I　共分散構造分析の速習

abstract　マーケティングサイエンスの代表的な手法である共分散構造分析について、分析を行ううえで最小限必要な知識を導入し、デモデータを用いてPython言語による計算例を紹介します。

In [1]:
```
# import文
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd

from semopy import Model, calc_stats
from semopy.plot import semplot
```

1　Introduction

共分散構造分析（Structural Equation Modeling）は、複数の変数間の関係性を分析する手法の一つです。マーケティング領域での代表的な応用例には、自社のブランドイメージや顧客の意識がどのように顧客の商品の購買意欲に結びついているのかを調査する分析課題が挙げられます。人事データ分析でも社員のモチベーションに、レクリエーションや社員研修がどのような影響を与えているのかを調べるために応用されています。もちろん、応用例はマーケティング領域や人事データ分析に限りません。

共分散構造分析の入力には、関係性を分析したい変数に関するデータに加え、分析者が事前に変数間の関係について仮説を考え、これを図に落とした**パス図**（path diagram）とよばれるものを準備します。この結果、共分散構造分析の計算を通して

- 変数間の関係性の強さを表す**パス係数**の推定
- 変数間の関係性の有無を検討するパス係数の仮説検定

を実施することができます。さらに、自分の仮説の妥当性を検討するヒントとして**適合度指標**とよばれる統計量を計算することができます。また、変数間の関係について複数の仮説がある場合には、どの仮説が正しいかを検討するヒントとして**情報量規準**を計算することができます。

変数間の関係性を分析する手法としては、相関係数・回帰分析・主成分分析などのほうがよく知られているかもしれません。しかし、共分散構造分析はこれらの手法と比較して、パス図によって分析者の仮説に対応した分析を柔軟に実施できるという長所があります。

さて、このノートでは共分散構造分析を使うために最小限必要な知識を解説します。具体的には以下に掲げる四つの内容です。

- 共分散構造分析の基本的な単語であるパス図とパス係数の解説
- 実際のデータ分析における共分散構造分析の考え方
- Python言語を用いた計算の方法

2　パス図とパス係数

共分散構造分析では、事前に分析者が興味のある変数の間の関係性について仮説を立て、これをパス図とよばれる図に表現します。そこでこの節では、パス図とは何かを説明します。パス図とは何かを理解して、簡単な仮説をパス図で表現できるようになりましょう。

2.1　パス図

パス図は変数の間の関係性について、分析者が考える仮説をグラフで表現したものです。変数を表す四角□と丸○、変数間の関係を表す両側の矢印↔と片側の矢印→を用いて表されます。

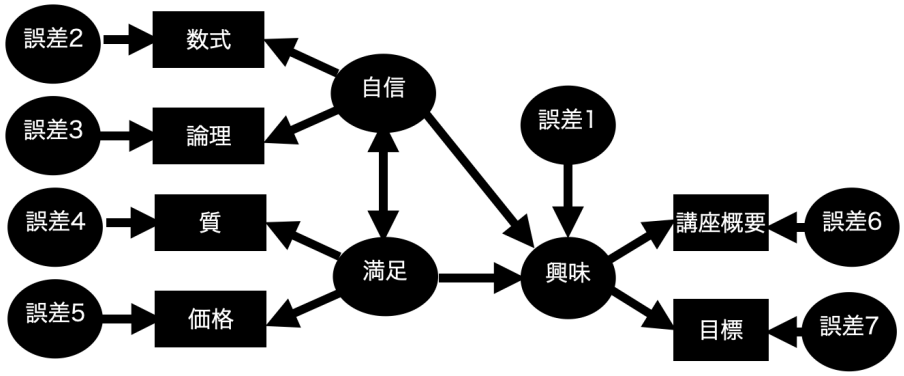


Fig 1　パス図

2.2　観測変数と潜在変数

変数は**観測変数**と**潜在変数**の二種類に分類されます。データのなかに値が含まれる変数を観測変数、含まれない変数を潜在変数と定義します。パス図では、観測変数を四角□、潜在変数を丸○で囲んで表現します。Fig 1では、以下のようになっています。

- 観測変数: 数式, 論理, 質, 価格, 講座概要, 目標
- 潜在変数: 自信, 満足, 興味, 誤差（7個）

Remark　必ずしも潜在変数を考えないといけないわけではありません。■

2.3　相関関係と回帰関係

変数間の関係性は**相関関係**と**回帰関係**の二種類に分類されます。相関関係は片方の変数の値が大きければもう片方も大きいというような大小関係の類似性に注目したものです。相関関係は以下のように両側の矢印↔を用いてグラフで表現されます。Fig 1では「自信」と「満足」の関係が相関関係として定義されています。

