データマイニング

第12回　共分散構造分析

2023年春学期

宮津和弘

本日の講義・演習

日付	講義・演習内容
04/14/23	(1) イントロダクション
04/21/23	(2) ビジネスシミュレーション
04/28/23	(3) ID-POSデータ分析
05/12/23	(4) 対応分析
05/19/23	(5) クラスター分析
05/26/23	(6) 自己組織化マップ
06/02/23	(7) 線形判別分析
06/09/23	(8) 非線形判別分析
06/16/23	(9) ツリーモデル
06/23/23	(10) 集団学習
06/30/23	休講（※黒門祭のため）
07/04/23	(11) サポートベクターマシン
07/14/23	**(12) 共分散構造分析**
07/21/23	(13) テキスト分析
07/28/23	(14) まとめ（ポートフォリオ）

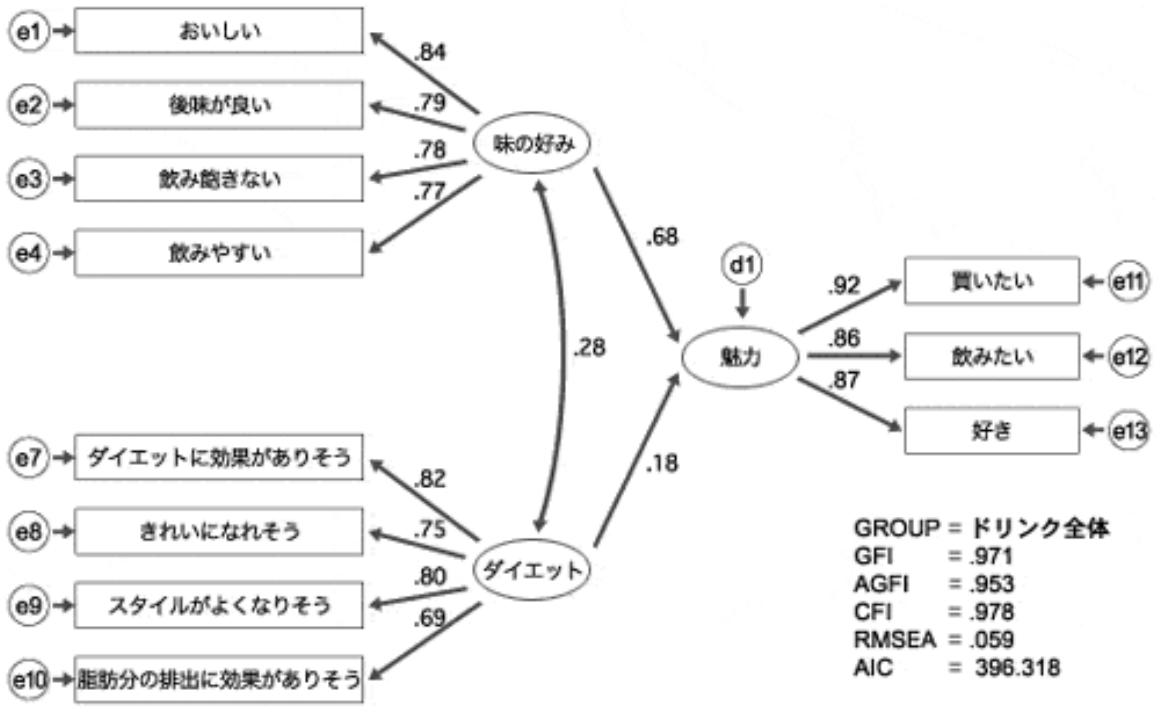
本日の演習概要とポイント

- **因子分析**

  → 主要５教科テスト点数のデータ

- **共分散構造分析**

  → 顧客満足度調査データ



（復習）因子分析

観測されない潜在因子

例）主要５教科テスト点数の背後に理系および文系の能力因子を仮定すると …

	数学	理科	英語	国語	社会
A	89	91	67	46	53
B	57	69	80	85	91
C	80	93	35	41	51
D	41	61	53	45	55
E	78	87	47	51	63
F	53	66	81	73	86
G	90	86	89	91	98

理系因子	文系因子
数理:平均	英国社:平均
90.0	53.3
63.0	85.3
86.5	42.3
51.0	37.7
82.5	53.7
59.5	80.0
88.0	92.7

