ヘドニック型価格指数へのリッジ回帰推定量の適用

丸山 祐造

東京大学空間情報科学研究センター

時空間社会経済システム研究部門

CSIS シンポジウム 9 月 19 日

0. きっかけ

昨年 12 月 CSIS 内の不動産関連のシンポジウムでのリクルートの清水千弘さんの発表

ヘドニック法 (回帰分析) に基づいて「性質の良い品質調整済住宅価格指数 (RRPI)，あるいはヘドニック型価格指数を求める」ことが一つの重要なテーマ

RRPI とは? 各期毎に売買事例のデータセットが得られるとき異時点間における住宅価格の変化を捉えるための価格指数

1. ヘドニック回帰と指数

首都圏における中古マンションに対する価格指数を作成

データセット (各月毎，月平均 1000 件程度のサンプル，1989.04–2003.03 全部で $T = 168$ 期)

被説明変数

中古マンションの価格

説明変数

最寄駅までの距離，都心までの接近性，専有面積，築後年数，バルコニー面積，管理費，南向きダミー，沿線ダミー，など

データのイメージを掴む

$n = 158,955$	平均	標準偏差	最小	最大
マンション価格 (万円)	3750	1802	1000	9998
最寄駅までの距離 (分)	7.6	4.2	1.0	20.0
都心までの接近性 (分)	25.2	5.0	16.3	77.5
専有面積 (m^2)	55.3	17.9	15.0	120
築年数	14.1	7.0	1.0	35.0
市場滞留時間 (月)	88.0	85.4	1.0	700

他の経済インデックスとの違い

日経平均，株価指数など (基本的に) 各期毎に同じ商品が常に市場にある．

中古マンションの例では，全く同じ条件の物件が各期毎に存在することはあり得ない!

異時点間で条件の異なる商品価格の推移から，価格変動を把握したい．

ヘドニック回帰モデル

第 1 期

$$y_{i1} = \beta_{01} + \beta_{11} x_{i1} + \cdots + \beta_{p1} x_{ip} + \epsilon,\ i = 1, \ldots, N_1$$

$$\vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots$$

第 T 期

$$y_{iT} = \beta_{0T} + \beta_{1T} x_{iT} + \cdots + \beta_{pT} x_{ip} + \epsilon,\ i = 1, \ldots, N_T$$

ϵ は誤差項で，平均 0 分散 σ^2 の正規分布
また物件間の誤差項は互いに独立であると仮定

既存の指数

構造制約型

全期間を通じて不動産市場の価格メカニズムの構造 (つまり回帰係数) は変化しない

$$\beta_{11} = \cdots = \beta_{1T}, \quad \ldots \quad \beta_{p1} = \cdots = \beta_{pT}$$

全てのデータをプール
各期のダミー変数 (1 or 0) を説明変数 $z_1, \ldots z_T$ として加えて，最小二乗法

$i = 1, \ldots, N(= \sum_{i=1}^{T} N_i)$ **に対して**

$$
\begin{aligned}
y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \\
p_1 z_1 + \cdots + p_T z_T + \epsilon
\end{aligned}
$$