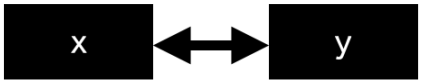


Fig 2　相関関係

回帰関係は、以下の二つ

- 変数 x の値が大きくなると変数 y はいくら大きくなる傾向にあるか
- 変数 x を使って変数 y の値の大きさをどれだけ正確に予測できるか

に注目した関係です。数式では $y = \alpha + \beta x + \epsilon$ に対応しています。変数 ϵ は変数 x で説明しきれなかった変数 y の大きさで、**誤差**とよばれる潜在変数です。回帰関係は以下のように片側の矢印→を用いてグラフで表現されます。Fig 1では「興味」を「自信」と「満足」の二変数で説明する回帰関係が定義されています。

Fig 3　回帰関係

パス図は、この相関関係のグラフと回帰関係のグラフを組み合わせて作ります。実際、Fig 1に掲げたパス図もこの組み合わせでできています。

Remark　実際の分析時に二つの変数間の関係性をどう設定するかは、必ずしも相関関係と回帰関係のどちらが正しいかといった「正しさ」の問題ではなく、分析者がどちらの関係に興味があるかを基準に設定すると考えて差し支えありまん。■

2.4　パス係数と誤差の分散

相関関係・回帰関係といった変数間の関係性にはその強さを表す値があります。これを**パス係数**（path coefficient）といいます。共分散構造分析は本質的には、データからパス係数と誤差の分散の推定値を計算する手法といえます。Fig 4において ρ は相関関係のパス係数、β は回帰関係のパス係数、σ^2 は誤差の分散です。

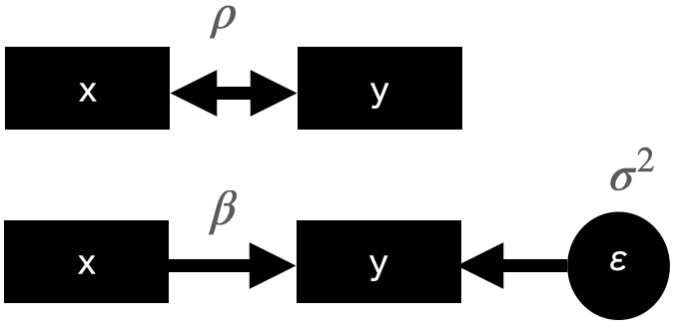


Fig 4　パス係数と誤差の分散

2.5　標準化解と非標準化解

パス図に現れる変数を標準化した場合に得られるパス係数の推定値を**標準化解**、そうでない場合のパス係数の推定値を**非標準化解**といいます。以下、標準化解と非標準化解それぞれにおけるパス係数の解釈について説明します。

標準化解の場合のパス係数の解釈を説明します。相関関係のパス係数は相関係数になります。回帰関係のパス係数はどの変数間の関係性が強いかを相対的に比較できる値として出力されます。誤差の分散は 0 以上 1 以下の間に収まり、0 に近いほど回帰関係が強く、正確に予測できると解釈することができます。

非標準化解の場合を説明します。回帰関係のパス係数は、変数 x の値を大きくするに従って変数 y の値がどれだけ大きくなる傾向にあるのかを読み取れる値として出力されます。誤差の分散は、変数の実際の値が回帰関係によって計算できる予測値からどれだけ離れうるかを表します。相関関係のパス係数は、共分散といって一般には解釈が難しい値になります。

標準化解、非標準化解がどうしてこのように解釈できるのか、気になる方もいらっしゃるでしょう。実は、パス図がどんな数式に対応しているのかを理解することで、このように解釈できることを自力で説明できるようになります。このことは、次回以降詳しく解説します。

Remark　標準化解の誤差の分散はその回帰関係の1-決定係数（1-分散説明率）の値であると考えると解釈して差し支えありません。■

3　実際の分析の流れ

共分散構造分析の実際の分析の流れを、具体的な分析課題を設定して説明します。ここで設定した分析課題は第4節で実際に分析します。

3.1　実際の分析の流れ

共分散構造分析は一般に、以下の流れに従って実施されます。

1. 分析者の仮説をパス図で表現する。
2. 変数間の相関関係を把握し、パス図との整合性を検討する。
3. R言語やPython言語を用いて共分散構造分析の計算を実施する。
4. 共分散構造分析の結果に対して適合度指標、情報量規準を計算する。

