2011年7月14日　統計数理研究所　オープンハウス

2項モデルによる予測と閾値の選択

川崎　能典　モデリング研究系　准教授

概要: 本報告では, 2項モデルで表現される保険・与信契約といった事象にかかわるリスク総額最小化問題を総利潤の側から捉え, 総利潤を確率収束の意味で最大化する最適2項予測を与え，その形が簡潔な閾値モデルに帰着することを示す．現実には, 不連続関数である指示関数に最尤推定値を代入して予測を行う．こうした際でも, 次期に実現する最大化総利潤の予想区間ないし最小予想区間が, 通常の漸近論の結果から構成される．数値実験では, 我々の漸近論の結果は有限標本における近似も良好である．[本研究は，赤司健太郎氏（現・学習院大学，2011年3月までリスク解析戦略研究センター特任研究員）との共同研究である．]

1. 問題設定

- y_i $(i = 1, \cdots, N)$: 2値を取る観測値．例えば被保険者の事故発生時に1，無事故で0とコード化.
- $(\mathbf{x}_i, \boldsymbol{\beta})$はそれぞれ説明変数(顧客属性など)と係数ベクトル.
- 指示関数でモデル化: $y_i = \mathbb{1}\{\ \boldsymbol{\beta}'\mathbf{x}_i + \epsilon_i \geq 0\ \}$
- $\mathbf{x}_i$と誤差項ϵ_i独立のとき, Fで$-\epsilon_i$の累積分布関数を表せば事故率は, $\Pr(y_i = 1|\mathbf{x}_i) = F(\boldsymbol{\beta}'\mathbf{x}_i)$
- ここではϵ_iにロジスティック分布を仮定(ロジットモデル)．未知係数$\boldsymbol{\beta}$は最尤法で推定.
- $(\mathbf{x}_i, \epsilon_i)$は互いに独立$(i = 1, \cdots, N)$．一方$\mathbf{x}_i$は同一分布に従わなくてもよい. 貸し手にとっては$F$は既知.
- $r_i(> 0)$: 貸し手である企業側が, 無事故の際に受け取る元本と利子, あるいは保険料などの利益
- $d_i(> 0)$: 事故発生時の貸し倒れや事故補償などの損失額
- z_i: 貸し手が貸出しや契約を結ぶか否か(貸す時は1, その他で0).
- 指示関数でモデル化: $z_i = \mathbb{1}\{\ F(\boldsymbol{\beta}'\mathbf{x}_i) \leq c_i\ \}$
- c_iは$0 < c_i < 1$となる閾値(cut-off point)．事故率が閾値以下であれば貸出や契約を行う.

問題

どのようなc_iを選べばz_iを含む損失関数が最適化されるか

2. 最適2項予測と最大化総利潤の予想区間

利潤$\pi(z_i)$を以下の通り定式化し,

$$\begin{aligned}\pi(z_i) &= r_i(1 - y_i)z_i - d_i y_i z_i \\ \max_{z_i} \mathcal{E}[\pi(z_i)] &= \max_{z_i} \mathcal{E}[r_i(1 - y_i)z_i - d_i y_i z_i]\end{aligned}$$

により期待利潤の最大化問題を考える．このとき

密度関数の微分可能性や正値性を仮定して，期待利潤最大を与える定数閾値c_i^* $(i = 1, \cdots, N)$は,

- (r_i, d_i)が定数(r_{0i}, d_{0i})の場合，(r_i, d_i)と$(\mathbf{x}_i', \epsilon_i)$が独立である場合は, それぞれ$c_i^* = \frac{r_{0i}}{r_{0i}+d_{0i}}$, $c_i^* = \frac{\mathcal{E}[r_i]}{\mathcal{E}[r_i]+\mathcal{E}[d_i]}$ で与えられる．(赤司・川崎 (2011), 定理1)
- 説明変数$\mathbf{x}_i$の一部に(r_i, d_i)の関数が含まれることを認めても，依然最適閾値は$c_i^* = \frac{r_i}{r_i+d_i}$で与えられる．(赤司・川崎 (2011), 定理2)

すなわち現実的な設定下においては，最適な意思決定ルールz_i^*は

$$z_i^* = \mathbb{1}\{\ r_i[1 - F(\boldsymbol{\beta}'\mathbf{x}_i)] \geq d_i F(\boldsymbol{\beta}'\mathbf{x}_i)\ \} = \mathbb{1}\{\ F(\boldsymbol{\beta}'\mathbf{x}_i) \leq \frac{r_i}{r_i + d_i}\ \}$$

で与えられる．また，総利潤$\sum_{i=1}^N \pi(z_i^*)$の分布に関しても，事故率や保険料，損害賠償額に関する幾つかの仮定のもとで，それが正規分布で近似できること，利用可能なデータを使って構成する予想区間の中では, それが最小予想区間となることが示される．(赤司・川崎 (2011), 定理3)