因子分析モデル

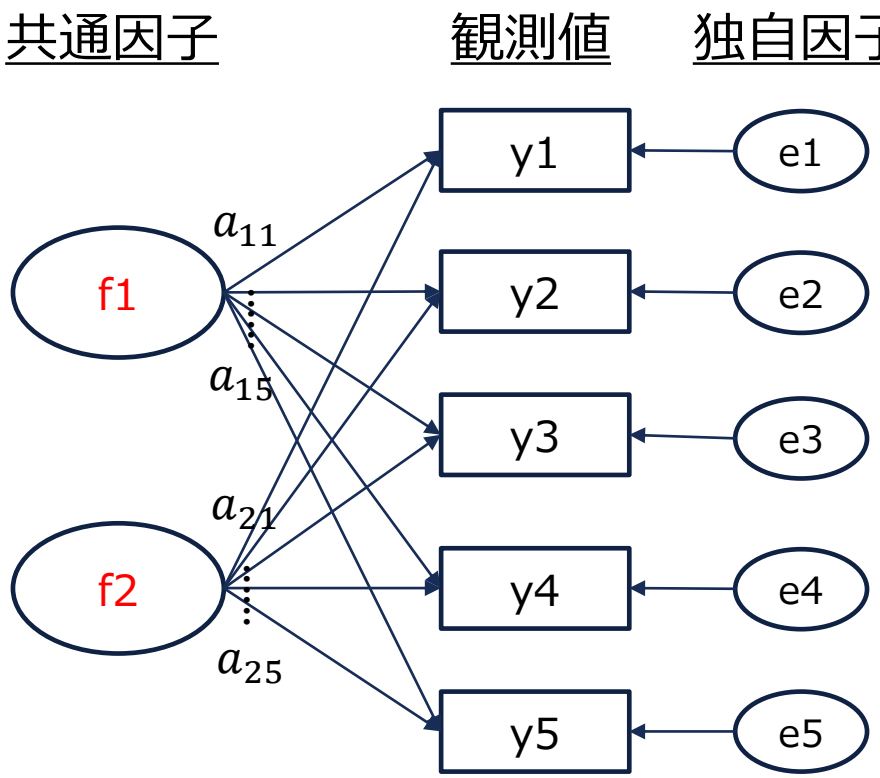
観測値の背後に共通因子と独自因子を仮定して、次元を減少する

$$\begin{pmatrix} y_{i,1} \\ y_{i,2} \\ y_{i,3} \\ y_{i,4} \\ y_{i,5} \end{pmatrix} = \begin{pmatrix} a_{11} a_{21} \\ a_{12} a_{22} \\ a_{13} a_{23} \\ a_{14} a_{24} \\ a_{15} a_{25} \end{pmatrix} \begin{pmatrix} f_{i,1} \\ f_{i,2} \end{pmatrix} + \begin{pmatrix} e_{i,1} \\ e_{i,2} \\ e_{i,3} \\ e_{i,4} \\ e_{i,5} \end{pmatrix}, \qquad \mathrm{i} = 1,2,3,\ldots,\mathrm{n}$$

観測値：$\boldsymbol{y}_i = \boldsymbol{A} \cdot \boldsymbol{f}_i + \boldsymbol{e}_i$

因子負荷量　因子得点　独自因子

因子分析モデルのパス図



共通・独自因子に関する仮定

- f_iは平均0、分散1
- f_iとf_jは独立
- f_iとe_iは独立
- $e_i \sim N(0, \sigma_i^2)$　正規分布
- e_iとe_jは独立

因子分析モデルの推定（１）

$$y_k = a_{1,k} f_1 + a_{2,k} f_2 + e_k \quad (k = 1,2,3,4,5)$$

$$e_k \sim N(0, \sigma_k^2)$$

独自因子（Uniqueness）：

ここでは、２つの因子 f_1 と f_2 で表せない部分の大きさ

■ **モデル**（理論）より

分散

$$\begin{aligned} Var(y_k) &= Var(a_{1,k} f_1 + a_{2,k} f_2 + e_k) \\ &= a_{1,k}^2 Var(f_1) + a_{2,k}^2 Var(f_2) + Var(e_k) \\ &= a_{1,k}^2 + a_{2,k}^2 + \sigma_i^2 \end{aligned}$$

共分散

$$\begin{aligned} Cov(y_i, y_j) &= Cov(a_{1,i} f_1 + a_{2,i} f_2 + e_i,\ a_{1,j} f_1 + a_{2,j} f_2 + e_j) \\ &= a_{1,i} a_{1,j} Var(f_1) + a_{2,i} a_{2,j} Var(f_2) + Cov(e_i, e_j) \\ &= a_{1,i} a_{1,j} + a_{2,i} a_{2,j} \end{aligned}$$

因子分析モデルの推定（２）

$$y_k = a_{1,k} f_1 + a_{2,k} f_2 + e_k \quad (k = 1,2,3,4,5)$$

$$e_k \sim N(0, \sigma_k^2)$$

独自因子（Uniqueness）：

ここでは、２つの因子 f_1 と f_2 で表せない部分の大きさ

■ **データ** より

$$s_{i,j} = \frac{\sum_{k=1}^{n} (y_{i,k} - \bar{y_i})\,(y_{j,k} - \bar{y_j})}{n-1}$$

ただし、$\overline{y_k} = \frac{1}{n}\sum_{j=1}^{n} y_i$ （平均値）

因子分析モデルの推定

■ モデルより

$$\Sigma = \begin{pmatrix} a_{11}^2 + a_{21}^2 + \sigma_1^2 & a_{11}a_{21} + a_{12}a_{22} & \cdots\cdots \\ a_{11}a_{21} + a_{12}a_{22} & a_{12}^2 + a_{22}^2 + \sigma_2^2 & \vdots \\ \vdots & \ddots\ddots & \\ & \cdots\cdots & a_{15}^2 + a_{25}^2 + \sigma_5^2 \end{pmatrix}$$

■ データより

$$S = \begin{pmatrix} s_{11} & s_{12} & \cdots \\ s_{21} & \ddots & \vdots \\ \vdots & \cdots\cdots & s_{55} \end{pmatrix}$$

$\min|\Sigma - S|$とする a_{ij}, f_i, σ_i を求める

5教科成績データの読込み

① Rstudio起動する

② ＞ library(Rcmdr)　　※コマンドラインから Rコマンダー を起動する

③ 演習ファイル “seiseki.csv” を読み込む

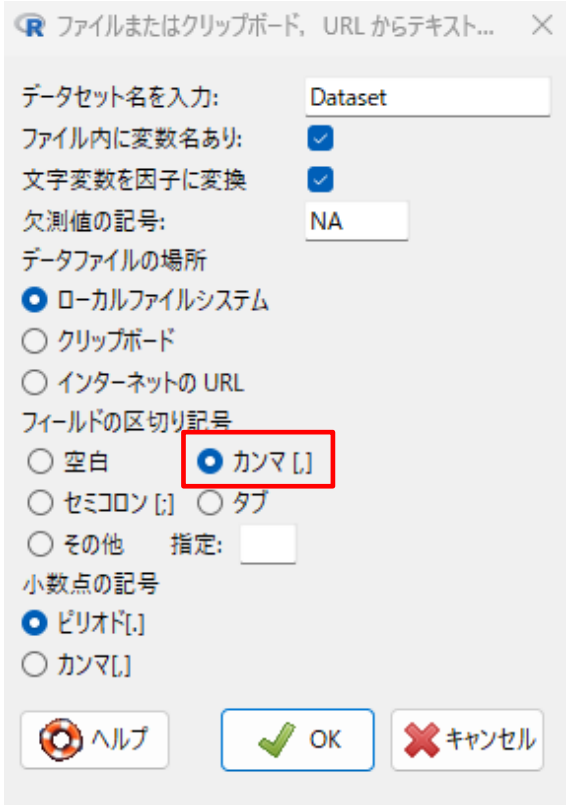
- Rstudio　＞ Dataset<-read.csv(“seiseki.csv”)

