - $E[\epsilon] = 0$, $Var[\epsilon] = 1$
- Unknowns
  - $H(\cdot)$
  - $\boldsymbol{\beta}, \boldsymbol{\gamma}$
  - CDF F of ϵ
- Approach:
  - Estimate H via kernel estimation
  - Estimate $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ via estimating equations

Estimating β and γ

- Authors propose set of 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

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

- Benefits
  - Closed-form solution for β
- Drawbacks
  - No closed-form solution for γ
  - Newton-Rhapson implementation will vary depending on form of σ
  - Mean-variance relationship?

Estimating H

- Note that Y depends on $\mathbf{X}$ through indices $Z_1 = \mathbf{X}'\beta$ and $Z_2 = \mathbf{X}'\gamma$
- Under the model, we have the following relationship between conditional CDF of Y, $G(y|z1, z2)$, and unknown CDF of error term F:

$$G(y|z1, z2) \quad = \quad F\left(\frac{H(y) - z_1}{\sigma(z_2)}\right)$$

- Taking derivatives with respect to y and z_1 yields

$$\begin{aligned} p(y|z_1, z_2) &= f\left(\frac{H(y) - z_1}{\sigma(z_2)}\right)\frac{H'(y)}{\sigma(z_2)} \text{ and} \\ g_1(y|z_1, z_2) &= -f\left(\frac{H(y) - z_1}{\sigma(z_2)}\right)\frac{1}{\sigma(z_2)} \end{aligned}$$

Estimating H continued

- These derivatives give us the relationship between $p(y|z_1, z_2)$ and $g_1(y|z_1, z_2)$:

$$p(y|z_1, z_2) \quad = \quad -g_1(y|z_1, z_2)H^{'}(y)$$

- By replacing z_1, z_2 with Z_{1i}, Z_{2i} and summing over all observations, we obtain

$$H^{'}(y) \quad = \quad -\frac{\sum p(y|Z_{1i}, Z_{2i})p(Z_{1i}, Z_{2i})}{\sum g_1(y|Z_{1i}, Z_{2i})p(Z_{1i}, Z_{2i})}$$

- Integrating both sides yields an expression for H:

$$H(y) \quad = \quad -\int_{y_0}^{y} \frac{p(y|Z_{1i}, Z_{2i})p(Z_{1i}, Z_{2i})}{\sum g_1(y|Z_{1i}, Z_{2i})p(Z_{1i}, Z_{2i})}$$

- Estimator for H is given by replacing unknown functions $p(y|z_1, z_2), g_1(y|z_1, z_2), p(z_1, z_2)$ with estimates obtained through kernel estimation

Estimating H continued

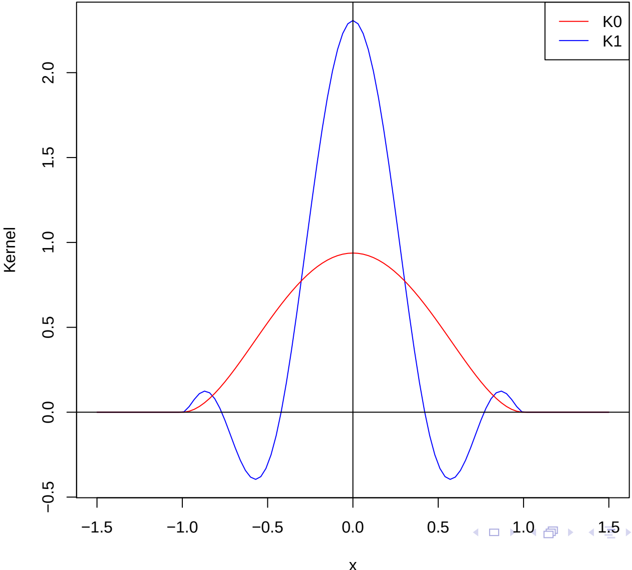
$$
\begin{aligned}
p_n(y|z_1, z_2) &= \frac{1}{nh_0 h_1 h_2} \sum_{i=1}^{n} K_0\left(\frac{Y_i - y}{h0}\right) K_1\left(\frac{Z_{1i} - z_1}{h1}\right) \\
&\quad \times K_2\left(\frac{Z_{2i} - z_2}{h_2}\right) \\
G_n(y|z_1, z_2) &= \frac{1}{nh_1 h_2 p_n(z_1, z_2)} \sum_{i=1}^{n} I\left(Y_i \le y\right) K_1\left(\frac{Z_{1i} - z_1}{h_1}\right) \\
&\quad \times K_2\left(\frac{Z_{2i} - z_2}{h_2}\right) \\
p_n(z_1, z_2) &= \frac{1}{nh_1 h_2} \sum_{i=1}^{n} K_1\left(\frac{Z_{1i} - z_1}{h_1}\right) K_2\left(\frac{Z_{2i} - z_2}{h_2}\right)
\end{aligned}
$$

