Estimation of Transformation Model

for Mortgage Prepayment Data

Junyi Zhang

Department of STATS and CIS
Baruch College

Aug 12^{th} 2015

[1] Joint work with Dr. Ying and Dr. Jin, Columbia Univ., Dr. Shao, NYU, and Mr. Gao.

Outlines

- Single Family Loan-Level Dataset
- Cox Proportional Hazard Model for Prepayment Time
- Transformation Model
- Self-Induced Smoothing
- Results

Data by Freddie Mac

- Provided by Freddie Mac
- Two groups of data
  1. origination data file:
     note rate, maturity, first payment date, unpaid balance, loan-to-value ratio, loan purpose (purchase/refi) MIP (mortgage insurance), first-time homebuyer flag, credit score (FICO), debt-to-income, occupancy status (own/invest/second), state, postal code
  2. monthly loan performance data filep:
     current UPB, delinquency status, loan age, remaining terms, zero balance code(prepaid/foreclosure/repurchse), MI recoveries, net sales proceeds

Data by Freddie Mac

- Data is split by origination year:
  year vintage 1999 through year vintage 2014
- Missing in the loan performance data:
  first 6 month performance, origination date
- Dynamic data:
  origination data: new mortgage;
  performance data: continuing UPB or becomes zero balance
- Performance Cut-off date in the analysis: September 2014 (right censored data).

Macro-economic data

- Home Price Index: state-wide monthly data by Freddie
- Market CMM (constant maturity mortgage) rate by Freddie
- To build major prepayment risk factors:
  1. Loan-level SATO (spread-at-origin):
     Note rate - CMM rate at origination (origination date is missing)
  2. Loan-level HPI:
     monthly HPI for the corresponding state (time-varying)
  3. Loan-level Financial Incentive:
     current actual unpaid-balance(f(note rate))/refinanced unpaid-balance(f(current market CMM rate)).

Goal of Study

- Prepayment risk: a major risk factor in pricing MBS (mortgage backed securities)
- Loan-level prepayment risk prediction (multilogit model by Calhoun and Deng 2002)
- Decompose the prepayment risk

Cox PH models

- Let T_i be time to prepayment.
- Cox's proportional hazard regression:

$$\lambda(T_i = t|X_i) = \lambda_0(t|X_0) \times \exp\{X_i\beta\}$$

- Choosing covariates $X_i = FI_i \times [1, SATO_i, FICO_i, LoanSize_i, LTV_i \times \mathrm{I}_{(LTV_i <= 80)}, LTV_i \times \mathrm{I}_{(LTV_i > 80)}, DTI_i, HPI_i]$

Class of semiparametric models

- Linear Regression:
$$Y_i = X_i'\beta + \epsilon_i$$
where ϵ_i are iid $N(\mu, \sigma^2)$

- Drop normality:

  ϵ_i's are i.i.d. with an unknown distribution. Then
$$E(Y|X) = X'\beta.$$

Class of semiparametric models

- Semiparametric linear transformation model:

$$Y_i = H(X_i'\beta + \epsilon_i)$$

where H is monotone and ϵ_i are iid with a completely specified distribution. Examples:

ϵ	dist. property	Model
extreme value	$\lambda(y\|X) = \exp(H^{-1}(y) - X'\beta)$	Cox's PH
logistic	$O(y\|X) = \exp(H^{-1}(y) - X'\beta)$	proportional odds

Class of semiparametric models

- General class of models:

$$Y_i = H \circ F(X_i'\beta, \epsilon_i)$$

  1. I.I.D. ϵ_i's.
  2. I.I.D. X_i's; and independent with ϵ_i's.
  3. Monotone increasing function $H(\cdot)$.
  4. Function $F(\cdot, \cdot)$ is strictly increasing in each of its arguments.

- Model identifiability: assume $\beta = (\theta, 1)'$ and $H^{-1}(y_0) = 0$.

Rank Correlation Function

- Define rank correlation (Kendall Tau; $(Y_i, X_i'\beta)$)

$$Q_n(\theta) = \frac{1}{n(n-1)} \sum_{i \neq j} \mathrm{I}(Y_i > Y_j)\mathrm{I}(X_i'\beta > X_j'\beta).$$

- Define the MRC estimator (Han, 1987) as

$$\beta_n(\theta_n) = \arg\max_{\theta} Q_n(\beta(\theta)).$$

The MRCE's large-sample properties

- Strong consistency: HAN, A.K.(1987), *J. Econometrics*
- $\sqrt{n}$-consistency and normality: Sherman, R. (1993), *Econometrica*
- Asymptotic covariance matrix:
  - Asymptotic variance is $D_0 = A^{-1}VA^{-1}$,

    where $2A = E\nabla_2\tau$, $V = E(\nabla_1\tau)^{\otimes 2}$ and

    $$\tau(y, x, \theta) = E^{y,x}\left[\mathrm{I}_{y>Y}\mathrm{I}_{(x-X)'\beta>0} + \mathrm{I}_{Y>y}\mathrm{I}_{(X-x)'\beta>0}\right].$$

Difficulties

- Rank correlation is discontinuous in θ.
- Possible approaches to estimate Σ_{MRC}:
  1. Finite difference (Sherman; bandwidth selection problem).
  2. Boostrap method (Subbotin, 2007; expansive computation).
  3. Stochastic perturbation (Jin et al., 2001; computation).

