

# 第19回 共分散構造分析（後半）

## 1. 概要と例題

今回は共分散構造分析のワンポイントアドバイスや注意点を紹介します。

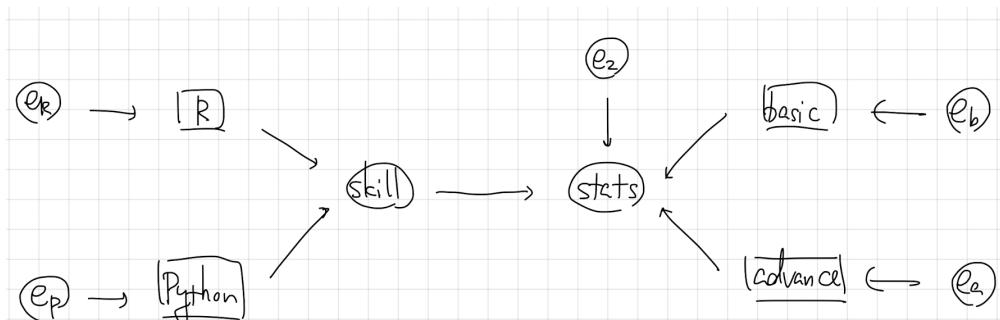
1. 共分散構造分析に現れる変数は必ずしもデータに記録されている変数（観測変数）だけとは限りません。探索的因子分析に現れる共通因子のように、観測変数の共通性を意味する変数を作ることができます。これを**潜在変数**といいます。そこで、潜在変数を含む共分散構造分析のやり方を紹介します。
2. どんな構造方程式に対しても共分散構造分析の結果が得られるとは限りません。パラメータが推定可能かを調べる一つの方法である**自由度**を紹介します。
3. 共分散構造分析によって得られたモデルがどれだけデータに当てはまっているかを調べる方法として**適合度指標**を紹介します。

**例題** S大学のデータサイエンス学部では、プログラミング言語系の講義として「基礎からのR」「基礎からのPython」、統計学の講義として「初級統計学」「続初級統計学」が開講されています。I教授はプログラミングのスキルが統計学の理解度にどれだけ影響があるのかを調べようと考えています。以下の間に答えてください。

- [1] データに観測されている変数を**観測変数**、観測されていない変数を**潜在変数**といいます。潜在変数「プログラミングのスキル」「統計学の理解度」および観測変数「基礎からのR」「基礎からのPython」「初級統計学」「続初級統計学」の間の関係を表すパス図を示してください。
- [2] 潜在変数「プログラミングのスキル」「統計学の理解度」を定義する測定方程式と、これらの関係を示す構造方程式を述べてください。また測定方程式と構造方程式のパラメータ（外生変数の分散、パス係数と誤差の分散）をすべて挙げてください。
- [3] このモデルの自由度を述べよ。
- [4] 今回、25名の学生について4科目の成績を `data` ディレクトリに `skill_stats.csv` という名前のcsvファイルに記録しました。このファイルとR言語の `lavaan` パッケージにある `sem` 関数を用いて測定方程式と構造方程式のパラメータを推定してください。
- [5] モデルがデータにどれだけあてはまっているかを確認する方法に、**適合度指標**があります。適合度指標の例にGFIがあります。GFIの最小二乗法の場合の計算方法を説明してください。

## 2. 潜在変数を含む場合のパス図

潜在変数はパス図では、丸印で記されます。また潜在変数はデータに記録されていないので、観測変数を説明する関係を定義することで測定します。例題の[1]の場合は、次のような解答になります。



### 3. 測定方程式

潜在変数が観測変数を説明する式を**測定方程式**といいます。例えば「プログラミングのスキル」（以下、skillと略）という潜在変数は2つの観測変数「基礎からのR」（以下、Rと略）と「基礎からのPython」（以下、Pythonと略）に対して、

$$\begin{aligned} R &= w_{R0} + w_{R1} \times \text{skill} + e_R \\ \text{Python} &= w_{P0} + w_{P1} \times \text{skill} + e_P \end{aligned}$$