Kernels

- Kernels taken from Muller (1984)

$$
\begin{aligned}
K_0 &= \frac{15}{16}\left(1 - 2x^2 + x^4\right) \\
K_1, K_2 &= \frac{315}{2048}\left(15 - 140x^2 + 378x^4 - 396x^6 + 143x^8\right)
\end{aligned}
$$

Kernels



Final Algorithm

- Note interdependence of $\hat{H}$ and $\hat{\boldsymbol{\beta}}, \hat{\gamma}$
- Iterative algorithm combines the two estimation procedures

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

Asymptotic Behavior

- Authors show that
  - $\sqrt{n}\left(\hat{H}(y) - H(y)\right)$ asymptotically normal
  - $\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$ asymptotically normal
  - $\sqrt{n}\left(\hat{\gamma} - \gamma\right)$ asymptotically normal
- Asymptotic covariances are complicated and depend on other unknown functions
  - More things to estimate!

Getting Estimate on Original Scale

- Given final estimates $\hat{H}, \hat{\boldsymbol{\beta}}$, and $\hat{\boldsymbol{\gamma}}$, estimate on original scale is

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

- Compare with Duan's smearing estimator:

$$\hat{\mu}(\mathbf{x}) \quad = \quad \frac{1}{n}\sum_{i=1}^{n} H^{-1}(x\hat{\beta} + \hat{\epsilon}_i)$$

Simulations: Setup

- Generate data according to model

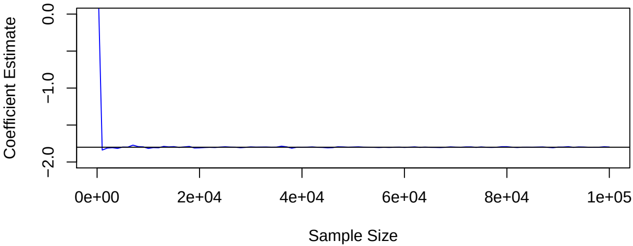
$$H(Y) \;=\; \mathbf{X}'\boldsymbol{\beta} + \sqrt{\mathbf{X}'\boldsymbol{\gamma}}\epsilon$$

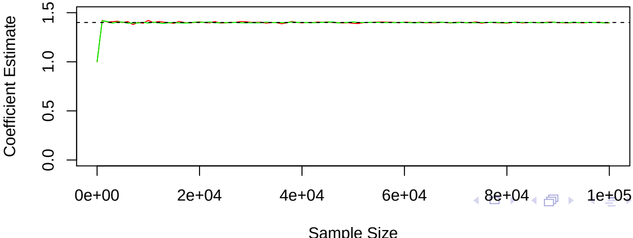
where

  - $\boldsymbol{\beta} = (-1.8, 1.4, 1.4)$
  - $\boldsymbol{\gamma} = (0.4, -0.35, 0)$
  - $X_1 \sim \textit{Bernoulli}(.5)$
  - $X_2 \sim \textit{Unif}\,(0, 2)$
  - $\epsilon \sim N(0, 1)$
  - H is related to Y via

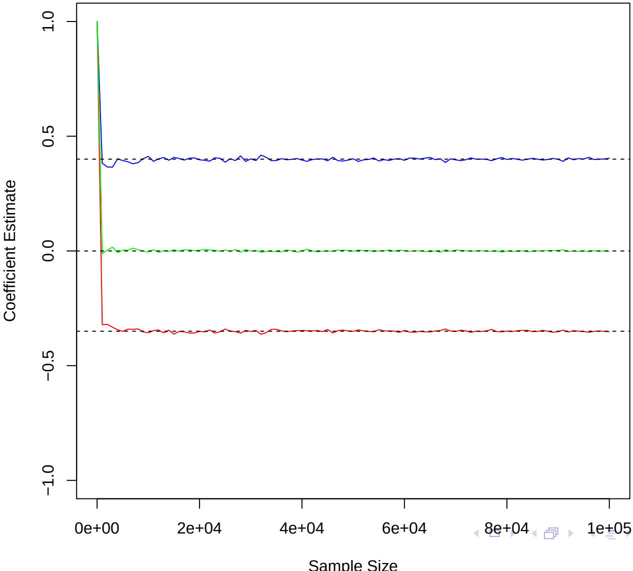
$$H(y) = \Phi^{-1}\left(\exp(y - 10)\right)$$

Simulations: Consistency of β estimates





Simulations: Consistency of γ estimates



Simulations: Duan's Smearing Estimator vs. Proposed Estimator

$\mathbf{x}$	$\mu(\mathbf{x})$	Method	Bias
(0,1)	8.795	Proposed	0.008
		Duan	0.06
(0,2)	9.753	Proposed	0.008
		Duan	0.03
(1,1)	9.818	Proposed	0.001
		Duan	0.03
(1,2)	9.990	Proposed	< 0.001
		Duan	0.006

Discussion and Critique

- Statistical Contribution
  - Extends previous methods to address issues commonly encountered in these types of data
- Scientific Contribution
  - Provides more accurate estimation of health care costs
- Implementation is slow
- Approach is somewhat unintuitive
  - Explaining to non-statistical collaborators might be difficult
- Do we really need to transform data?

Thanks for your time!