3.2　具体例: 第4節のための分析課題の設定

例題　あなたは動画教材コンテンツを販売しているS社のデータ分析者です。今回、新しいシリーズの動画教材が開発されることになり、新シリーズがどのような意識を持つ顧客から新商品へ興味を持ってもらえるのかを分析する課題に取り組んでいます。そこで、事前に以下のような仮説を考えました。

> 新シリーズの動画教材に関心を持つ顧客は、数学への苦手意識を持っている方、そしてこれまでにS社への満足度を感じている方ではないか？

さて、この仮説を検討するためにアンケート調査でデータを収集しました。アンケートは以下の六つの質問項目からなります。

- `x1`：数式への苦手意識はありますか？
- `x2`：論理的な思考への苦手意識はありますか？
- `x3`：これまでの講座の質には満足していますか？
- `x4`：これまでの講座の価格設定は高いと思いますか、安いと思いますか？
- `x5`：新シリーズの動画教材の講座概要に興味が持てましたか？
- `x6`：新シリーズの動画教材が掲げる学習目標は魅力的に感じますか？

なお、質問項目 `x1` と `x2` は数学への苦手意識、質問項目 `x3` と `x4` はS社への満足度、質問項目 `x5` と `x6` は新シリーズの動画教材への関心を測定するために準備したものです。質問項目への回答はすべて五件法で得られます。400名に調査を実施し、301件の有効回答を得ました。この結果は `data` ディレクトリの `demand_survey.csv` に記録されています。

In [2]:
```
# データの読み込み
data = pd.read_csv("./data/demand_survey.csv")
data.head(n = 5)
```

Out[2]:

	x1	x2	x3	x4	x5	x6
0	2	4	2	4	2	3
1	3	3	2	2	3	3
2	3	3	2	2	2	1
3	3	4	3	3	2	2
4	3	2	3	3	3	3

以下の問いに答えながら、顧客の意識と新商品への興味の関係を共分散構造分析を用いて調査しましょう。

[1]　以下の手順で仮説に対応するパス図を作ってください。

1. 変数 **苦手意識**，**満足度**，**新シリーズへの興味**，`x1`，...，`x6` それぞれについて、潜在変数と観測変数のどちらに該当するかを答えてください。
2. 三変数 **苦手意識**，**満足度**，**新シリーズへの興味** に想定している関係を、仮説に従ってパス図で表現してください。これを**構造方程式**といいます。また、共分散構造分析からどのような結果が出力されると想定しているのかを、パス係数という言葉を用いて説明してください。
3. 潜在変数を観測変数から定義するパス図（またはそれに対応する数式）を**測定方程式**といいます。測定方程式に対応するパス図をすべて列挙してください。
4. 1から3の結果をまとめて一つのパス図で表現してください。
5. `semopy` パッケージのルールに従って、パス図をコーディングしてください。

[2]　データの相関行列を計算して、パス図との整合性を検討してください。

[3]　`semopy` パッケージを用いて共分散構造分析の計算を実施してください。特に、出力結果を小問[1]の2で想定した内容と比較してください。

[4]　[3]で得られた結果の適合度指標と情報量規準を計算してください。

4　Python言語を用いた計算

Python言語には共分散構造分析を計算するためのパッケージ `semopy` があります。この節では、 `semopy` パッケージの使い方を説明しながら、実際のデータに対する共分散構造分析の計算例を紹介します。

4.1　パス図の作成（[1] 1-4の解答）

まずはパス図に現れる観測変数と潜在変数を整理しましょう。変数 `x1` , ..., `x6` はアンケートで調査される質問項目なので、実際に値を得ることができます。つまり観測変数です。一方で、 **`苦手意識`** , **`満足度`** , **`新シリーズへの興味`** は実際の値を得ることはせず、観測変数 `x1` , ..., `x6` を通して測定する変数なので潜在変数です。

仮説の上では三変数 **`苦手意識`** , **`満足度`** , **`新シリーズへの興味`** に対して、 **`新シリーズへの興味`** を **`苦手意識`** , **`満足度`** の二変数で説明する回帰関係を想定しています。つまり、構造方程式は以下のようなパス図で表現されます。特に、 **`苦手意識`** から **`新シリーズへの興味`** へのパス係数、 **`満足度`** から **`新シリーズへの興味`** へのパス係数が正の値になることを想定しています。

