経済データ分析における多重比較の問題と **knockoff** 法

植松　良公

2020 年 7 月 29 日

東北大学大学院経済学研究科

はじめに

回帰分析と t 検定に関する注意点について，その現代的な解決方法を含め概観する.

1. 線形回帰モデルと最小 2 乗法
2. 各係数の有意性検定と検定の誤り
3. 多重検定（多重比較）の問題
4. Family-Wise Error Rate (FWER) のコントロール
    - Bonferroni 法
5. False Discovery Rate (FDR) のコントロール
    - Benjamini-Hochberg 法
    - Knockoff 法

線形回帰モデルと最小 2 乗法

線形回帰モデルを考える：

$$y = x_1\beta_1 + \cdots + x_k\beta_k + u.$$

データセット $\{y_i, x_{i1}, \ldots, x_{ik}\}_{i=1}^{n}$ が得られたとき，このモデルの係数の最小 2 乗推定量は，

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\mathbf{y}.$$

ただし

$$\mathbf{X} = \begin{pmatrix} x_{11} & \ldots & x_{1k} \\ \vdots & & \vdots \\ x_{n1} & \ldots & x_{nk} \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}.$$

有意性検定

係数 β_j の有意性検定を考える．仮説は，

$$H_0 : \beta_j = 0 \quad v.s. \quad H_1 : \beta_j \neq 0.$$

係数 β_j の t 検定統計量は，

$$T_j = \frac{\hat{\beta}_j}{\hat{s}_j}.$$

ただし $\hat{s}_j^2$ は $\hat{\mathbf{S}} = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$ の (j, j) 要素で $\hat{\sigma}^2$ は誤差項 u の分散の推定量．

サンプルサイズ n が十分に大きいとき，中心極限定理より，H_0 の下での t 統計量 T_n は漸近的に標準正規分布に従う：

$$T_j \overset{a}{\sim} N(0, 1).$$

有意性検定

漸近正規性に基づいた検定を考える．

有意水準を α とすると，帰無分布 $N(0,1)$ の（両側）棄却域は $C = (-\infty, -\bar{t}_\alpha] \cup [\bar{t}_\alpha, \infty)$ と書ける．データから T_j の実現値 t_j を計算し，以下のいずれかを得る：

- $t_j \in C$ のとき，H_0 を棄却し H_1 を採択する．このとき，β_j は有意に 0 とは異なるという．
- $t_j \notin C$ のとき，H_0 を採択する．このとき，統計学的には $\beta_j = 0$ を否定できない．これを「有意ではない」という．

有意性検定

2 通りの t 検定の使われ方：

(A) 事前に検定したい仮説を考慮してモデルを作り，興味ある係数の t 検定を考える．

- 教育年数 x_1 が賃金 y に与える影響を調べたい．教育年数以外の「個人の能力」を $x_2,\ldots,x_k$ として回帰モデル $y=\alpha+\beta_1x_1+\beta_2x_2+\cdots+x_k\beta_k+u$ を立て，β_1 の t 検定を行う．

(B) 回帰モデルを推定した後，t 検定により事後的に有意な変数を選ぶ．

- 多くの経済指標 $x_1,\ldots,x_k$ の中から，経済成長 y の要因となるものを t 検定で探す．
- 多くの投資戦略 $x_1,\ldots,x_k$ の中から，ベンチマーク y を上回るものを t 検定で選ぶ．
- ある財の価格 y の決定要因を，関連のありそうな $x_1,\ldots,x_k$ の中から t 検定で探す．

検定の誤り

仮説検定には 2 つの誤りがある：

- 第一種の過誤：正しい H_0 を棄却する誤り．検定の定義より，第一種の過誤の確率 α は有意水準であり，分析者がコントロールできる．
- 第二種の過誤：正しくない H_0 を棄却できない誤り．第二種の過誤の確率を β としたとき，$1-\beta$ を検出力と呼ぶ：

$$\begin{aligned}1-\beta &= 1-\mathbb{P}\left(H_0 \text{ を採択} \mid H_1 \text{ は真}\right)\\ &= \mathbb{P}\left(H_0 \text{ を棄却} \mid H_1 \text{ は真}\right).\end{aligned}$$

(B) の場合，第一種の過誤が積み上がる．これを多重検定の問題という．

多重検定

複数の仮説検定を考える：

$$H_0^j : \beta_j = 0 \quad v.s. \quad H_1^j : \beta_j \neq 0, \quad j = 1, \ldots, k.$$

各仮説の t 検定を $j = 1, \ldots, k$ と繰り返すとき，以下の表を得る：

	H_0^j は真	H_0^j は偽	合計
H_0^j を棄却	E_1	$s - E_2$	
H_0^j を採択	$p - s - E_1$	E_2	
合計	$p - s$	s	p

- E_1 = 選ばれた「不要な」変数の数 = 第一種の過誤の数.
- E_2 = 選ばれなかった「重要な」変数の数 = 第二種の過誤の数.