3. β の推定と漸近論

以上の議論では，未知係数ベクトル$\boldsymbol{\beta}$の真値は所与．　ここからは$\boldsymbol{\beta}$の推定値をプラグインすることの影響を考慮した結果を与える．例えば$(y_j, \mathbf{x}_j)_{j=1}^{N_1}$を前期のデータとして係数の推定に用い, 今期の総利潤$\sum_{i=1}^{N_2} \pi(z_i^*)$の予測を行う，というような場合．実際の判断は推定量$\hat{\boldsymbol{\beta}}$を用いた次の2項予測に基づくことになる．特に, 下の第1式右辺のc_iを$c_i^* = r_i/(r_i + d_i)$で与えるとき, 左辺を$\hat{z}_i^*$と表す.

$$\hat{z}_i = \mathbb{1}\{\ F(\hat{\boldsymbol{\beta}}'\mathbf{x}_i) \leq c_i\ \} \quad \pi(\hat{z}_i) = r_i(1 - y_i)\hat{z}_i - d_i y_i \hat{z}_i$$

定理3の仮定に加え，予測評価標本での期待値の存在と，データが閾値周辺にばかり集中する病理的ケースの排除を仮定すると，以下が成り立つ．(赤司・川崎 (2011), 定理4)

- $\hat{z}_i^*$に基づく予測は，確率収束の意味で総利潤を最大化する.

$$\frac{\sum_{i=1}^{N_2} \pi(\hat{z}_i)}{\sum_{i=1}^{N_2} \pi(\hat{z}_i^*)} \xrightarrow{p} \lim_{N_2 \to \infty} \frac{\sum_{i=1}^{N_2} \mathcal{E}[\pi(z_i)]}{\sum_{i=1}^{N_2} \mathcal{E}[\pi(z_i^*)]} \leq 1$$

- $N_2/N_1 \to \eta$ $(0 \leq \eta < \infty)$とすると，$\sum_{i=1}^{N_2} \pi(z_i^*)$に対する漸近的予想区間を，依然として正規分布で構成できる．ただし，その予想区間幅を与える誤差分散は，定理3で導かれる形にプラグインの効果が上乗せされて，一般には広くなる.

4. 数値実験

- $N_1 = N_2 = 10000$としてロジットモデル$F(\boldsymbol{\beta}'\mathbf{x}_i) = \frac{\exp(\beta_1+\beta_2 x_i)}{1+\exp(\beta_1+\beta_2 x_i)}$ からデータ生成．$\beta_2 = -1$, x_iは自由度15のχ^2分布から生成.
- 事故率$(1/N_2)\sum_{i=1}^{N_2} y_i$は$\beta_1$を調整し，1% $(\beta_1 = 4.2)$と3% $(\beta_1 = 5.8)$の2通りを考慮. (一般に, 債務不履行や保険の支払いは稀な事象．)
- 損失は, i $(i = 1, ..., N_2)$に対して共通で$r_i = 0.1$と$d_i = 2.0$とする.
- 利得とこのとき最適な閾値は, $c_i^* = 0.1/(0.1 + 2.0) \simeq 0.05$.

考察

- 最大化総利潤(c_i^*の場合)の予想区間に関る結果が下の表．丸括弧内の数値は真値$\boldsymbol{\beta} = (\beta_1, \beta_2)'$を所与とした結果.
- 95%予想区間は事故率が低いほど狭く, 事故率が低い方がその水準自体も高くなっている.
- 括弧内は真値$\boldsymbol{\beta}$を所与とした結果なので, 定理4に従い推定値を用いるといずれの事故率の下でも約30%予想区間の幅を広げなければならない.
- 繰り返し数3000のうち予想区間に入った割合(%)を示す被覆係数については, プラグイン効果を無視すると予想区間が狭くなり, 実際には80%台(表中未報告)に.
- プラグイン効果を追加することで, 名目の95%被覆係数に対して表の実現値は95%近くに．漸近論の近似は有限標本の下でも良好.

	事故率1%	事故率3%
95%予想区間	[885, 934]　([890, 927])	[799, 855]　([805, 849])
95%被覆係数	94.9 (94.7)	94.5 (94.5)

なお，赤司・川崎(2011)では，南ドイツ銀行の個人向けローンのデータを用いた実証分析で, 総利潤の実績値を数%上回る結果が2項予測の閾値の最適化によって得られることを示している.

参考文献

赤司健太郎，川崎能典 (2011)．2項モデルの予測による金融リスク最小化：理論と応用，統計数理, Vol. 59, No. 1, 29–40.