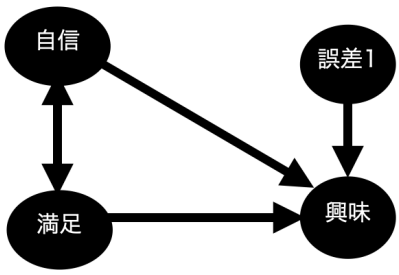


Fig 5　構造方程式に対応するパス図

測定方程式は、潜在変数を用いて観測変数を説明するような回帰関係を考えます。複数の観測変数を説明するような変数として潜在変数を定義することで、観測変数の共通因子を抽出するようなイメージです。今回の場合、潜在変数 **`苦手意識`** が観測変数 `x1` , `x2` を、潜在変数 **`満足度`** が観測変数 `x3` , `x4` を、潜在変数 **`新シリーズへの興味`** が観測変数 `x5` , `x6` を説明するような回帰関係を測定方程式として立てます。

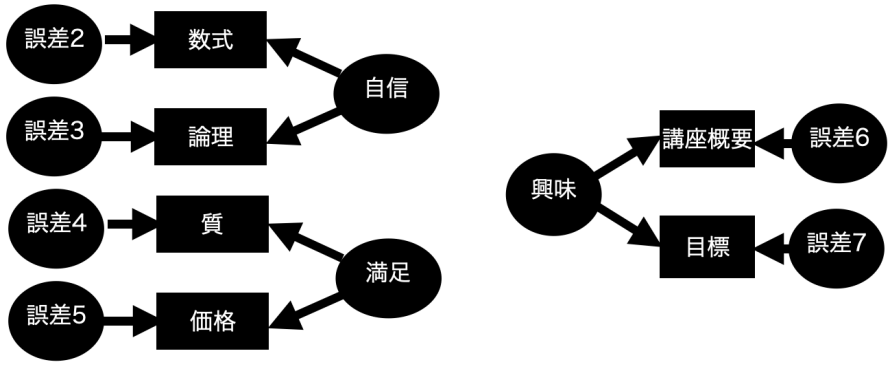


Fig 6　測定方程式に対応するパス図

以上をまとめると、仮説を表現するパス図は以下のようになることがわかります。

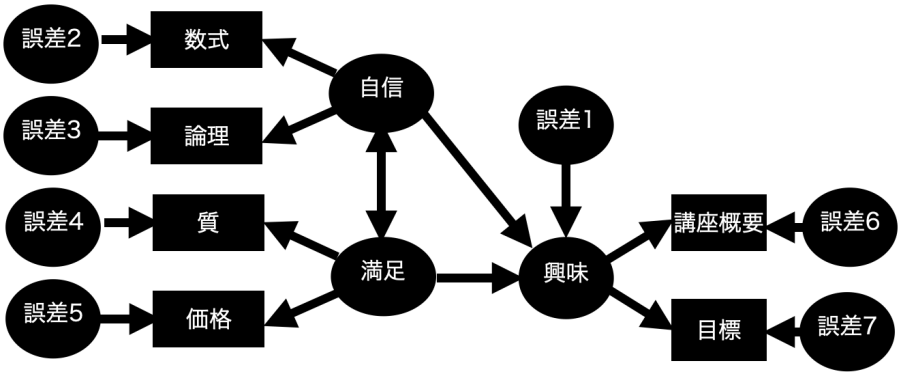


Fig 7　仮説に対応するパス図（Fig 1と同様）

Remark　共分散構造分析では、潜在変数 `z` から観測変数を説明するような回帰関係のことを変数 `z` の測定方程式といいます。一方、その他の関係をすべて構造方程式といいます。■

4.2　パス図のコーディング（[1] 5の解答）

パス図を作成したら、`semopy` パッケージに入力できるようにパス図をコーディングしましょう。`semopy` では以下の3つのシンタックスを用いてパス図をコーディングする約束になっています。

- `x~~y` : 変数 `x` と変数 `y` の相関関係
- `y~x` : 変数 `y` を変数 `x` で説明する回帰関係
- `z=~x` : 潜在変数 `z` を観測変数 `x` で回帰する測定方程式

それぞれcovariance operator、regression operator、measurement operatorとよばれています。演算子 `+` を用いることで、より複雑な関係を定義することもできます。例えば、