各期ダミーの回帰係数の推定量を $\hat{p}_t$
基準期のダミーの回帰係数の推定量 $\hat{p}_1$

$\hat{p}_t / \hat{p}_1 (t = 1, \ldots, T)$ **が指数**

構造非制約型

各期毎に構造は変化する

各期毎に最小二乗法 $\hat{\beta}_t$

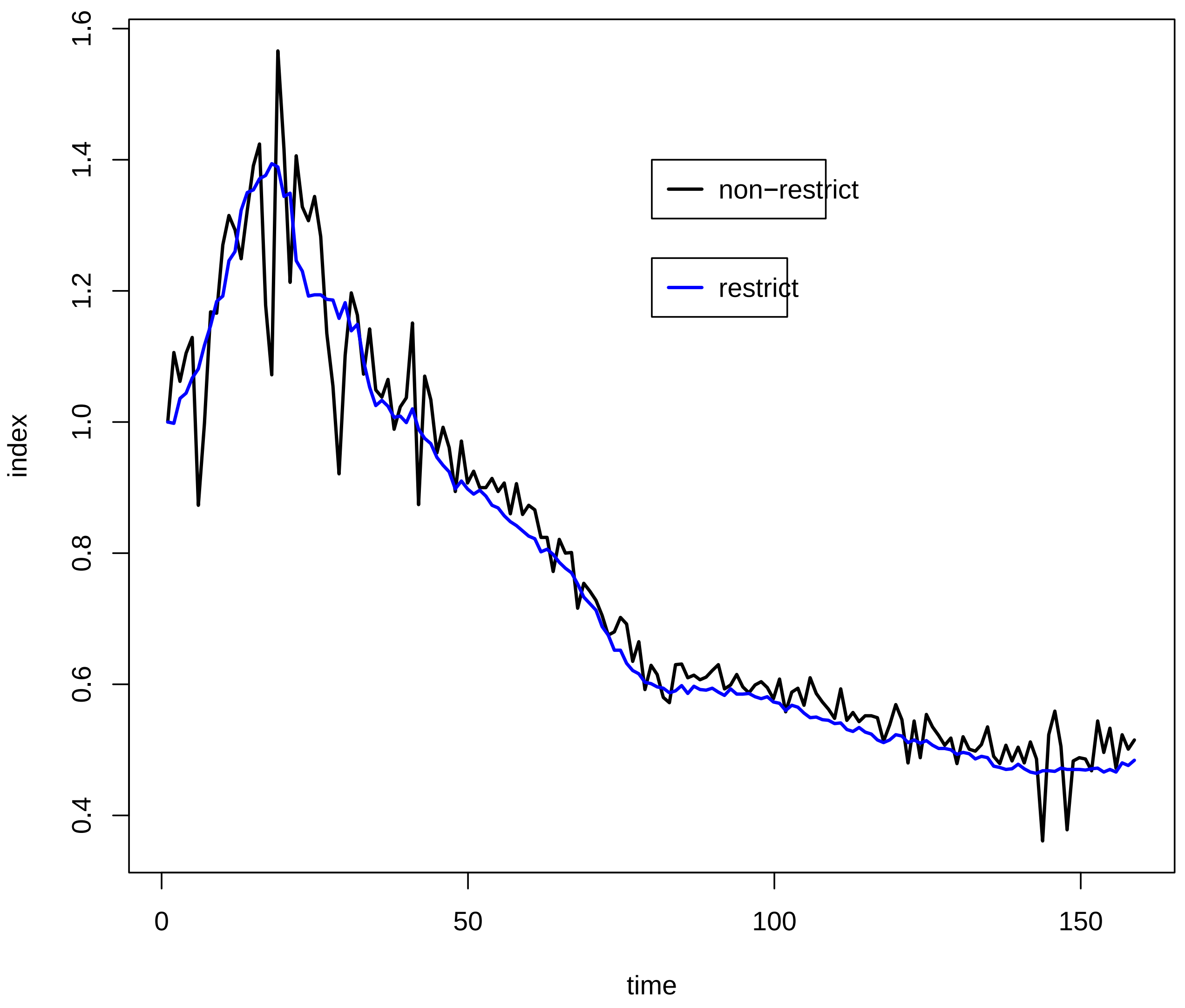
注目する品質 $\boldsymbol{x}' = (x_1, \ldots, x_p)$

(例：桜上水徒歩 10 分，65 平米，3LDK，築 3 年，南向きなど)

を各期毎に代入した注目品質の推定値 $\hat{y}_t = \boldsymbol{x}'\hat{\beta}_t$

基準期の注目品質の推定値 $y_1 = \boldsymbol{x}'\hat{\beta}_1$

$y_t/y_1(t = 1, \ldots, T)$ **が指数**



index
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0
50
100
150
time
non−restrict
restrict

制約型の欠点

- 指数変化は比較的滑らかだが，全期間を通じて構造変化がないというのはムチャな話
- 特定の回帰係数の推移にも関心がある場合に対応不可能
- 逐次的に最新のデータを得て指数を得る場合，更新毎に過去の指数が変化

非制約型の欠点

- 指数変化がかなりギザギザ
- 近接する期とは構造が近いはずでその情報を使わないのはあまりにもったいない

回帰分析の復習

$$y \doteq \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

回帰係数 β_i

説明変数 x_i を 1 単位増加させたときに (他の説明変数の値は固定した下で) 中古マンション価格 y が平均的に何円増えるか?