　　又は

- Rコマンダー　(データ) → (データインポート) → (テキストファイルまたはクリップボード・・・) →

　✔OKを選択して、seiseki.csv を指定する

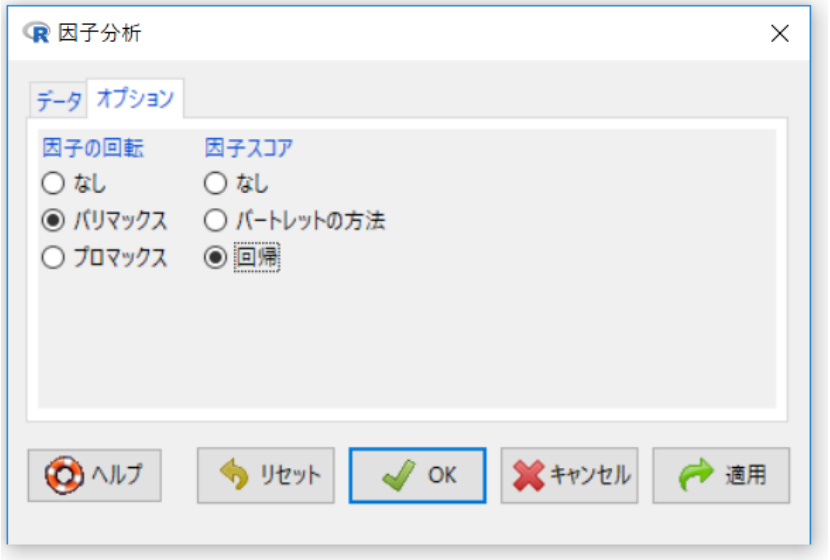
④ 演習データが Dataset に読込まれる



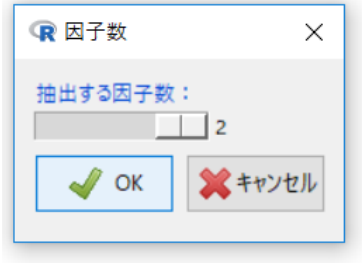
因子分析の推定

Rコマンダーから 【統計量】 → 【次元解析】 → 【因子分析】

変数として5項目全てを選択



オプションはバリマックスと回転を選択



因子数は２を設定

因子分析の結果

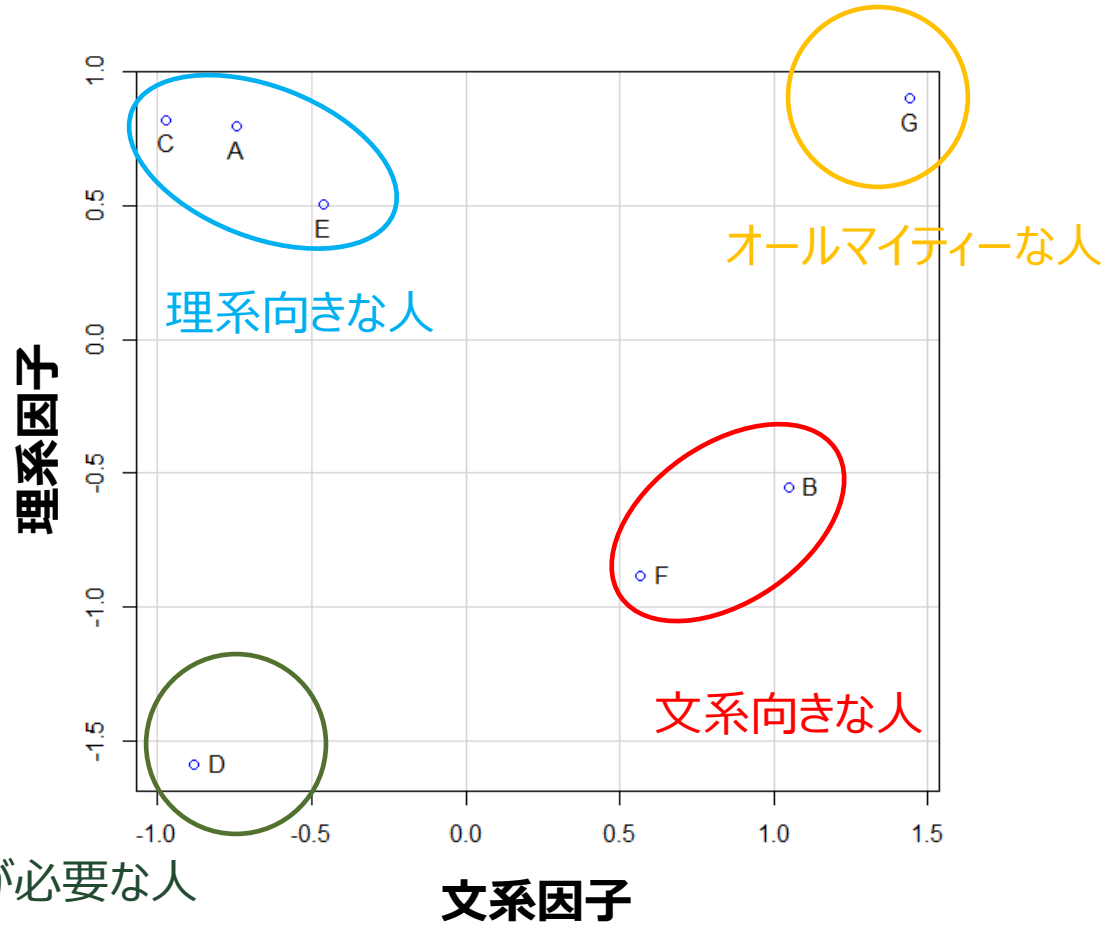
```
Call:
factanal(x = ~english + japanese + math + science + sociology,     factors = 2, data = Dataset, scores = "regression",
rotation = "varimax")

Uniquenesses:
  english  japanese      math   science sociology
    0.217     0.005     0.033     0.005     0.010

Loadings:
           Factor1 Factor2
english     0.877  -0.114
japanese    0.997
math                0.983
science    -0.229   0.971
sociology   0.992

                Factor1 Factor2
SS loadings       2.80   1.930
Proportion Var    0.56   0.386
Cumulative Var    0.56   0.946

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 3 on 1 degree of freedom.
The p-value is 0.0831
```

独自因子の大きさ

因子負荷量（各評価項目で共通）

２ファクターで全体の94.6%がカバーできる

潜在変数によるマッピング




```
Loadings:
          Factor1 Factor2
english     0.877  -0.114
japanese    0.997
math                0.983
science    -0.229   0.971
sociology   0.992
```