- `x~~y1+y2` : 変数 `x` は変数 `y1`, `y2` の両方と相関関係がある。
- `y~~x1+x2` : 変数 `y` は変数 `x1` と `x2` から説明される回帰関係がある。
- `z=~x1+x2` : 潜在変数 `z` は観測変数 `x1`, `x2` それぞれに対して回帰関係を想定した測定方程式がある。

このシンタックスに倣って、4.1節のパス図をコーディングすると以下のようになります。

In [3]:
```
# パス図のコーディング
path_diagram = '''
    # 測定方程式
    confidence =~ x1 + x2
    satisfaction =~ x3 + x4
    interest =~ x5 + x6

    # 構造方程式
    confidence ~ satisfaction
    interest ~ confidence + satisfaction
'''
```

4.3　データの相関行列の確認（[2]の解答）

相関行列（correlation matrix）とは、データに含まれる二変数間の相関係数を行列の形でまとめたものです。`pandas.DataFrame` オブジェクトの場合、`corr` メソッドを用いて計算することができます。

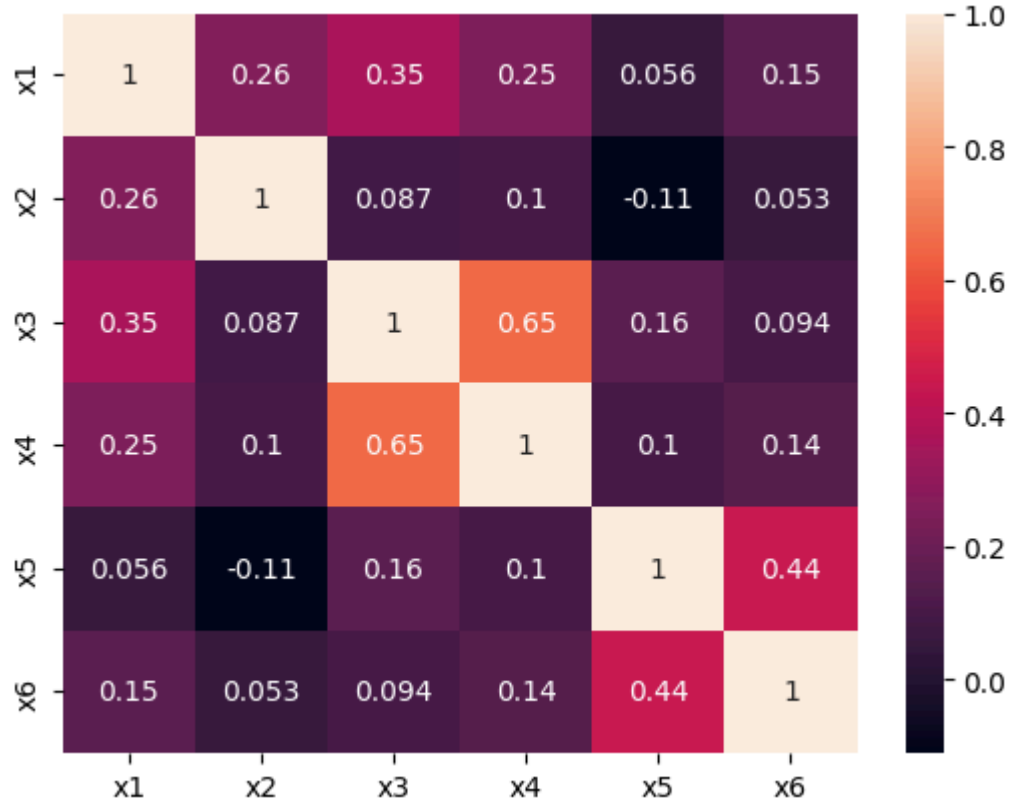
In [4]:
```
# 相関行列の確認
data.corr()
```

Out[4]:

	x1	x2	x3	x4	x5	x6
x1	1.000000	0.256330	0.354012	0.253765	0.055846	0.154977
x2	0.256330	1.000000	0.086873	0.104977	-0.110542	0.053176
x3	0.354012	0.086873	1.000000	0.651431	0.161940	0.094432
x4	0.253765	0.104977	0.651431	1.000000	0.104465	0.141943
x5	0.055846	-0.110542	0.161940	0.104465	1.000000	0.437009
x6	0.154977	0.053176	0.094432	0.141943	0.437009	1.000000