各期毎に回帰係数を推定することのメリット

例えば,,,,

南向きということに消費者がどれほどの価値を置く傾向にあるかということに関する推移を推測することも出来る.

総武線沿線であること，東急東横線沿線であることの効果の推移，あるいはそれらの効果の比較も出来る.

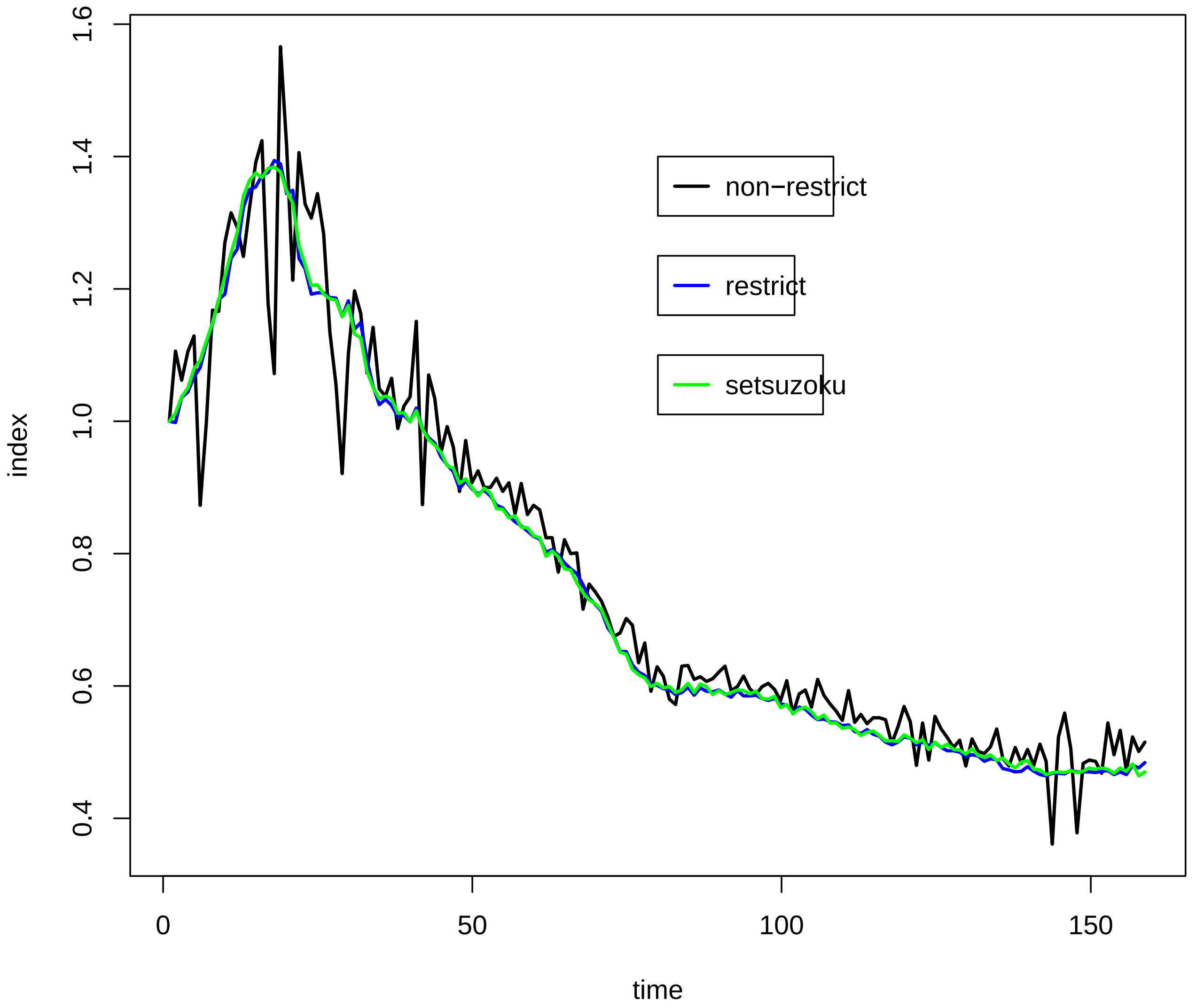
適度な構造制約を入れながら，滑らかに接続する指数が望ましい

(R 清水タイプ) 接続型

過去一定期間の構造は変化しないとして，構造制約型を適用し逐次的に指数を作成

欠点

- 構造制約型と同じく更新の度に過去の指数が変化．
- 過去の一定期間と言ってもその長さをどうやって決めるのか?


index
time
0.4
0.6
0.8
1.0
1.2
1.4
1.6
0
50
100
150
non−restrict
restrict
setsuzoku

本講演で提案する指数

「一期前の構造とは近い」という情報だけ．
基本は非制約型
ベイズ，カルマンフィルター，リッジ回帰のアイデアを組み合わせる

特徴

- 滑らかさ 制約と非制約の中間
- 新データ取得後も，過去の指数は変化なし
- 各回帰係数の推移も把握可能

2. 問題設定

とりあえず t 期だけに限定．後で時系列．

$$y_t = A_t\beta_t + \epsilon \quad t \text{ 期のヘドニックモデル}$$

$\epsilon \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}_{N_t})$, 分散 σ^2 未知，$\boldsymbol{I}$ は単位行列

β_t: p 次元回帰係数ベクトル 未知

A_t: $N_t \times p$ ランク p の説明変数行列 既知

回帰係数ベクトル β の良い推定量を求めたい

通常の推定量　最小二乗推定量

$$\hat{\beta} = (A'A)^{-1}A'y \sim N(\beta, \sigma^2(A'A)^{-1})$$

都合が悪い状況

- A(説明変数を p 個並べた行列) のいくつかが高い相関を持っている