という方程式を測定方程式として準備します。ここで注意してほしいのは、このままだと潜在変数skillの平均と分散が一意に決まらないことです。そこで、一般的には潜在変数の平均を0とします。また分散を決めるために、 $w_{r1}$ か $w_{p1}$ のいずれかの係数を1に固定します。

以上の議論から、測定方程式は次のようにになります。なお、以下では「統計学の理解度」をstats、「初級統計学」をbasic、「続初級統計学」をadvanceと略すことにします。

$$\begin{aligned} R &= w_{R0} + \text{skill} + e_R \\ \text{Python} &= w_{P0} + w_{P1} \times \text{skill} + e_P \\ \text{basic} &= w_{b0} \times \text{stats} + e_b \\ \text{advance} &= w_{a0} + w_{a1} \times \text{stats} + e_a \end{aligned}$$

なお、誤差 $e_R$ の分散を $\sigma_R^2$ と表すことにし、その他の誤差の分散も同様の記号を用いることにします。またskillの分散を $\sigma_1^2$ とします。構造方程式は、

$$\text{stats} = w_{21} \times \text{skill} + e_2$$

です。なお、誤差 $e_s$ の分散を $\sigma_2^2$ と表すことにします。切片項がないのは潜在変数の平均も誤差の平均も0と仮定することから、切片項が必ず0になることが従うためです。

また、測定方程式と構造方程式のパラメータは

- パス係数： $w_{P1}, w_{a1}, w_{21}$
- 誤差の分散： $\sigma_R^2, \sigma_P^2, \sigma_b^2, \sigma_a^2, \sigma_1^2, \sigma_2^2$

の9個あります。

### 3. 自由度

分散共分散行列の対角成分と上三角部分の要素の個数に比べて、推定するべきパラメータの個数が多いと、パラメータの値を推定することができなくなります。変数の個数を $d$ と表すとき、分散共分散行列の対角成分と上三角部分の要素の個数は $\frac{1}{2}d(d+1)$ です。推定するべきパラメータの個数を $p$ と表すとき、

$$df = \frac{1}{2}d(d+1) - p$$

をモデルの**自由度**といいます。言い換えれば自由度が負ならば、パラメータが推定できません。

### 4. R言語を用いた共分散構造分析の計算

#### 4.1 lavaanパッケージの読み込み

Hide

```
# パッケージの読み込み
library(lavaan)
library(semPlot)
```

## 4.2 データの読み込み

今回、学生の4科目の成績を `skill_stats.csv` という名前のcsvファイルに記録しました。このファイルを読み込んで、先頭5行を確認してみましょう。

Hide

```
# データの読み込み
dat <- read.csv("./data/skill_stats.csv", fileEncoding = "utf-8")
head(dat, n = 5)
```

	r <dbl>	python <dbl>	basic <dbl>	advance <dbl>
1	3.2	4.8	3.2	4.0
2	3.5	3.7	2.9	2.5
3	6.1	4.9	5.5	6.6
4	3.5	3.5	2.3	3.7
5	6.7	5.5	5.4	5.5

5 rows

データの分散共分散行列も確認しておきましょう。

Hide

```
# データの分散共分散行列
cov(dat)
```

```
      r     python    basic  advance
r  2.109167 1.2765833 1.2381667 1.796667
python 1.276583 1.2922333 0.9079667 1.210917
basic  1.238167 0.9079667 1.2446000 1.361000
advance 1.796667 1.2109167 1.3610000 2.745000
```

## 4.3 構造方程式のコーディング

`sem` 関数には構造方程式を渡す必要があります。相関関係は `~`、回帰の関係は `~` を用いて表します。今回の構造方程式は次のようにコーディングできます。

Hide

```
model <- "
  skill =~ 1*r + python
  stats =~ 1*basic + advance

  stats ~ skill
"
```

## 4.4 構造方程式のパラメータの推定

`sem` 関数にデータ `dat` と構造方程式のコード `model` を渡すと、構造方程式のパラメータが計算できます。なお、`sem` 関数では最尤法を用いています。

[Hide](#)

```
result <- sem(model = model, data = dat)
summary(result)
```

lavaan 0.6-9 ended normally after 26 iterations

Estimator	ML
Optimization method	NLMINB
Number of model parameters	9
Number of observations	25