相関行列のパターンを確認しやすくするために、**ヒートマップ**を書いてみましょう。ヒートマップは `seaborn` パッケージを使うと簡単にかくことができます。

In [5]:
```
# 相関行列のヒートマップ
sns.heatmap(data.corr(), annot = True)
plt.show()
```



4.1節で考えたパス図が正しい場合、以下のことが従うはずです。

- 観測変数 `x1`, `x2` は潜在変数 **`自信`** を共通因子として相関する。
- 観測変数 `x3`, `x4` は潜在変数 **`満足`** を共通因子として相関する。
- 観測変数 `x5`, `x6` は潜在変数 **`興味`** を共通因子として相関する。

観測変数 `x1`, `x2` の間の相関は弱いですが、他の関係はパス図と整合性があることを確認することができます。

また、 **`自信`** と **`満足`** が **`興味`** を説明するような回帰関係が強いのであれば、 **`自信`** または **`満足`** を定義する観測変数 `x1`,..., `x4` と **`興味`** を定義する観測変数 `x5`, `x6` との間に相関が認められることが予想されますが、このような関係は相関行列からは確認できません。従って、仮説のような回帰関係は存在したとしても弱いのではないかということが推察されます。

4.4　共分散構造分析の計算（[3]の解答）

第1節では、共分散構造分析にデータとパス図を入力することで、以下の二つの分析が実施できることを説明しました。

- 変数間の関係性の強さを表すパス係数の推定
- 変数間の関係性の有無を検討するパス係数の仮説検定

`semopy` パッケージを用いて、実際にこの結果を出力してみましょう。`semopy` パッケージの `Model` クラスを用いて、共分散構造分析のインスタンスをたてます。`Model` クラスには4.2節で作った、パス図のコードを渡します。

In [6]:
```
# インスタンスを建てる
sem = Model(path_diagram)
sem
```

Out[6]:
```
<semopy.model.Model at 0x299117e50>
```

`Model` クラスで建てたインスタンス `sem` は `fit` メソッドを持っていて、このメソッドにデータを渡すことで、共分散構造分析の計算を実施することができます。

In [7]:
```
# 共分散構造分析の計算
sem.fit(data)
```

Out[7]:
```
SolverResult(fun=0.05731081785604619, success=True, n_it=33, x=array([0.
2842902 , 0.84351696, 0.94448617, 0.32112603, 0.07545032,
       0.10995942, 0.48445187, 0.22481404, 0.43434144, 0.08889804,
       0.23818576, 0.34708729, 0.36168598, 0.        , 0.58397632]), mes
sage='Optimization terminated successfully', name_method='SLSQP', name_o
bj='MLW')
```

`fit` メソッドを実行したあとの `sem` オブジェクトは、`inspect` メソッドを用いてパス係数の推定値や仮説検定の結果を確認することができます。ここでパス係数の仮説検定は、以下の検定問題に対して、帰無分布が標準正規分布の検定統計量 Z を計算したものになっています。

$$H_0 : \text{パス係数} = 0 \text{ v.s. } H_1 : \text{パス係数} \neq 0$$

出力結果は左から順に、変数1, operator, 変数2, パス係数の推定値（非標準化解）, 標準誤差, 標準化解, 検定統計量 Z の値, 仮説検定の $p-$値です。

In [8]:
```
# パス係数に対する推定値や仮説検定の結果
sem.inspect(std_est = True)    # std_est=True: 標準化解も出力
```

Out[8]:

	lval	op	rval	Estimate	Est. Std	Std. Err	z-value	p-value
0	confidence	~	satisfaction	0.321126	0.377803	0.064225	5.000054	0.000001
1	interest	~	confidence	0.075450	0.100586	0.079388	0.950396	0.341911
2	interest	~	satisfaction	0.109959	0.172465	0.063814	1.723124	0.084866
3	x1	~	confidence	1.000000	1.000000	-	-	-
4	x2	~	confidence	0.284290	0.256433	0.170185	1.670478	0.094825
5	x3	~	satisfaction	1.000000	0.931602	-	-	-
6	x4	~	satisfaction	0.843517	0.699229	0.134568	6.268352	0.0
7	x5	~	interest	1.000000	0.637304	-	-	-
8	x6	~	interest	0.944486	0.686031	0.393359	2.401082	0.016347
9	interest	~~	interest	0.224814	0.947030	0.099337	2.26314	0.023627
10	confidence	~~	confidence	0.361686	0.857265	0.237362	1.523772	0.127566
11	satisfaction	~~	satisfaction	0.583976	1.000000	0.102267	5.710287	0.0
12	x2	~~	x2	0.484452	0.934242	0.043832	11.052536	0.0
13	x4	~~	x4	0.434341	0.511078	0.071077	6.1109	0.0
14	x3	~~	x3	0.088898	0.132117	0.086921	1.022744	0.306429
15	x6	~~	x6	0.238186	0.529362	0.089526	2.660518	0.007802
16	x5	~~	x5	0.347087	0.593844	0.101974	3.403671	0.000665
17	x1	~~	x1	0.000000	0.000000	0.235346	0.0	1.0