- $A'A$ は極めて小さい固有値を持つ
- 分散共分散行列が $(A'A)^{-1}$ であることに注意すると $\hat{\beta}$ は分散が極めて大きい (精度が低く，不安定な) 推定量
- 例えば $E[\hat{\beta}'\hat{\beta}]$ はとても大きい可能性がある

実データではどうなっているか?

例えば専有面積とバルコニー面積，管理費は比較的高い相関を持つ．

	97	98	99	00	01	02	03
専バ	0.49	0.5	0.51	0.49	0.53	0.52	0.49
専管	0.68	0.67	0.67	0.65	0.57	0.64	0.67

相関係数が驚く程安定していることにも注意

3. リッジ回帰推定量とベイズの解説

Hoerl and Kennard (1970)

制約つき最小二乗法

$$\min_{\boldsymbol{\beta}}\{(\boldsymbol{y}-\boldsymbol{A\beta})'(\boldsymbol{y}-\boldsymbol{A\beta})\} \quad \text{s.t. } \boldsymbol{\beta}'\boldsymbol{\beta} \le M$$

$$\begin{aligned}\hat{\boldsymbol{\beta}}^R(\lambda) &= [\boldsymbol{A}'\boldsymbol{A}+k\boldsymbol{I}]^{-1}\boldsymbol{A}'\boldsymbol{y} \\ &= (\boldsymbol{I}-[\boldsymbol{I}+\lambda\boldsymbol{A}'\boldsymbol{A}]^{-1})\hat{\boldsymbol{\beta}}\end{aligned}$$

ここで $\lambda = 1/k$ **である.**

リッジ回帰推定量のベイズ的解釈

リッジ回帰推定量　事前分布 $\beta \sim N(\mathbf{0}, \lambda\sigma^2 \boldsymbol{I}_p)$ に関するベイズ推定量