文系因子

理系因子

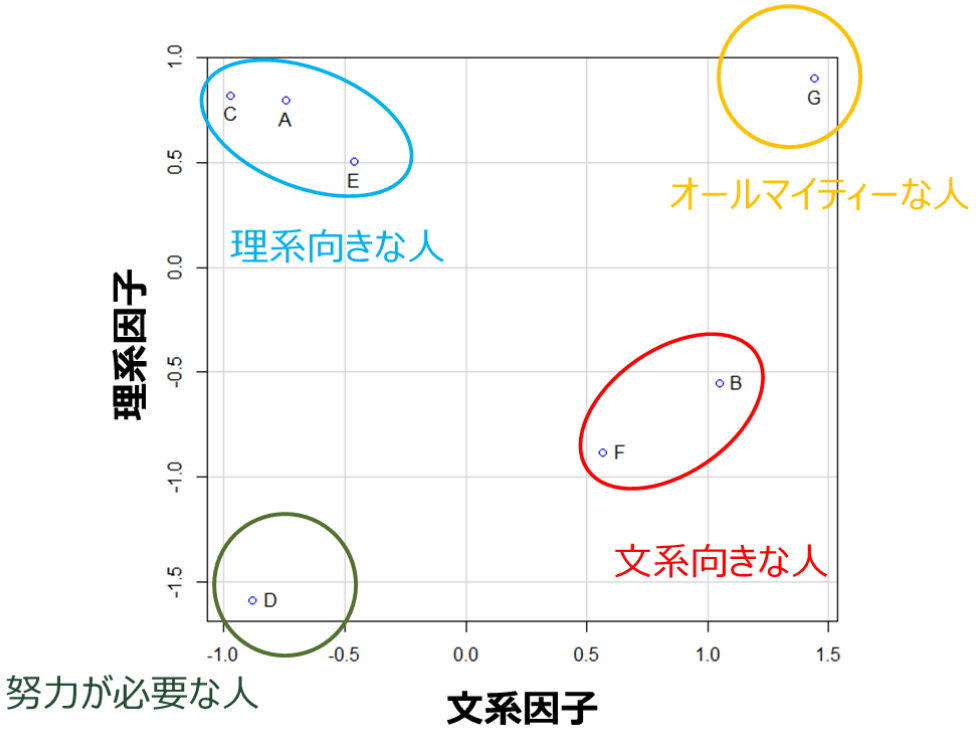
グラフは散布図から、F1とF2を指定する

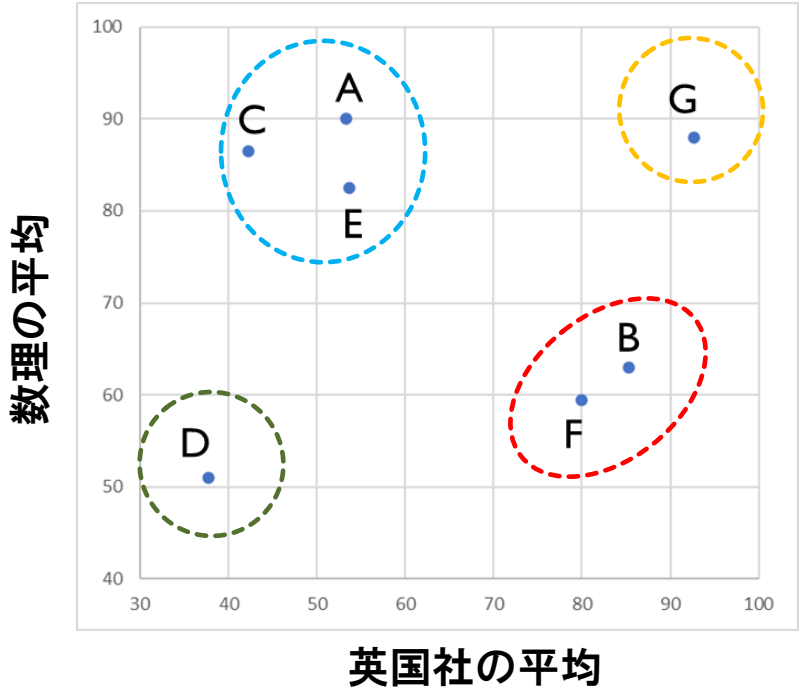
<u>ブランド名の表示</u>

Rコマンダーから　【データ】
→ 【アクティブデータセット】
→ 【ケース名の設定】

因子得点　VS. 平均値

ほとんど同じ特性を示すが、若干異なるものの同様な配置が確認できる

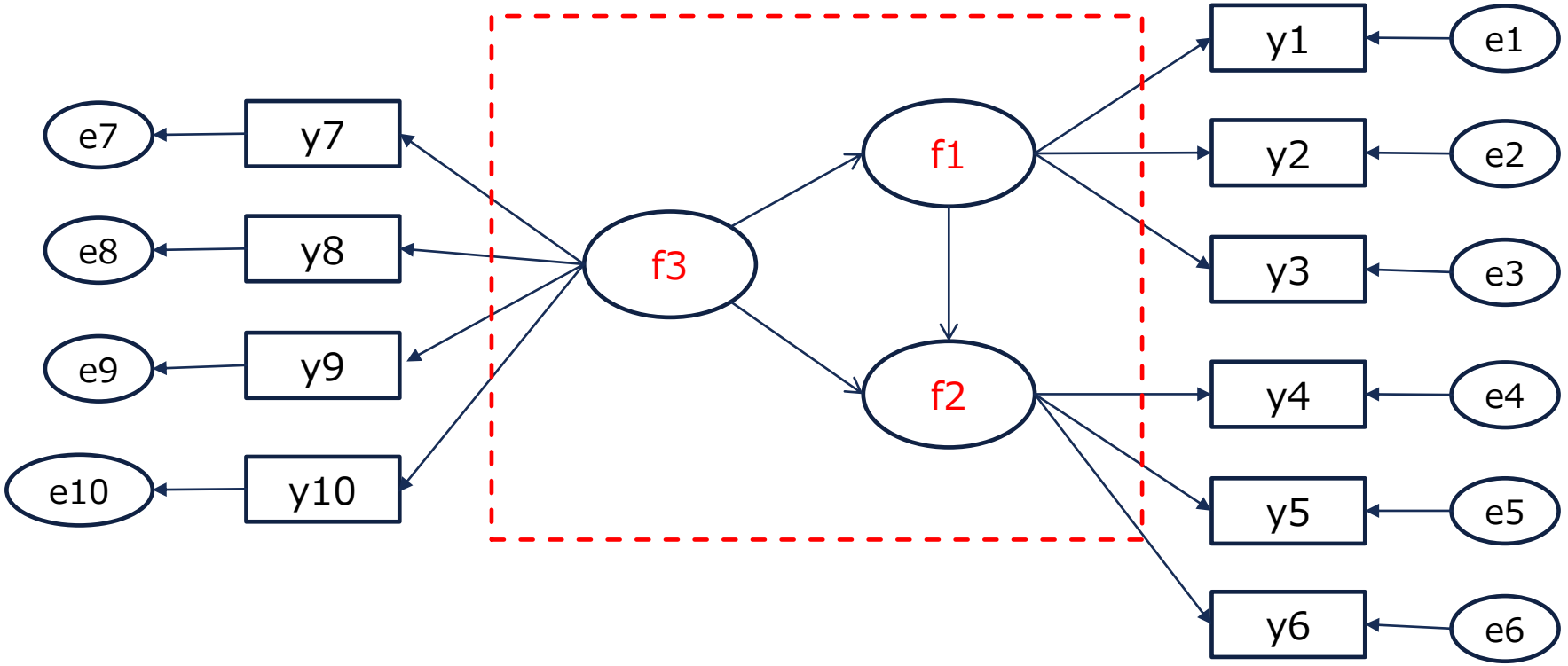




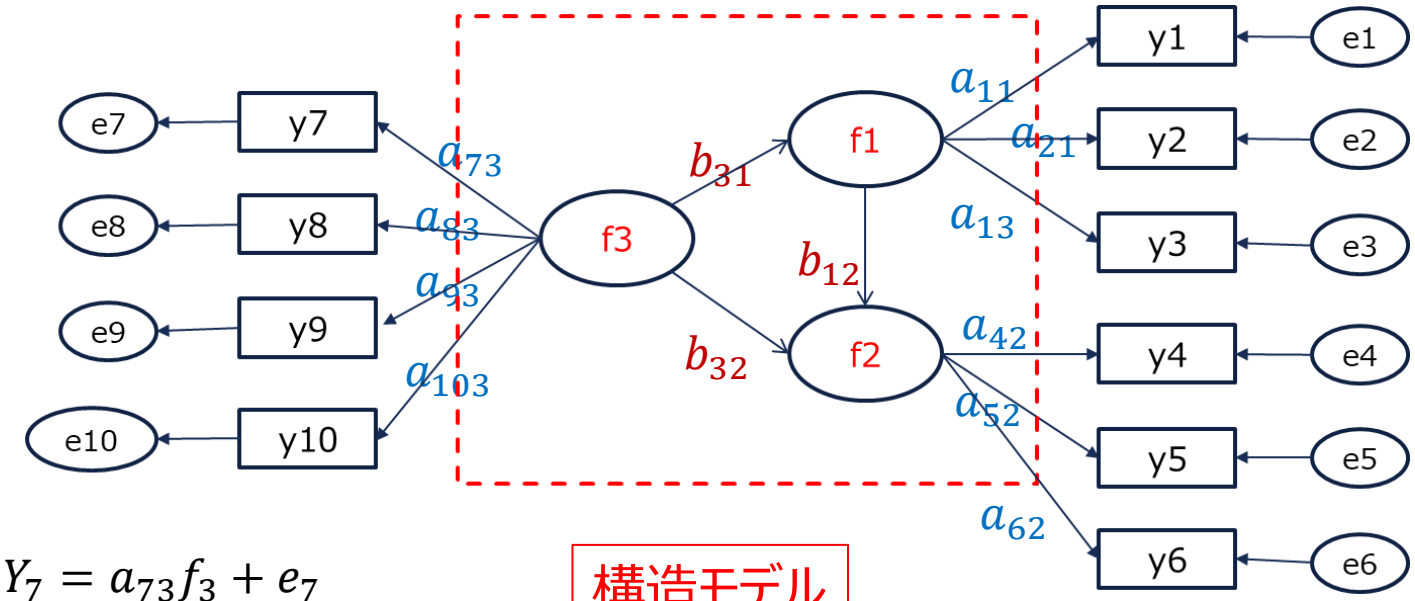
共分散構造分析

共分散構造分析のパス図

共分散構造分析では**潜在因子間での構造**も含めてモデリングすることができる
⇒　但し、**因子間構造の仮説**が必要となり、いくつかのモデル構築後に評価する