この出力結果を解釈してみましょう。今回、特に注目したいのは以下の結果です。

1. **自信** と **満足** は **興味** を説明するか？するとしたら、どれくらい説明できるか。
2. **自信** と **満足** のどちらのほうが **興味** につながっているか？

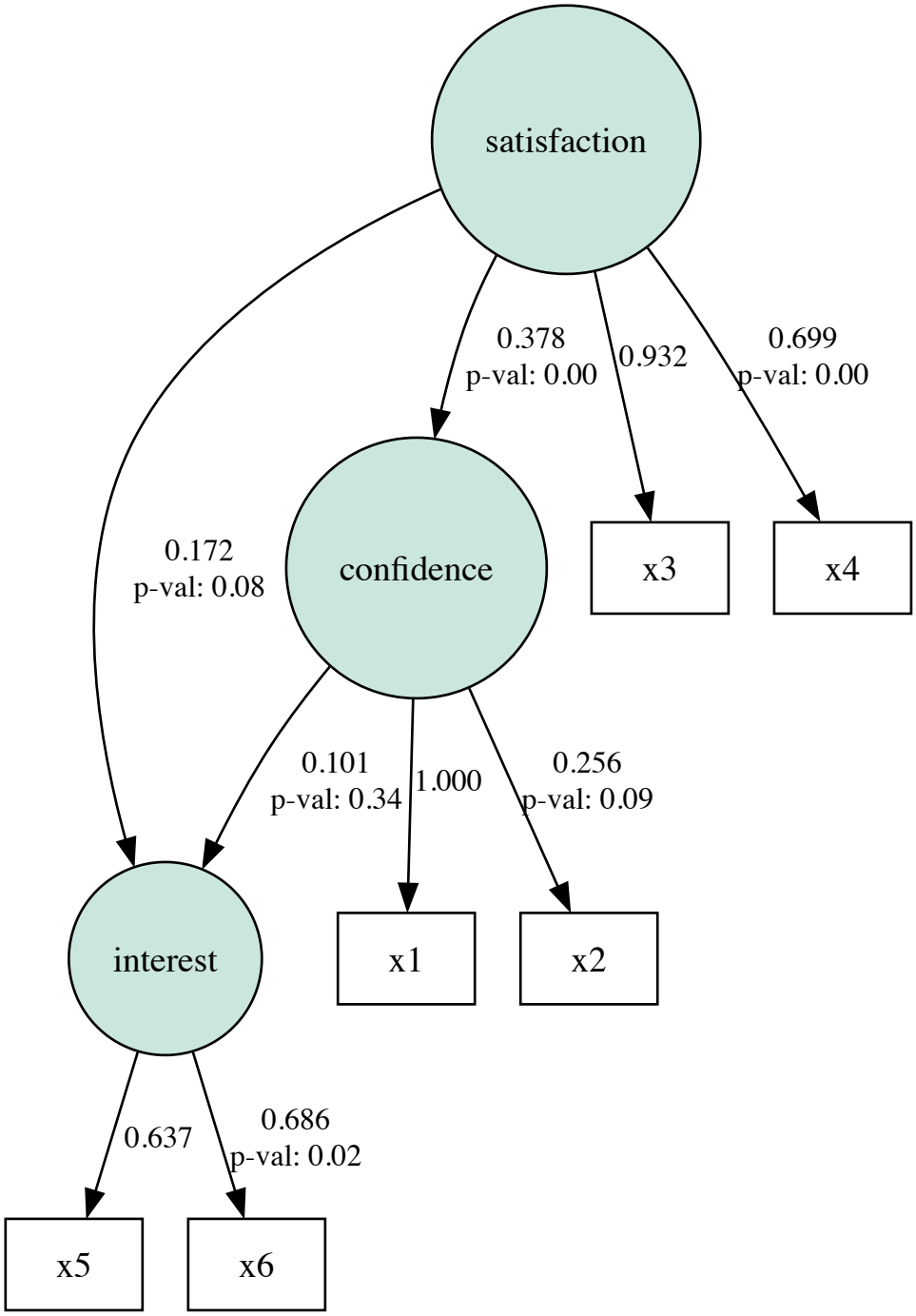
まず、**興味** の誤差の分散に注目してみましょう。標準化解の推定値は `0.947` であり、**自信** と **満足** からは説明できない部分が大きいことを示しています。決定係数でいうと、$1-0.947=0.053$ です。次に、**自信** から **興味** への回帰関係のパス係数と **満足** から **興味** への回帰関係のパス係数を標準化解で比較します。**自信** からのパス係数は 0.101、**満足** からのパス係数は 0.172 なので、**興味** はどちらかというと **満足** から説明できることがわかります。