実は 0 でなくても任意の実数ベクトルで定義可能 (これが後で効いてくる)

- 問題は λ の決め方
- また事前分布の分散共分散行列が単位行列の定数倍であるのは，自然ではない．

新たなベイズ推定量を提案

ベイズ統計・ベイズ推定量 超速習

簡単のため一次元 $X_1, \dots, X_n \sim N(\theta, \sigma^2)$

通常の自然な推定量 標本平均 $\bar{X} = (1/n) \sum X_i$

ベイズ　事前分布 $\theta \sim N(\mu, \tau^2)$ を想定

ベイズ推定量は重み付き平均

$$\frac{\frac{n}{\sigma^2}\bar{X} + \frac{1}{\tau^2}\mu}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} = c\bar{X} + (1-c)\mu, \quad c = \frac{\frac{n}{\sigma^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}$$

c 非常に自然な係数

サンプルが十分情報を含めば，($\Leftrightarrow n$ 大，σ^2 小)，$c \to 1$ つまりベイズは $\bar{X}$ に近い

逆に事前情報が正確であれば ($\Leftrightarrow$ τ^2 小)，$c \to 0$ つまりベイズ推定量は μ に近い．

ただ σ^2 や τ^2 は未知なので，つまり重み係数 c は未知なのでデータから推定する必要がある．

4. 新たなリッジ回帰型推定量

問題設定の再確認 (t 期だけに限定)

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{y} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{A}'\boldsymbol{A})^{-1})$$

$$S/\sigma^2 \sim \chi^2_n \quad \textbf{where} \quad S = (\boldsymbol{y} - \boldsymbol{A}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{A}\hat{\boldsymbol{\beta}})$$

事前分布 $\boldsymbol{\beta}|\eta, \lambda \sim N(\boldsymbol{0}, \eta^{-1}\{\lambda^{-1}d_1\boldsymbol{I}_p - (\boldsymbol{A}'\boldsymbol{A})^{-1}\})$

$$\lambda \sim \lambda^a(1-\lambda)^b I_{(0,1)}(\lambda), \quad \eta \sim \eta^c I_{(0,\infty)}(\eta), \quad \eta = \sigma^{-2}$$

d_1 は、$(\boldsymbol{A}'\boldsymbol{A})^{-1}$ の最大固有値

提案するベイズ推定量

$b = n/2 + c - a - 1, \gamma = (p/2 + a + 1)/(n/2 + c - a)$

$$\delta_{\gamma}^{SB} = \left(\boldsymbol{I} - \frac{\gamma}{w + \gamma + 1} d_1^{-1} (\boldsymbol{A}'\boldsymbol{A})^{-1} \right) \hat{\boldsymbol{\beta}}$$

$w = \hat{\boldsymbol{\beta}}' \hat{\boldsymbol{\beta}} / (d_1 S)$

- **リッジ回帰推定量の一般化**
- **最小二乗推定量 $\hat{\beta}$ と事前情報の平均値 0 の重み付き平均**
- **γ が小さい程，$\hat{\beta}$ に近い**

γ の決め方

期を固定すれば数学的には容易．

基準 $(\delta-\beta)'(\delta-\beta)/\sigma^2$ の期待値 $R(\delta,\beta,\sigma^2)$

$0 \le \gamma \le 2(\sum_{i=1}^{p} d_i^2/d_1^2 - 2)/(n+2)$ とすると，

$$R(\delta_{\gamma}^{SB},\beta,\sigma^2) \le R(\hat{\beta},\beta,\sigma^2) \quad \forall \beta,\ \sigma^2$$

例えば $\gamma = (p-2)/(n+2)$ とする

(固有値がすべて等しい場合の最適な γ)

各期毎に primitive な最小二乗推定量 (OLS) よりも良い

接続を考えた場合，最適な γ の選択はやや困難

5. スムーズな接続に向けて

事前分布の平均ベクトルが 0 だったのは本質的ではない

一期前の推定量を事前分布の平均ベクトルとして想定する

$$\boldsymbol{\beta}_t|\eta,\lambda \sim N(\hat{\boldsymbol{\beta}}_{t-1},\eta^{-1}\{\lambda^{-1}d_1\boldsymbol{I}_p-(\boldsymbol{A}'\boldsymbol{A})^{-1}\})$$

$$\lambda \sim \lambda^a(1-\lambda)^b I_{(0,1)}(\lambda), \quad \eta \sim \eta^c I_{(0,\infty)}(\eta)$$

