分散関数回帰分析（Variance Function Regression）

麦山 亮太（東京大学大学院人文社会系研究科博士課程）

【文献】Western, Bruce and Deirdre Bloome. 2009. “Variance Function Regressions for Studying Inequality.” *Sociological Methodology* 39(1): 293–326.

要約

1 Introduction

不平等を研究する際には、集団間の違い（differences between groups）を集団内の違い（differences within groups）から分離することが目指される。社会学の理論も、集団間の不平等に関する仮説を導くものが多い。しかし、集団内の分散はしばしば集団間の分散よりも大きく、社会学の分析にとっても重要な意味を持っている。

本稿は、分散関数回帰分析（variance function regression、以下 VFR）という、共変量が集団内の不平等と集団間の不平等に与える効果を推定する統計モデルを提示する。

2 Between-group and within-group inequality in sociology

単純な OLS 回帰分析を考える。

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

ここで、従属変数 y の分散 $V(y_i)$ は、集団間の分散 $V(\hat{y}_i)$ と、残差＝集団内の分散 $V(\hat{e}_i)$ に分解される。

伝統的に $V(\hat{e}_i)$ は、本質的に重要でないが、除外変数によって説明されるであろう部分として位置づけられてきた。しかし近年、とりわけ不平等研究の文脈では、$V(\hat{e}_i)$ は観察されないスキルや運として解釈され、これらが不平等の拡大に寄与しているという議論がなされている。

社会学的には、集団内の不平等の拡大は最低賃金や労働組合、大企業のキャリアラダーといった、所得を安定化させ、市場の力から賃金を守る役割を果たしてきた諸制度が縮小してきた結果として、集団内の不平等が拡大してきたと解釈できる。このような労働市場の脱制度化 deinstitutionalization に関する議論は、とくに教育水準の低い労働者の不安定性 insecurity を高め、賃金の不平等を拡大することを予測している。

このように、近年は社会学においても、こうした残差を明示的に説明の対象とする研究が現れてきている。しかしこれらの研究は、(1) ミクロ水準の変数を用いた回帰分析を行い、(2) 得られた残差をマクロ水準の変数で回帰する、という（アド・ホックな）手順で分析を行っている。

これに対して VFR は、集団内分散と集団間分散を同時推定し、ミクロ変数、マクロ変数のいずれも集団内の不平等に影響すること、および変数が集団内分散・集団間分散の両者に影響を与えることを許容する。

3 Formalizing and estimating the model

VFR では、以下の 2 つの式を同時推定する。

$$\hat{y}_i = \boldsymbol{x}_i'\boldsymbol{\beta}$$
$$\log \sigma_i^2 = \boldsymbol{z}_i'\boldsymbol{\lambda}$$

$\boldsymbol{x}_i'$ と $\boldsymbol{z}_i'$ には同じ変数が含まれていてもよい（すべて同じであってもよい）。
係数 β_k は通常の回帰分析と同じく、独立変数 x_k1 単位の変化に対する y の平均的な変化量を意味する。そして係数 λ_j は、z_j1 単位の変化に対する対数分散の変化量 を意味する。

VFR は、homoskedastic な分散を仮定する通常の回帰分析と異なり、heteroskedastic な分散を許容し、かつ、分散が共変量と関連を持っていてもよい、とするモデルである。

3.1 Estimation

推定に当たっては、2 段階推定（two-stage）、最尤法（ML）、制限付き最尤法（REML）、ベイズ（Bayes）の 4 つの方法がある。もっともバイアスが大きいのは 2 段階推定。制限付き最尤法およびベイズはサンプルサイズが小さくてもよい推定値が得られやすい。実装がもっとも容易なのは最尤法。