Smoothing Cont'd

- Induced smoothing for score functions. (Brown and Wang, 2005)
- For $\sqrt{n}$-consistent $\hat{\theta}$, $\theta - \hat{\theta}$ is approximately a Gaussian noise $Z/\sqrt{n}$ where $Z \sim N(0, \Sigma)$ and Σ is the limiting covariance matrix of the MRC estimator.
- Self-induced smoothing:
  - Smoothed rank correlation:

$$\begin{aligned}\widetilde{Q}_n(\theta) &= E_Z Q_n(\theta + Z/\sqrt{n}) \\ &= \frac{1}{n(n-1)} \sum_{i \neq j} I[Y_i > Y_j] \Phi\left(\frac{\sqrt{n} X_{ij}' \beta(\theta)}{\sigma_{ij}}\right),\end{aligned}$$

where $X_{ij} = X_i - X_j$, $\sigma_{ij} = \sqrt{(X_{ij}^{(1)})' \Sigma X_{ij}^{(1)}}$.

Smoothing Cont'd

- The smoothed MRC estimator: $\widetilde{\theta}_n = \arg\max_{\theta} \widetilde{Q}_n(\theta)$.
- Variance estimator: $\widehat{D}_n(\theta, \Sigma) = \widehat{A}_n^{-1}(\theta, \Sigma)\widehat{V}_n(\theta, \Sigma)\widehat{A}_n^{-1}(\theta, \Sigma)$, where

$$\widehat{A}_n(\theta, \Sigma) = \frac{1}{2n(n-1)} \sum_{i \neq j} \left\{ H_{ij} \dot{\phi} \left(\frac{\sqrt{n} X_{ij}' \beta}{\sigma_{ij}} \right) \left[\frac{\sqrt{n} X_{ij}^{(1)}}{\sigma_{ij}} \right]^{\otimes 2} \right\},$$

and

$$\widehat{V}_n(\theta, \Sigma) = \frac{1}{n^3} \sum_{i=1}^{n} \left\{ \sum_j \left[H_{ij} \phi\left(\frac{\sqrt{n} X_{ij}' \beta}{\sigma_{ij}}\right) \frac{\sqrt{n} X_{ij}^{(1)}}{\sigma_{ij}} \right] \right\}^{\otimes 2}.$$

Here $H_{ij} = \mathrm{sgn}(Y_i - Y_j)$, and $\dot{\phi}(z) = -z\phi(z)$.

Iterative algorithm

- The limiting covariance matrix Σ is unknown in $\hat{D}_n(\theta, \Sigma)$.
- An iterative algorithm:

1. Compute the MRC estimator $\hat{\theta}_n$ and set $\hat{\Sigma}^{(0)}$ to be the identity matrix.
2. Update variance-covariance matrix $\hat{\Sigma}_n^{(k)} = \hat{D}_n(\hat{\theta}_n, \hat{\Sigma}_n^{(k-1)})$. Smooth the rank correlation $Q_n(\theta)$ using covariance matrix $\hat{\Sigma}_n^{(k)}$. Maximize the resulting smoothed rank correlation to get an estimator $\hat{\theta}_n^{(k)}$.
3. Repeat step 2 until $\hat{\theta}_n^{(k)}$ converge.

Asymptotic Equivalency

Theorem

For any positive definite matrix Σ, under certain regularity conditions, the smoothed MRC estimator $\widetilde{\theta}_n$ is consistent, $\widetilde{\theta}_n \rightarrow \theta_0$ a.s., and asymptotically normal,

$$\sqrt{n}(\widetilde{\theta}_n - \theta_0) \Rightarrow N(0, A^{-1}VA^{-1}).$$

In addition, the SMRCE $\widetilde{\theta}_n$ is asymptotically equivalent to the MRCE $\hat{\theta}_n$ in the sense that

$$\widetilde{\theta}_n = \hat{\theta}_n + o_p(n^{-1/2}).$$

Consistent Variance Estimator

Theorem

For any positive definite matrix Σ, the variance estimator $\hat{D}_n(\hat{\theta}_n, \Sigma)$ converges in probability to D_0, the limiting variance-covariance matrix of the MRC estimator $\hat{\theta}_n$.