実は離散時間カルマンフィルターと同じこと

第 $t-1$ 期の回帰係数 β の推定量 $\hat{\boldsymbol{\beta}}_{t-1}$
第 t 期の OLS $\hat{\boldsymbol{\beta}}_{OLS,t}$，残差平方和 S_t

提案するリッジ回帰型ベイズ推定量

$$\hat{\boldsymbol{\beta}}_t^R = (\boldsymbol{I} - \boldsymbol{C})\hat{\boldsymbol{\beta}}_{OLS,t} + \boldsymbol{C}\hat{\boldsymbol{\beta}}_{t-1}$$

ここで $p \times p$ 行列 C_t は

$$\boldsymbol{C} = \frac{\gamma S_i}{(\hat{\boldsymbol{\beta}}_{OLS,i} - \hat{\boldsymbol{\beta}}_{i-1})'(\hat{\boldsymbol{\beta}}_{OLS,i} - \hat{\boldsymbol{\beta}}_{i-1}) + (\gamma + 1)d_1 S_i}(\boldsymbol{A}'\boldsymbol{A})^{-1}$$

A は t 期の説明変数行列

提案するリッジ回帰型ベイズ推定量 (正式版)

$$\hat{\boldsymbol{\beta}}^R_t = (\boldsymbol{I} - \boldsymbol{C})\hat{\boldsymbol{\beta}}_{OLS,t} + \boldsymbol{C}\hat{\boldsymbol{\beta}}^R_{t-1}, \quad t = 2, \cdots$$

ここで $p \times p$ 行列 C_t は

$$\boldsymbol{C} = \frac{\gamma S_i}{(\hat{\boldsymbol{\beta}}_{OLS,i} - \hat{\boldsymbol{\beta}}^R_{i-1})'(\hat{\boldsymbol{\beta}}_{OLS,i} - \hat{\boldsymbol{\beta}}^R_{i-1}) + (\gamma + 1)d_1 S_i}(\boldsymbol{A}'\boldsymbol{A})^{-1}$$

推定量の性質

- 多重共線性による弊害を防ぐ
- 一期前の推定量と今期の OLS の重みつき平均
- γ が大きい程，一期前の推定量に近付く (収束先は別だが,,)
- OLS の精度の低い component をより一期前の推定に近づけ，精度の高い component はあまり近づけない (自動的に調整)
- γ を小さく適切に決めると，各期毎に OLS よりも良い

