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)

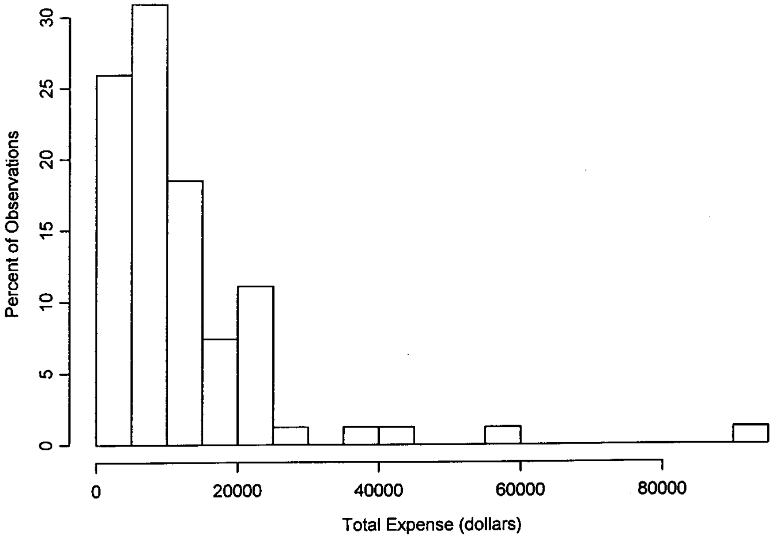
Scott Coggeshall

May 14, 2013

Motivation: Issues with Health Care Cost Data

- Predictions about individuals' health care costs important for many applications
- Predictions needed on original scale
- Transformation models are popular in this area
  - Need to specify transformation, assume homoscedasticity
  - Introduce bias

Motivation: Issues with Health Care Cost Data



Proposed Model

- Let
  - Y - an individual's observed health care cost
  - $\mathbf{X}$ - a $q \times 1$ vector of observed explanatory variables
  - $\boldsymbol{\beta}, \gamma$ - $q \times 1$ vectors of unknown parameters, to be estimated
  - $H(\cdot)$ - an unknown function, to be estimated
  - $\sigma(\cdot)$ - a known variance function
  - ϵ - an error term with mean 0, variance 1
- Proposed Model:

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

Implementation

- Recall algorithm for reaching final estimates of $H, \boldsymbol{\beta}, \gamma$:
  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

Implementation: Estimating β and γ

- Estimating equation for β:

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

- Estimating equation for γ:

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

Implementation: Estimating $\boldsymbol{\beta}$ and γ

- Estimator for $\boldsymbol{\beta}$ is easy...

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

- But γ is a bit trickier...
  - Closed form solution?
  - Newton-Rhapson?

Simulation Setup

- Generate $X_1 \sim Bernoulli(p), p = 0.5$
- Generate $X_2 \sim Unif(0, 2)$
- Let

$$\sigma(\mathbf{X}'\gamma) = \sqrt{0.4 + X_1\gamma}$$

- Generate $H(Y)$ according to

$$H(Y) = \beta_0 + X_1\beta_1 + X_2\beta_2 + \sqrt{0.4 + \gamma X_1}\epsilon$$

where $\epsilon \sim N(0, 1)$.

Estimating β and γ

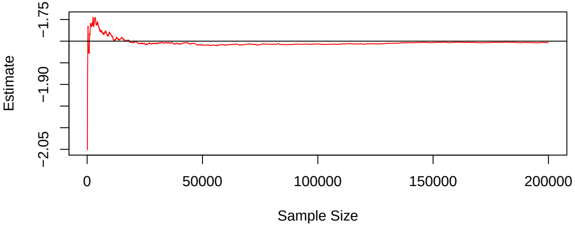
- Propose initial value of γ
- Estimate β via closed-form solution
- Think of estimating equations for γ as function $f : \mathbb{R} \to \mathbb{R}^3$
- Find derivative vector $J = \begin{pmatrix} \partial f_1/\partial\gamma \\ \partial f_2/\partial\gamma \\ \partial f_3/\partial\gamma \end{pmatrix}$
- Estimate γ with Newton Rhapson:

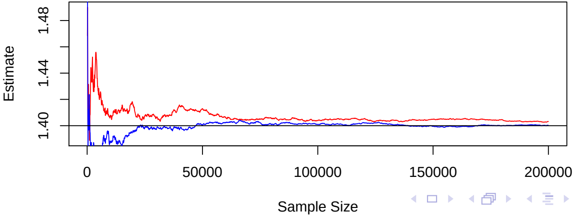
$$\gamma^{(n+1)} = \gamma^{(n)} - \left((J^T J)^{-1} J^T\right) f(\gamma^{(n)})$$