最尤法の手続きは以下のようになる。

(1) y_i を x_i で回帰し、係数 $\hat{\boldsymbol{\beta}}$ および残差の予測値 $\hat{e}_i$ を得る。
(2) 残差の予測値の 2 乗を $\boldsymbol{z}_i$ で回帰（ガンマ回帰を用いる）し、係数 $\hat{\boldsymbol{\lambda}}$ および $\hat{\sigma}_i^2$ を得る。
(3) $1/\hat{\sigma}_i^2$ で重みづけしたうえでふたたび式 (1) を推定し、残差の予測値 $\hat{e}_i$ の値を更新する。
(4) 以上の手順を対数尤度が十分に高くなるまで繰り返し、最終的な係数 $\hat{\boldsymbol{\beta}}$、$\hat{\boldsymbol{\lambda}}$ を得る。

【補足】ガンマ回帰（gamma regression）

確率変数 y の取り得る範囲が 0 以上の連続確率分布のことをガンマ分布という。

$$p(y|s,r) = \frac{r^s}{\Gamma(s)} y^{s-1} \exp(-ry)$$

$$\text{ただし} \Gamma(s) = \int_0^\infty t^{s-1} \exp(-t) dt \quad \text{（ガンマ関数）}$$

ガンマ分布の平均は s/r、分散は s/r^2 であり、確率変数 y が常に 0 より大きく、かつ平均値の上昇にしたがって分散も増大していくような場合に適している。

4 Application I: Incarceration and earnings insecurity

投獄が賃金の平均および対数分散に与える効果を推定する。結果は以下のようになる。

	β	λ
Intercept	.085	-.149
Previously imprisoned	-.329	.435
Currently imprisoned	-.462	.178
Years of schooling	.038	-.107
Work experience	.010	-.017

出所）Western and Bloome (2009)、Table 3。従属変数は 1 年間の勤労所得の対数値。分析に際して観測値から個人内平均を減じている（固定効果モデル）。ベイズ推定の結果のみを提示。標準誤差は省略。

Previously imprisoned の係数 β より、投獄経験は他の変数を一定としたうえで賃金を約 28% ($= 1 - \exp(-.329)$) 引き下げる。加えて、係数 λ より、投獄経験を持つことは、他の変数を一定としたうえで賃金の対数分散を約 54%（$= \exp(.435)$）増加させる。結果の解釈としては、投獄経験を持つ場合は全体として賃金は低下するが、それに加えて、高いスキルを有する場合はより有利に、他方でスキルの低い場合は顕著に不利になるということを意味している。

5 Decomposing trends in inequality

対数分散 variance of the logarithms は、不平等の指標として好ましい性質を持っている（Allison, Paul D., 1978, “Measures of Inequality.” *American Sociological Review,* 43(6): 865–80. を参照のこと）

$y_i = \log Y_i$ とする。対数分散は、以下のように集団間分散と集団内分散に分解することができる。r_c は集団間分散 $\hat{y}_c - \overline{y}$、σ_c^2 はセル（集団）内分散を表す。

$$\begin{aligned} V &= B + W \\ &= \sum_{c=1}^{C} \pi_c r_c^2 + \sum_{c=1}^{C} \pi_c \sigma_c^2, \end{aligned}$$

時点 $t = 0, 1$ における 2 時点間の分散の変化は、以下の 3 つの要因（compositional effect,

between-group effect, within-group effect）に分解することができる。

$$V_1 - V_0 = \sum_{c=1}^{C}(\pi_{1c} - \pi_{0c})(r_{1c}^2 + \sigma_{1c}^2) + \sum_{c=1}^{C}\pi_{0c}(r_{1c}^2 - r_{0c}^2) + \sum_{c=1}^{C}\pi_{0c}(\sigma_{1c}^2 - \sigma_{0c}^2)$$

一般に、時点 $t = 0, \cdots, T$ における総分散は $V_t = \sum_{c=1}^{C}\pi_{tc}(r_{tc}^2 + \sigma_{tc}^2)$ と表すことができる。これを踏まえたうえで、以下の 3 つの counterfactual な変化をプロットする方法も有用である。

$$V_t^C = \sum_{c=1}^{C}\pi_{0c}(r_{tc}^2 + \sigma_{tc}^2) \quad \text{(composition fixed)}$$
$$V_t^B = \sum_{c=1}^{C}\pi_{tc}(r_{0c}^2 + \sigma_{tc}^2) \quad \text{(between variance fixed)}$$
$$V_t^W = \sum_{c=1}^{C}\pi_{tc}(r_{tc}^2 + \sigma_{0c}^2) \quad \text{(within variance fixed)}$$