共分散構造モデリング



観測モデル

$$Y_7 = a_{73}f_3 + e_7$$
$$Y_8 = a_{83}f_3 + e_8$$
$$Y_9 = a_{93}f_3 + e_9$$
$$Y_{10} = a_{93}f_3 + e_{10}$$

構造モデル

$$f_1 = b_{31}f_3 + \epsilon_1$$
$$f_2 = b_{32}f_3 + b_{12}f_1 + \epsilon_2$$

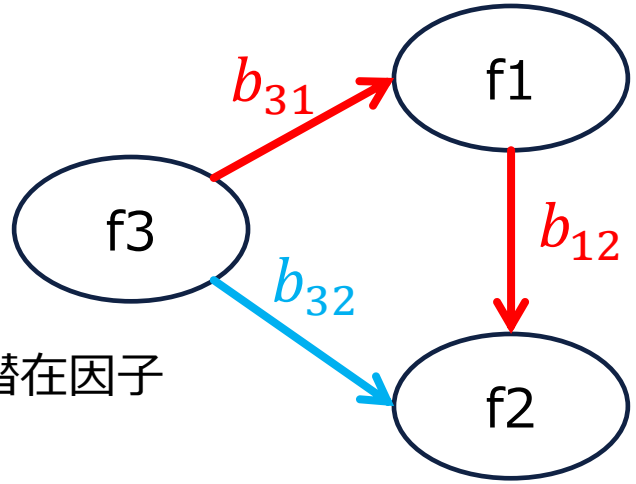
観測モデル

$$Y_1 = a_{11}f_1 + e_1$$
$$Y_2 = a_{21}f_1 + e_2$$
$$Y_3 = a_{31}f_1 + e_3$$
$$Y_4 = a_{42}f_2 + e_4$$
$$Y_5 = a_{52}f_2 + e_5$$
$$Y_6 = a_{63}f_3 + e_6$$