- Repeat updating process until both estimates converge

Results

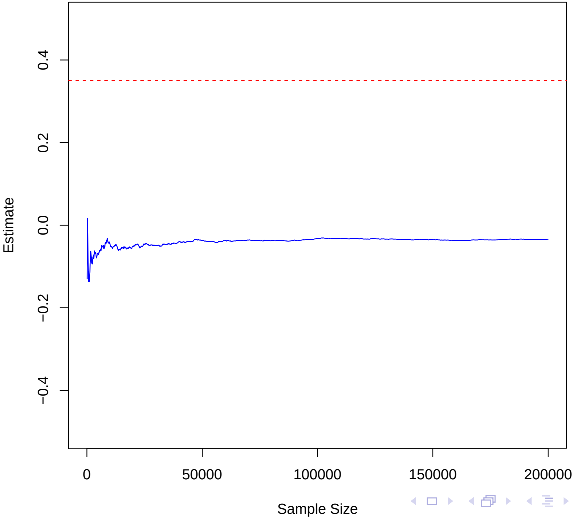
- Estimates for β are OK





Results

- Estimates for γ are way off



Implementation: Estimating H

- For each observation, define the two indices

$$Z_1 = \mathbf{X}'\beta$$
$$Z_2 = \mathbf{X}'\gamma$$

- Estimator of H is a function of $p(z_1, z_2)$ and $G(y|z_1, z_2)$ (and its derivatives)

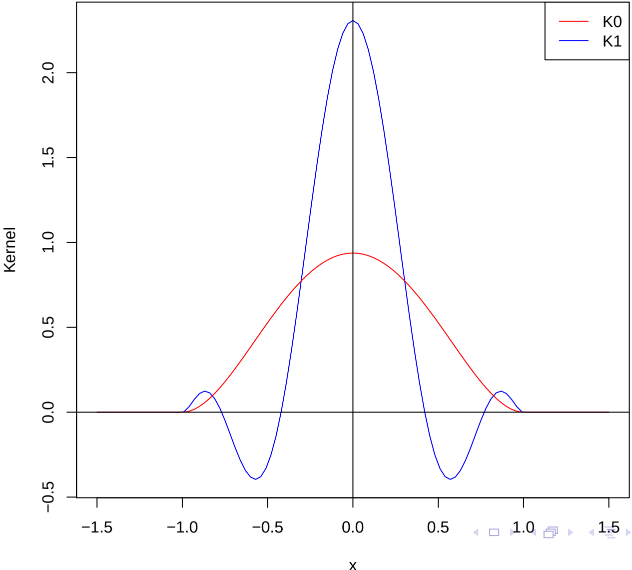
$$H(y) = -\int_{y_0}^{y} \frac{\sum_{i=1}^{n} p(u|Z_{1i}, Z_{2i})p(Z_{1i}, Z_{2i})}{\sum_{i=1}^{n} g_1(u|Z_{1i}, Z_{2i})p(Z_{1i}, Z_{2i})} du$$

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



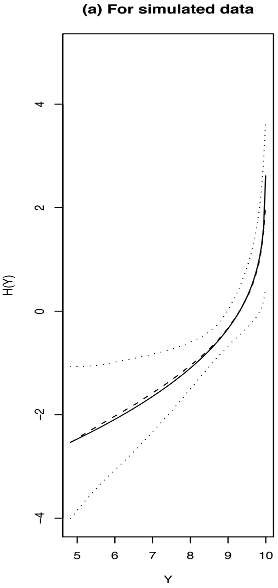
Simulation Setup

- Same as before, except no longer observe $H(Y)$
- Observe Y where Y is related to H via

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

- Goal: Estimate H, assuming we know $\boldsymbol{\beta}$ and γ

What It Should Look Like...



...and What I'm Getting

R Console

~/Desktop/plots 5-10-13

Help Search

```
> n <- 2000
> X1 <- rbinom(n,1,.5)
> X2 <- runif(n,min=0,max=2)
> errors <- rnorm(n)
> HY <- b0 + b1*X1 + b2*X2 + sqrt(0.4 + gam*X1)*errors
> Y <- 10 + log(pnorm(HY))
>
> Z1 <- b0 + X1*b1 + X2*b2
> Z2 <- 0.4 + X1*gam
>
> H_n(min(Y),median(Y),Y,Z1,Z2,h)
```

Summary and Next Steps

- Implementing the procedure is proving difficult
- Look for solution to convergence problem for estimating γ
- Speed up code for estimating H
  - Re-write code in C

Question Time

Thanks for your attention!