参考として、仮説検定の結果にも注目してみましょう。**自信** から **興味** を説明する回帰関係のパス係数の $p-$値は 0.342 です。5% 有意水準を採用するなら、**自信** は **興味** を説明していないという判断になります。同様に、**満足** から **興味** を説明する回帰関係のパス係数の $p-$値は 0.085 です。5% 有意水準を採用するなら、**満足** は **興味** を説明していないという判断になります。

なお、`semopy.plot` モジュールの `semplot` 関数を用いることで、出力結果を以下のような図として出力することもできます。

In [9]:
```
# 推定結果
semplot(sem, "sem_result.pdf", std_ests = True)
```

Out[9]:



4.5　適合度指標と情報量規準（[4]の解答）

共分散構造分析の出力結果に対しては、データと比較して整合性があるかを数値で表現した**適合度指標**が計算できます。適合度指標が悪い場合は、パス図が妥当ではなかった可能性があり、仮説を再検討することが一般的です。適合度指標にはさまざまなものがありますが、以下の四つがよく用いられています。

- **TLI**（Tucker-Lewis指標）
- **CFI**（Comparative fit index）
- **GFI**（Goodness of Fit Index）
- **RMSEA**（Root Mean Squared Error of Approximation）

それぞれの適合度指標の定義や詳しい意味については、次回以降に詳しく解説します。TLI、CFI、GFIは 1 に近い数値であるほど、RMSEAは 0 に近い数値であるほど良いとされる統計量です。

情報量規準（Information Criteria）は複数の異なるパス図で共分散構造分析を実施したとき、どの結果を採用するかを検討する際に参考になります。代表的なものに**赤池情報量規準**（Akaike Infomation Criteria, AIC）と**ベイズ情報量規準**（Bayesian Information Criteria, BIC）があります。いずれも値自体に意味があるものではなく、値が小さいものほど良いと考えられる統計量です。

`fit` メソッドを実施したあとのオブジェクト `sem` を `calc_stats` 関数に渡すことで、適合度指標や情報量規準の値を計算することができます。今回、TLI, CFI, GFIの値は 1 に近く、RMSEAの値は 0 に近いため、仮説またはパス図がデータと整合していなかったというわけではなさそうです。

In [10]:
```
# 適合度指標や情報量規準の計算
calc_stats(sem)
```

Out[10]:

	DoF	DoF Baseline	chi2	chi2 p-value	chi2 Baseline	CFI	GFI	AGFI	
Value	6	15	17.250556	0.008405	317.54459	0.962814	0.945675	0.864188	0.9

まとめ

今回は、共分散構造分析について分析に最小限必要な知識を学びました。特に、以下のことについて勉強しました。

- 共分散構造分析では、データとパス図を入力して、パス係数の推定や仮説検定を実施する。
- パス図とは、分析者が考える変数間の仮説を表現するグラフのことをいう。
- 観測変数とは、データに値が記録されている変数のこと。そうでない変数を潜在変数という。パス図では観測変数が四角、潜在変数が丸で表現される。
- 値の大小関係が類似している関係を相関関係といい、変数間の説明関係を回帰関係という。パス図では相関関係が双方向の矢印、回帰関係が片方の矢印で表現される。
- パス図上で変数間の関係の強さを表す数値をパス係数という。
- パス係数には標準化解と非標準化解がある。変数をすべて標準化した場合のパス係数が標準化解、そうでない場合のパス係数が非標準化解。解釈によって使い分ける必要がある。
- 共分散構造分析の分析の流れと `semopy` パッケージの使い方を学んだ。