直接効果と間接効果

外生潜在因子

⇒ 原因となっている潜在因子



内生潜在因子

⇒ 他の潜在因子から決定される

f3によるf2への**直接効果**はb32で表す

f3によるf2への**間接効果**はf1を経由してb31とb12の積で表す

⇒ 最終的には、f2への**総合効果**は直接効果と間接効果の和として表す

➡ b32＋b31・b12

モデルの識別性

構造モデル

$$f_1 = b_{31}f_3 + \epsilon_1$$

$$f_2 = b_{32}f_3 + b_{12}f_1 + \epsilon_2$$

$$\epsilon_1 \sim N(0,1) \qquad \epsilon_2 \sim N(0,1)$$

構造モデルを構築する潜在変数は、そもそも観測することはできないため、スケール変換について**識別性のないモデル**といわれる。

通常は、構造モデルの誤差分散を予め１と定義して識別性を持たせる。

さらに、同一の潜在変数を影響要因に有する観測モデルのパラメータ１つを１と基準化して識別性を担保する。

顧客満足度(CSI)データの読込み

① Rstudio起動する

② > library(Rcmdr)　※コマンドラインから Rコマンダー を起動する

③ 演習ファイル “CSI.csv” を読み込む

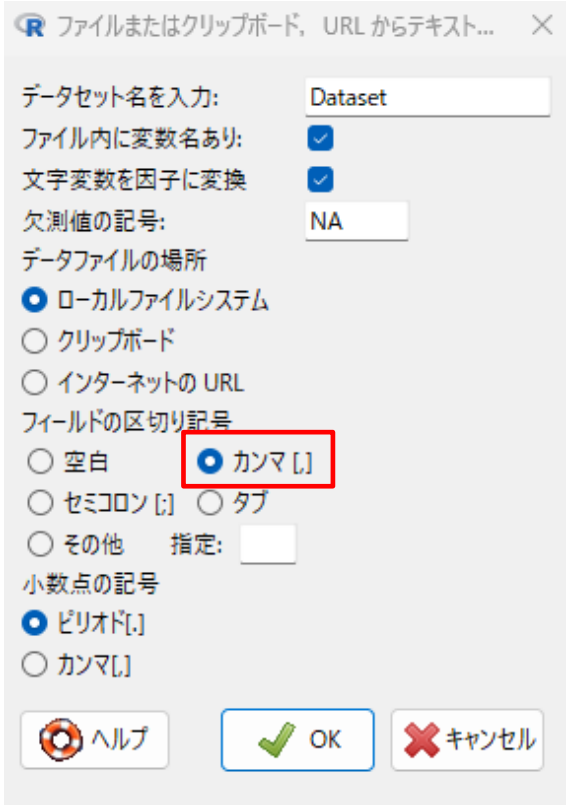
- Rstudio　> Dataset<-read.csv(“CSI.csv”)

  又は

- Rコマンダー　(データ) → (データインポート) → (テキストファイルまたはクリップボード・・・) →

  ✔OKを選択して、CSI.csv を指定する

④ 演習データが Dataset に読込まれる



CSIデータ

y1	y2	y3	y4	y5	y6	y7	y8	y9	y10
8	8	7	9	8	5	5	8	7	6
5	5	5	5	10	6	5	5	4	4
8	8	7	10	8	7	7	8	7	8
10	10	10	10	10	10	10	10	10	9
7	9	7	9	6	6	8	8	7	7
9	9	8	9	8	8	8	9	9	8
6	7	7	6	5	6	7	7	6	4
7	4	5	4	7	6	7	5	4	4
9	9	9	10	10	5	9	10	7	8
7	7	7	7	6	6	6	7	7	7
10	10	10	10	10	10	9	10	6	8
8	8	6	6	8	6	8	7	6	7
10	10	10	10	10	10	10	10	10	10
8	8	8	8	8	8	8	7	7	7
7	9	7	4	5	7	8	7	8	7
9	10	10	10	9	9	10	9	6	6
8	8	8	8	7	6	7	8	8	8
8	8	6	6	7	5	5	7	6	7
9	9	9	9	9	6	9	9	9	8

ある製品に対して、製品の購買前後での評価について
10段階評価のアンケート調査した結果（100人分）

Y1	購買後のデザイン性評価
Y2	購買後の信頼性評価
Y3	カスタマイズに対する評価
Y4	利用操作性に対する評価
Y5	購買前の品質に対する期待
Y6	購買前の要求に対する期待
Y7	購買前の信頼性に対する期待
Y8	総合的な評価
Y9	期待との不一致性
Y10	期待された性能との差異

```
co<- cor(data[,1:10])
co[upper.tri(co)] <- 0

#モデルの作成#

#測定方程式
   #ラベル
model <- specifyModel()
  知覚品質 -> y1, NA,  1        #測定方程式, 識別性制約のため係数を 1 に固定
  知覚品質 -> y2, b12, NA
  知覚品質 -> y3, b13, NA
  知覚品質 -> y4, b14, NA
  顧客期待 -> y5, NA,  1
  顧客期待 -> y6, b22, NA
  顧客期待 -> y7, b23, NA
  顧客満足 -> y8, NA, 1
  顧客満足 -> y9, b32, NA
  顧客満足 -> y10, b33, NA
  y1 <-> y1, e01, NA            #測定方程式の分散設定
  y2 <-> y2, e02, NA
  y3 <-> y3, e03, NA
  y4 <-> y4, e04, NA
  y5 <-> y5, e05, NA
  y6 <-> y6, e06, NA
  y7 <-> y7, e07, NA
  y8 <-> y8, e08, NA
  y9 <-> y9, e09, NA
  y10 <-> y10, e10, NA
  知覚品質     -> 顧客満足,  b1,NA        #構造方程式
  顧客期待     -> 顧客満足,  b2,NA
  顧客期待     -> 知覚品質,  b4,NA
  知覚品質    <-> 知覚品質 , NA, 1         #構造方程式の分散設定
  顧客期待    <-> 顧客期待 , NA, 1
  顧客満足    <-> 顧客満足 , NA, 1
```