VFR を用いた要因分解法は、通常の要因分解法とくらべて以下の 3 つの利点を持つ。第 1 に、個人レベルの共変量が集団間分散に与える影響と、集団内分散に与える影響とを区別できる。第 2 に、ある年で観察されていないセルがあっても分析ができる。第 3 に、ベイズ推定を用いた場合、パラメータの事後分布から変化の信頼区間を求めることができる（ということを多分言っている）。

6 Application Ⅱ: Decomposing trends in hourly wages

1970 年以降、賃金の不平等は拡大傾向にある。こうした不平等の拡大を、(1) 教育水準間の不平等、(2) 教育水準内の不平等、(3) 教育水準の構成の変化、という 3 つの要因に分解する。分析対象はフルタイムの男性労働者である。
時点 t における VFR を以下のように定義する。

$$\hat{y}_{ti} = \boldsymbol{x}_{ti}'\boldsymbol{\gamma}_t + \boldsymbol{e}_{ti}'\boldsymbol{\beta}_t$$
$$\log\sigma_{ti}^2 = \boldsymbol{x}_{ti}'\boldsymbol{\theta}_t + \boldsymbol{e}_{ti}'\boldsymbol{\lambda}_t$$

$\boldsymbol{x}_{ti}'$ は人種・エスニシティおよび就業年数を示すダミー変数群、$\boldsymbol{e}_{ti}'$ は 5 つの教育水準を表す 4 つのダミー変数群を意味する。以下の 4 つの統計量を計算する。
教育水準間の格差を 1970 年時点に固定

$$V_t^\beta = \sum_{c=1}^{C}\pi_{tc}\left[(\boldsymbol{x}_c'\boldsymbol{\gamma}_t + \boldsymbol{e}_c'\boldsymbol{\beta}_{1970} - \overline{y}_t)^2 + \sigma_{tc}^2\right]$$

教育水準内の格差を 1970 年時点に固定

$$V_t^\lambda = \sum_{c=1}^{C}\pi_{tc}\left[r_{tc}^2 + \exp(\boldsymbol{x}_c'\boldsymbol{\theta}_t + \boldsymbol{e}_c'\boldsymbol{\lambda}_{1970})\right]$$

教育水準間・教育水準内の格差を 1970 年時点に固定

$$V_t^{\beta\lambda} = \sum_{c=1}^{C} \pi_{tc} \left[(\boldsymbol{x}_c' \boldsymbol{\gamma}_t + \boldsymbol{e}_c' \boldsymbol{\beta}_{1970} - \overline{y}_t)^2 + \exp(\boldsymbol{x}_c' \boldsymbol{\theta}_t + \boldsymbol{e}_c' \boldsymbol{\lambda}_{1970}) \right]$$

教育水準の構成（周辺分布）を 1970 年時点に固定

$$V_t^{\pi} = \sum_{c=1}^{C} \frac{p_{1970c}}{p_{tc}} \pi_{tc} [r_{tc}^2 + \sigma_{tc}^2]$$

各年の値をプロットする他、任意の 2 時点を選び、各要因がどの程度不平等の拡大に寄与したのかを計算することも有用である。計算の結果は以下の表のようになる。今回の場合、教育水準間 β の変化（between education）が不平等の変化のうちの 27.2% を、教育水準内 λ の変化が不平等の変化のうちの 25.3% を説明することが分かる（計算の順序を逆にすると寄与度の値は多少異なると思われる。しかし、β と λ のいずれも同程度に不平等の拡大に寄与しているということは変わらないだろう）。