Model Test User Model:

Test statistic	0.373
Degrees of freedom	1
P-value (Chi-square)	0.541

Parameter Estimates:

Standard errors	Standard
Information	Expected
Information saturated (h1) model	Structured

Latent Variables:

	Estimate	Std.Err	z-value	P(> z )
skill =~				
r	1.000			
python	0.709	0.125	5.666	0.000
stats =~				
basic	1.000			
advance	1.417	0.268	5.291	0.000

Regressions:

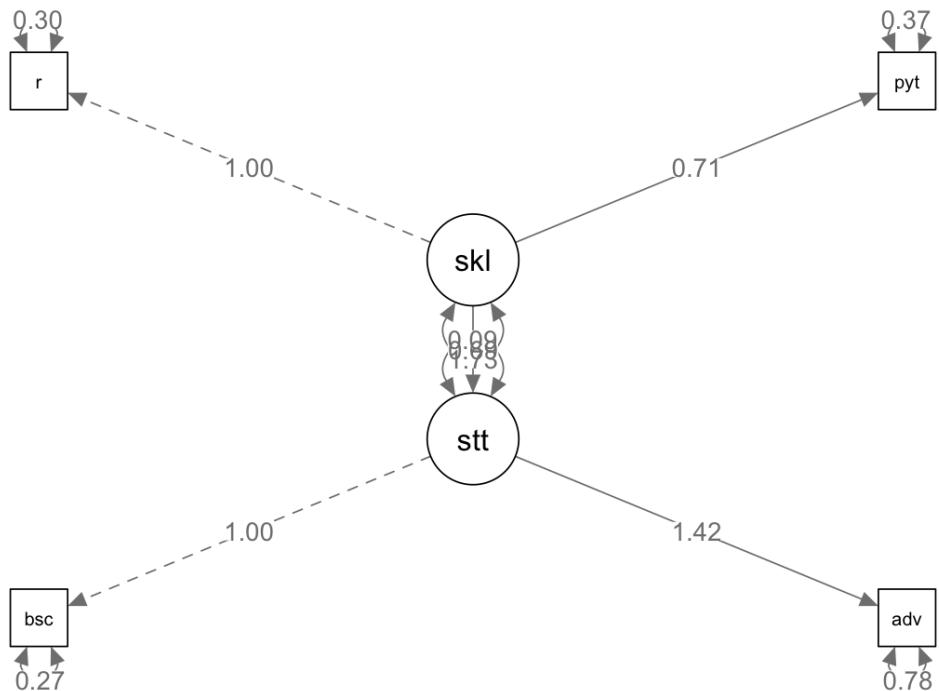
	Estimate	Std.Err	z-value	P(> z )
stats ~				
skill	0.694	0.126	5.513	0.000

Variances:

	Estimate	Std.Err	z-value	P(> z )
.r	0.295	0.193	1.531	0.126
.python	0.372	0.137	2.721	0.006
.basic	0.273	0.127	2.147	0.032
.advance	0.784	0.300	2.611	0.009
skill	1.730	0.593	2.918	0.004
.stats	0.089	0.121	0.733	0.464

[Hide](#)

```
semPaths(result, "model", "est", sizeMan = 5, edge.label.cex = 1.0)
```



## 4.5 適合度指標の計算

Hide

```
fitmeasures(result, fit.measures = "gfi")
```

*gfi*  
0.993

## 5. 適合度指標

定義した測定方程式と構造方程式の妥当性を検討するヒントとして、適合度指標とよばれるものを参考にすることがあります。適合度指標の代表例はGFI (Goodness of Fit Index) で、最小二乗法によって得られるモデルの場合、これは

$$GFI = 1 - \frac{\text{標本分散共分散とモデルの分散共分散の要素の差の2乗和}}{\text{標本分散共分散の要素の2乗和}}$$

で定義されます。決定係数とよく似た定義です。0以上1以下の値をとり、1に近いほど標本によくあてはまっているモデルと考えることができます。他にも、似たような定義のもとにSRMR (Standardized Root Mean Squared Error) などがあります。