Algorithm Convergence

Theorem

Let $\hat{\Sigma}_n^{(k)}$ be defined as in the iterative algorithm. Under certain regularity conditions, there exist Σ_n^*, $n \geq 1$, such that for any $\epsilon > 0$, there exists N, such that for all $n > N$,

$$P(\lim_{k\to\infty} \hat{\Sigma}_n^{(k)} = \Sigma_n^*, \;\; \|\Sigma_n^* - D_0\| < \epsilon) > 1 - \epsilon.$$

Model and Chen's Method

- An equivalent transformation model (Λ is strictly monotone):

$$\Lambda(Y_i) = X_i'\beta + \epsilon_i$$

- Chen's (2002) rank-based estimate:

$$Q_n^{\Lambda}(y, \Lambda, b) = \frac{1}{n(n-1)} \sum_{i \neq j} (d_{iy} - d_{jy_0}) I[X_i'b - X_j'b \geq \Lambda],$$

where $d_{iy} = I[Y_i \leq y] = I[X_i'\beta + \epsilon_i \leq \Lambda_0(y)]$.
Define

$$\hat{\Lambda}_n(y) = \arg\max_{\Lambda \in M_\Lambda} Q_n^{\Lambda}(y, \Lambda, b_n)$$

for any given $y \in [y_2, y_1]$, where b_n is the $\sqrt{n}$-consistent estimator for β, for example, Han's MRC estimator.

Smoothing

- The smoothed rank correlation function:

$$\tilde{Q}_n^{\Lambda}(y, \Lambda, b) = \frac{1}{n(n-1)} \sum_{i \neq j} (d_{iy} - d_{jy_0}) \Phi\left(\sqrt{n}(X_{ij}' b - \Lambda)\right).$$

- Define the smoothed rank estimator

$$\tilde{\Lambda}_n(y) = \arg\max_{\Lambda \in M_\Lambda} \tilde{Q}_n^{\Lambda}(y, \Lambda, b_n)$$

for any given $y \in [y_2, y_1]$.

Covariance function

- Define

$$\hat{V}_n^{\Lambda}(y, y', \Lambda, b) = \frac{1}{n^3} \sum_{i=1}^{n} \Bigg\{ \sum_{j} \Big\{ n(d_{iy} - d_{jy_0})(d_{iy'} - d_{jy_0}) \\ \phi\left(\sqrt{n}(X_{ij}'b - \Lambda(y))\right) \phi\left(\sqrt{n}(X_{ij}'b - \Lambda(y'))\right) \Big\} \Bigg\}$$

- Define

$$\hat{A}_n^{\Lambda}(y, \Lambda, b) = \frac{1}{2n(n-1)} \sum_{i \neq j} \left\{ n(d_{iy} - d_{jy_0}) \dot{\phi}\left(\sqrt{n}(X_{ij}'b - \Lambda(y))\right) \right\},$$

- Define

$$\hat{D}_n^{\Lambda}(y, y', \Lambda, b) = \left[\hat{A}_n^{\Lambda}(y, \Lambda, b)\right]^{-1} \hat{V}_n^{\Lambda}(y, y', \Lambda, b) \left[\hat{A}_n^{\Lambda}(y', \Lambda(y'), b)\right]^{-1}.$$

Large-sample properties

Theorem

Under certain regularity conditions,

(i) $\sup_{y_2 \le y \le y_1} |\tilde{\Lambda}_n(y) - \Lambda_0(y)| = o_p(1)$*;*

(ii) Uniformly over $y \in [y_2, y_1]$,

$$\sqrt{n}(\tilde{\Lambda}_n(y) - \Lambda_0(y)) \Rightarrow H_\Lambda(y_0, y)$$

where $H_\Lambda(y_0, y)$ is a Gaussian process with mean 0 and a covariance function $\Gamma^\Lambda(y, y'; y_0)$.

(iii) The limiting Gaussian process for $\sqrt{n}(\tilde{\Lambda}_n(y) - \Lambda_0(y))$ is the same as that for $\sqrt{n}(\hat{\Lambda}_n(y) - \Lambda_0(y))$.

Large-sample properties, Cont'd

Theorem

Under certain regularity conditions, The covariance estimate $\hat{D}_n^{\Lambda}(y, y', \tilde{\Lambda}_n, b_n)$ *converges in probability to the limiting covariance function* $\Gamma^{\Lambda}(y, y'; y_0)$ *uniformly over* $\{(y, y') : y \in [y_2, y_1], y' \in [y_2, y_1]\}$.

Data fact

- Fitting period: originated in 2008-2013.
- More than 5 millions mortgages.
- Censoring rate is about 50%.

SATO

FICO

LTV

DTI

HPI

FI

Origination Year 2008

Origination Year 2009

Origination Year 2010

Origination Year 2011

Origination Year 2012

Reference

- HAN, A. K. (1987). Non-parametric analysis of a generalized regression model. *J. Econometrics*, **35**, 303-316.
- SHERMAN, R. P. (1993). The limit distribution of the maximum rank correlation estimator. *Econometrica*, **61** 123-137.
- KHAN, S., TAMER, E. (2007). Partial rank estimation of duration models with general forms of censoring. *J. Econometrics*, **136**, 251-280.
- BROWN, B. M. AND WANG, Y. (2005) Standard errors and covariance matrices for smoothed rank estimators. *Biometrika*, **92**, 149-158.
- JIN, Z., YING, Z., WEI L. J. (2001). A simple resampling method by perturbing the minimand. *Biometrika*, **88**, 381-390.