	2005	Change from 1970 to 2005	Percentage of Change Explained
Observed Variance	.481	.179	—
Adjusted variance, fixing at 1970:			
Education effects, $\boldsymbol{\beta}$	.432	.131	27.2
Education effects, $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$	.387	.085	52.5
Educational attainment	.500	.199	-10.8
All within-group effects, $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$	.371	.069	61.4

出所）Western and Bloome (2009)、Table 4。標準誤差は省略。

7 Discussion

VFR は、2 つの式を用いているため、モデルの特定化はより注意深く行う必要がある。

- 平均値に関するモデルを適切に特定化できていない場合：通常の回帰分析と同様、係数 β の推定値にバイアスが生じる。加えて、残差 σ^2 は真の残差とはならないため、たとえ分散に関するモデルを適切に特定できたとしても、係数 λ の推定値にバイアスが生じる。
- 分散に関するモデルを適切に特定化できていない場合：平均値に関するモデルの標準誤差の大きさにバイアスが生じるが、こちらのモデルが適切に特定化されていれば、係数 β は不偏推定量となる。

通常の回帰分析は集団間の不平等のみを問題とするが、集団内の不平等もまた社会学的に重要な意味を持っている。加えて、全体の不平等を問題とする場合、集団内の不平等も重要な部分を占めており、それ自体も説明されるべき対象であることを最後に強調している。

実装

```
/*車のデータを利用します。*/
use http://www.stata-press.com/data/r13/auto ,clear

/*local 変数を定義します。x は式（1）の、z は式（2）の独立変数にそれぞれ対応しています。*/
local x = "weight mpg foreign"
local z = "weight mpg foreign"
/*price を x で回帰し、係数（beta）および残差 e の予測値を得ます。さらに残差を 2 乗して残差平方和を得ます。*/
reg price 'x'
predict R, r
gen R2=R^2

/*e の 2 乗をガンマ回帰し、係数（gamma）および sigma の 2 乗の予測値を得ます。その際、リンク関数として log を指定します。*/
glm R2 'z', family(gamma) link(log)
predict S2, mu

/*尤度関数を構築し、これを反復計算の初期値とします。*/
gen LOGLIK = -(1/2)*(ln(S2)+(R2/S2))
egen LL0 = sum(LOGLIK)
display LL0

/*以下、収束するまで計算を繰り返します。*/
gen DLL=1
while DLL > .00001 {
drop R
quietly reg price 'x' [aw=1/S2]
drop S2
predict R, r
replace R2=R^2
est store BETA
quietly glm R2 'z', family(gamma) link(log)
predict S2, mu
est store LAMBDA
replace LOGLIK = -(1/2)*(ln(S2)+(R2/S2))
egen LLN = sum(LOGLIK)
display LLN
replace DLL=LLN-LL0
replace LL0=LLN
drop LLN
}

/*無事収束したら、係数を確認します。*/
est tab BETA LAMBDA, b se stat(N r2)
```

参考文献

[1] 小川和孝，2016，「社会的属性と収入の不安定性」『理論と方法』31(1): 39–51.

[2] 瀧川裕貴，2013，「現代日本における所得の不平等――要因の多次元性に注目して」佐藤嘉倫・木村敏明編『不平等生成メカニズムの解明――格差・階層・公正』ミネルヴァ書房.

[3] Western, Bruce, Deirdre Bloome, and Christine Percheski, 2008, "Inequality among American Family with Children, 1975 to 2005," *American Sociological Review,* 73(6): 903–20.

[4] Western, Bruce and Jake Rosenfeld, 2011, "Unions, Norms, and the Rise in U.S. Wage Inequality," *American Sociological Review,* 76(4): 513–37.

[5] Williams, Mark, 2013, "Occupations and British Wage Inequality, 1970s-2000s," *European Sociological Review,* 29(4): 841–57.

[6] Zheng, Hui, Yang Yang, and Kenneth C. Land, 2011, "Variance Function Regression in Hierarchical Age-Period-Cohort Models: Applications to the Study of Self-Reported Health," *American sociological review,* 76(6): 955–83.