FWER と Bonferroni 法

FWER (Family-Wise Error Rate) とは，「少なくとも 1 回は正しい H_0 を棄却してしまう」確率のこと．つまり

$$
\begin{aligned}
\mathbf{FWER} &= \mathbb{P}(E_1 \geq 1) \\
&= 1 - \mathbb{P}(E_1 = 0) \\
&= 1 - \mathbb{P}\left(\text{正しい } H_0^j \text{ がすべて採択される}\right) \\
&= 1 - (1-\alpha)^{k-s} \quad \text{(独立ならば)} \\
&\leq 1 - (1-\alpha)^k.
\end{aligned}
$$

これは $\alpha = 0.05, k = 20$ のとき，$1 - (1-\alpha)^k \approx 0.64$.

FWER を α 以下にコントロールするためには，個々の仮説検定における有意水準を α/k とすればよいことが知られている．これをBonferroni の方法 (Dunn, 1961) という．

FWER と Bonferroni 法

Bonferroni 法の問題点：

- 第一種の過誤の指標として FWER は非常に保守的.
- 特に k が大きいとき，個々の H_0^j は棄却されにくくなる.
- その結果，検出力が上がらない.

多重検定における検出力：

$$\text{Power} = \mathbb{E}\left[\frac{s - E_2}{s}\right] = \mathbb{E}\left[\frac{\text{選ばれた「重要な」変数の数}}{\text{「重要な」変数の数}}\right].$$

（一般に，第一種の過誤を低くコントロールしつつ，検出力が高くなる手法が望ましい.）

FDR

Benjamini & Hochberg (1995) の提案：

第一種の過誤の指標として，FWER の代わりに以下のFDR（False Discovery Rate, 偽発見率）をコントロール目標とする：

$$\mathrm{FDR} = \mathbb{E}[\mathrm{FDP}].$$

ただしFDP (False Discovery Proportion) とは,

$$\mathrm{FDP} = \frac{E_1}{s+E_1-E_2} = \frac{\text{選ばれた「不要な」変数の数}}{\text{選ばれた変数の数}}.$$

Benjamini & Hochberg (1995) では，FDR を q 以下にコントロールする変数選択法（BH 法）を提案している.

BH 法

BH 法：

1. k 個の帰無仮説 $H_0^1,\dots,H_0^k$ に対応する p 値を $p_1,\dots,p_k$ とする．これらを小さい順に並べたものを $p_{(1)} \le \cdots \le p_{(k)}$ とする．
2. $p_{(j)} \le q \times j/k$ を満たす全ての $j = 1,\dots,k$ について，対応する $H_0^{(j)}$ を棄却（$x_{(j)}$ を選択）する．

いくつかの仮定のもと，BH 法で選ばれた変数たちの FDR は q 以下にコントロールされる．

一方で k が大きいとき（特に $k > n$ のとき），BH 法はうまく機能しないことが知られている．

Knockoff 法

最近では，変数の数 k がサンプルサイズ n よりも大きい場合（高次元）でも回帰分析を求められる場合が多々ある．

- そもそも $k > n$ のとき最小 2 乗法は機能しない．（代わりにLassoや Ridge 回帰が用いられるが，その場合の p 値の計算は安定しない．よって BH 法を用いるのは難しい．）

Barber & Candès (2015), Candes, Janson, Fan and Lv (2018) では，Knockoff 法という新しい FDR コントロール手法を提案した．これは p 値によらない手法で，高次元でも適用できる．

Knockoff 法

Knockoff 法：

1. データセット $\{y_i, x_{i1}, \dots, x_{ik}\}_{i=1}^n$ を用いて，「$\mathbf{X}$ に似ているが $\mathbf{y}$ とは独立な変数行列 $\tilde{\mathbf{X}}$」を作る（Knockoff 行列）.
2. $\mathbf{y}$ を $(\mathbf{X}, \tilde{\mathbf{X}})$ に回帰する.（$k > n$ のときは Lasso を用いる.）
3. 得られた係数の推定値 $\hat{\boldsymbol{\beta}}' = (\hat{\boldsymbol{\beta}}_1', \hat{\boldsymbol{\beta}}_2')'$ を用いて，$W_j = |\hat{\beta}_j| - |\hat{\beta}_{p+j}|$ を計算する.
4. ある方法で閾値 $\tau > 0$ を決め，$W_j > \tau$ を満たす j について H_0^j を棄却（x_j を選択）する.