SEMモデルの構築

※ パッケージインストール ＞ install.packages(“sem”)

$$Y_1 = 1 \cdot f_1 + e_1$$
$$Y_2 = a_{21} f_1 + e_2$$
$$Y_3 = a_{31} f_1 + e_3$$
$$Y_4 = a_{42} f_2 + e_4$$
$$Y_5 = 1 \cdot f_2 + e_5$$
$$Y_6 = a_{63} f_3 + e_6$$
$$Y_7 = a_{73} f_3 + e_7$$
$$Y_8 = 1 \cdot f_3 + e_8$$
$$Y_9 = a_{93} f_3 + e_9$$
$$Y_{10} = a_{103} f_3 + e_{10}$$

観測モデル

$$f_1 = b_{31} f_3 + \epsilon_1$$
$$f_2 = b_{32} f_3 + b_{12} f_1 + \epsilon_2$$

構造モデル

モデル推定結果

※ パッケージインストール ＞ install.packages(“Pathdiagram”)
＞ install.packages(“diagramR”)

```
#分析と出力#
result <- sem(model,co,N=100) #モデル,相関係数 or 共分散行列,サンプル数の並び
summary(result,
        fit.indices = c("GFI", "AGFI", "RMSEA","NFI", "NNFI",
                        "CFI", "RNI", "IFI", "SRMR", "AIC",
                        "AICc", "BIC", "CAIC"))
stdCoef(result)     #標準解の表示

#因子スコアの計算#
fs2<-fscores(result,data)

#パス図の作成#
pathDiagram(result, out.file="csi.txt", ignore.double=FALSE,
edge.labels="values", digits=3,
node.font=c("C:/WINDOWS/Fonts/msgothic.ttc",10))

#満足度得点の計算と表示#
cscore<- (fs2[,1]-min(fs2[,1]))/(max(fs2[,1])-min(fs2[,1]))*100
hist(cscore)
mscore<- mean(cscore)
mscore
```

```
 Model Chisquare =  195.6711   Df =  35 Pr(>Chisq) = 3.164196e-24
 Goodness-of-fit index =  0.748378
 Adjusted goodness-of-fit index =  0.604594
 RMSEA index =  0.2153363   90% CI: (NA, NA)
 Bentler-Bonett NFI =  0.767247
 Tucker-Lewis NNFI =  0.740377
 Bentler CFI =  0.798071
 Bentler RNI =  0.798071
 Bollen IFI =  0.8005773
 SRMR =  0.5259738
 AIC =  235.6711
 AICc =  206.304
 BIC =  34.49018
 CAIC =  -0.5098221

 Normalized Residuals
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-4.0893 -2.6985 -2.0772 -1.9622 -1.2448  0.2512

 R-square for Endogenous Variables
知覚品質       y1       y2       y3       y4       y5       y6       y7
  0.4418   0.8873   0.9119   0.8988   0.5656   0.5820   0.8040   0.7587
顧客満足       y8       y9      y10
  0.5381   0.9066   0.8433   0.8010
```

Parameter Estimates

	Estimate	Std Error	z value	Pr(>|z|)			
b12	0.9162105	0.04305966	21.277700	1.826866e-100	y2	<---	知覚品質
b13	0.9054438	0.04400241	20.577142	4.398638e-94	y3	<---	知覚品質
b14	0.6463132	0.06189876	10.441456	1.603325e-25	y4	<---	知覚品質
b22	1.0265356	0.08318691	12.340110	5.508078e-35	y6	<---	顧客期待
b23	0.9885477	0.08415097	11.747312	7.290076e-32	y7	<---	顧客期待
b32	0.8249091	0.04686889	17.600357	2.447748e-69	y9	<---	顧客満足
b33	0.7892953	0.04888956	16.144454	1.242427e-58	y10	<---	顧客満足
e01	0.2275598	0.04496172	5.061190	4.166485e-07	y1	<-->	y1
e02	0.1453096	0.03293616	4.411856	1.024884e-05	y2	<-->	y2
e03	0.1652789	0.03459975	4.776881	1.780353e-06	y3	<-->	y3
e04	0.5746909	0.08545514	6.725059	1.755209e-11	y4	<-->	y4
e05	0.7183356	0.11947607	6.012381	1.828183e-09	y5	<-->	y5
e06	0.2568796	0.06994935	3.672367	2.403146e-04	y6	<-->	y6
e07	0.3108622	0.07152943	4.345934	1.386840e-05	y7	<-->	y7
e08	0.2230454	0.06527017	3.417264	6.325381e-04	y8	<-->	y8
e09	0.2737678	0.05622600	4.869062	1.121294e-06	y9	<-->	y9
e10	0.3351215	0.06108596	5.486064	4.109881e-08	y10	<-->	y10
b1	0.6498263	0.11842667	5.487161	4.084438e-08	顧客満足	<---	知覚品質
b2	0.2836745	0.16096707	1.762314	7.801632e-02	顧客満足	<---	顧客期待
b4	0.8895948	0.12393267	7.178049	7.071336e-13	知覚品質	<---	顧客期待