リッジ回帰ヘドニック型価格指数

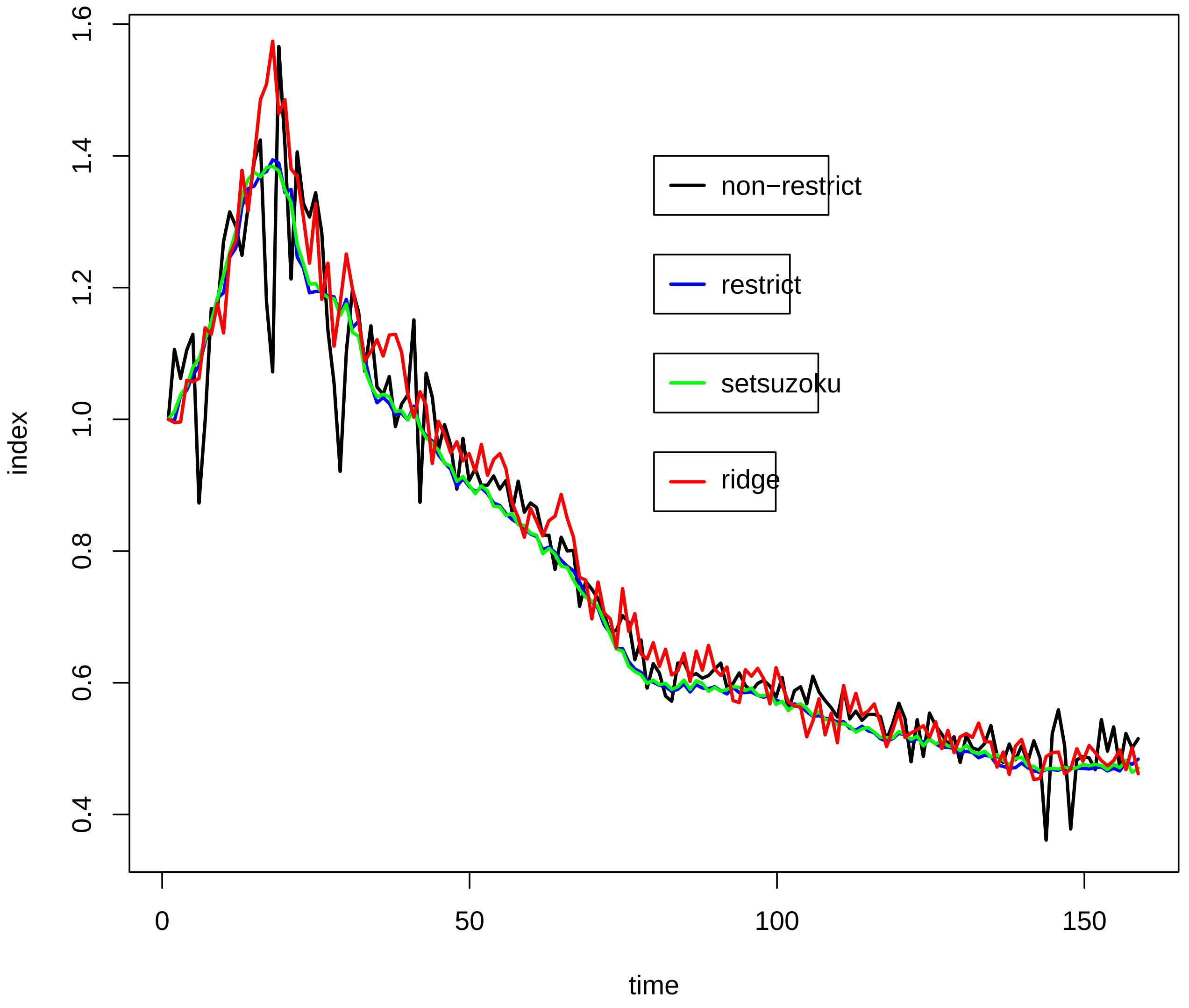
全 T 期の回帰係数

$$\hat{\beta}_1 \to \hat{\beta}_2^R \to \cdots \to \hat{\beta}_T^R$$

注目する品質 $\boldsymbol{x} = (x_1, \ldots, x_p)$

$$1 \to \frac{\boldsymbol{x}\hat{\beta}_2^R}{\boldsymbol{x}\hat{\beta}_1} \to \cdots \to \frac{\boldsymbol{x}\hat{\beta}_T^R}{\boldsymbol{x}\hat{\beta}_1}$$

- **非制約型価格指数よりも滑らかに推移**
- **急な変化に対してロバスト**


index
time
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0
50
100
150
non−restrict
restrict
setsuzoku
ridge

6. 今後の課題

- γ の選び方に自由度．今回は目で見て決めただけ．最適性の基準を考える．

trade off

・各期毎に OLS よりも良くするためには，γ は小さい方が良い．(但し前期の改良の条件は十分条件．γ が少々大きくても OLS よりも数値的には良いはず．)

・指数が滑らかに推移するためには γ が大きい方が良い．

他のデータへの適用

総務庁統計局の消費者物価指数

POS データを使ったパソコンの価格推移にヘドニック法を用いる

説明変数 (CPU，重さ，画面の大きさ，USB の個数，メーカーダミー，DVD ダミーなど)

半年経つと同じ製品はないと思ってよい