いくつかの仮定のもと，Knockoff 法で選ばれた変数たちの FDR は q 以下にコントロールされる.

一般的に Knockoff $\tilde{\mathbf{X}}$ を作るのが難しい.

Fan, Lv, Sharifvaghefi and Uematsu (2020) では，データがファクターモデル $\mathbf{X} = \mathbf{F}\mathbf{B}' + \mathbf{E}$ に従うという仮定の下で，Knockoff $\tilde{\mathbf{X}}$ を生成する方法を提案した.

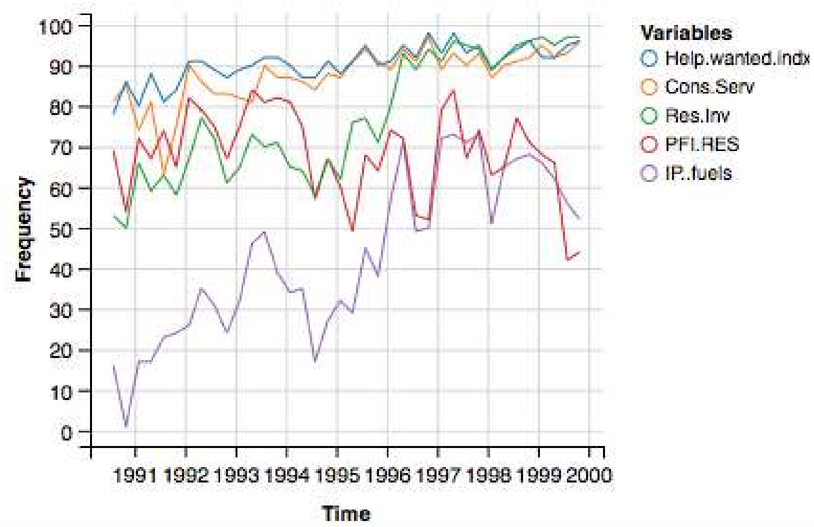

Figure 1. 1990–1999

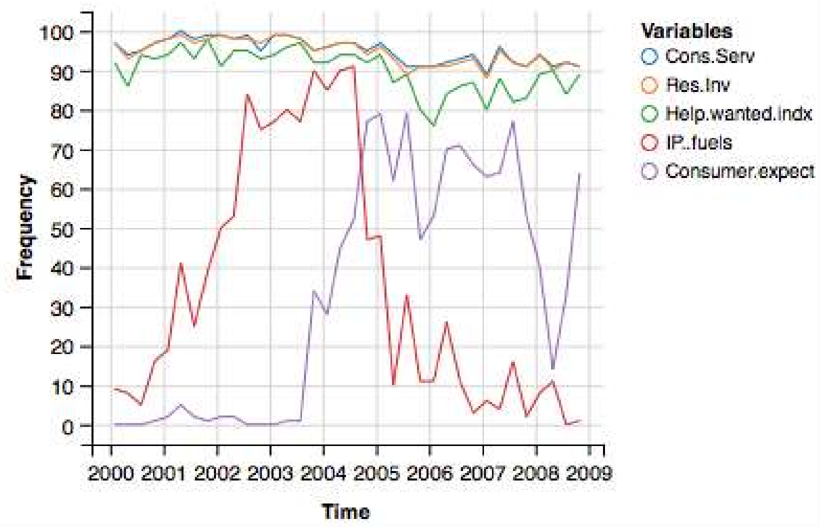

Figure 2. 2000-2008

参考文献

- Dunn, O. J. (1961). “Multiple comparisons among means.” *Journal of the American Statistical Association*, **56**, 52–64.
- Benjamini, Y. and Y. Hochberg. (1995). “Controlling the false discovery rate: a practical and powerful approach to multiple testing.” *Journal of the Royal Statistical Society Series B*, **57**, 289–300.
- Barber, R. F. and E. Candès (2015). “Controlling the false discovery rate via knockoffs.” *Annals of Statistics*, **43**, 2055–2085.
- Candès, E., L. Janson, Y. Fan and J. Lv (2018). “Panning for gold: model-X knockoffs for high-dimensional controlled variable selection.” *Journal of the Royal Statistical Society Series B*, **80**, 551–577.
- Fan, Y., J. Lv, M. Sharifvaghefi and Y. Uematsu (2020). “IPAD: stable interpretable forecasting with knockoffs inference” *Journal of the American Statistical Association*, forthcoming.