基準化済

b11=1

b21=1

b31=1

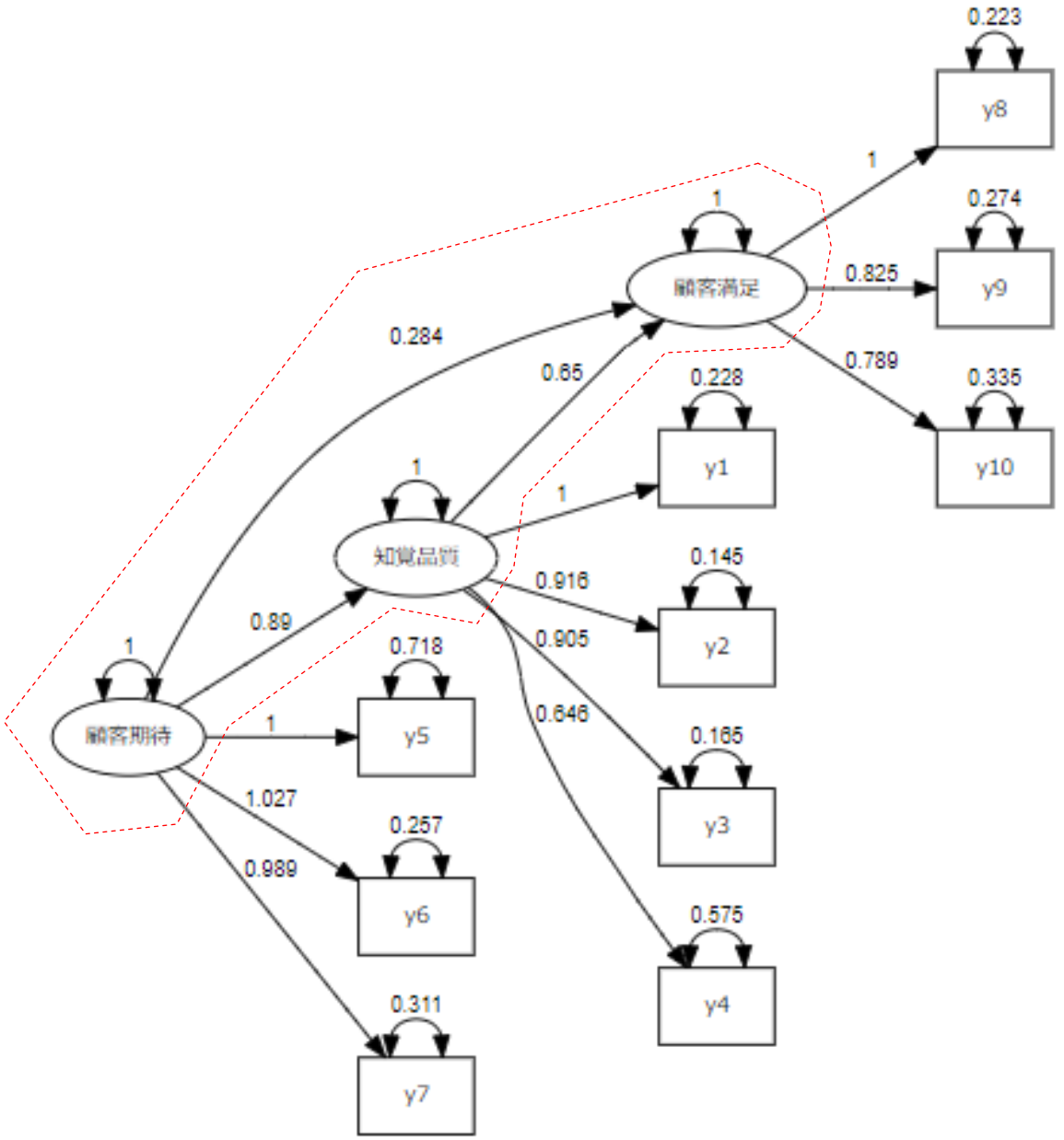
顧客満足に対する顧客期待の効果

直接効果：0.284

間接効果：0.89x0.65=0.579

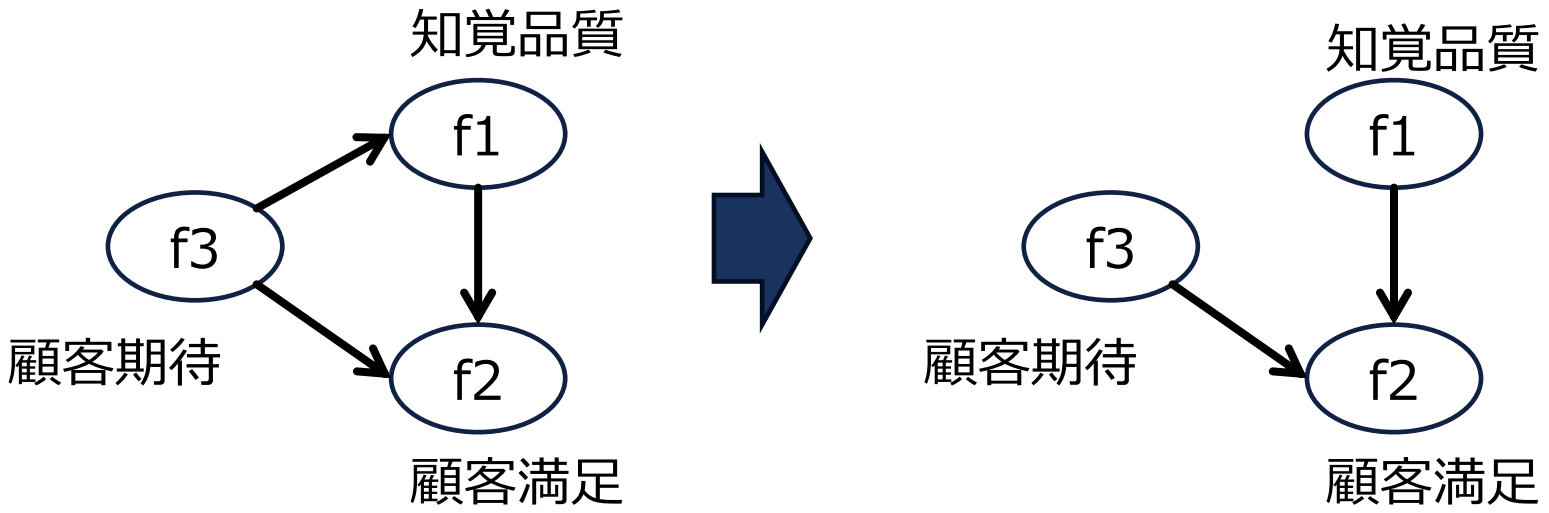
総合効果：0.863

```
Model Chisquare =  195.6711   Df =  35 Pr(>Chisq) = 3.164196e-24
Goodness-of-fit index =  0.748378
Adjusted goodness-of-fit index =  0.604594
RMSEA index =  0.2153363   90% CI: (NA, NA)
Bentler-Bonett NFI =  0.767247
Tucker-Lewis NNFI =  0.740377
Bentler CFI =  0.798071
Bentler RNI =  0.798071
Bollen IFI =  0.8005773
SRMR =  0.5259738
AIC =  235.6711
AICc =  206.304
BIC =  34.49018
CAIC =  -0.5098221
```



課題：CSIデータを用いたSEM

演習でCSIデータを用いたSEM分析で顧客満足は、顧客期待からの直接効果と知覚品質を介した間接効果で表現できると仮定した。本課題では、顧客期待と知覚品質は独立であるという仮定でSEM分析を実施し、本演習結果と比較して論じなさい。




データマイニングを楽しもう！