因果推論

NRIデジタル株式会社
データサイエンティスト

2020年08月08日

Share the Next Values!

ガイダンス

■ 以下の講義内容で進めていきます。
- 主には基礎の考え方を中心に、プログラミングで実演していきます。

■ 質疑応答は、チャットなどでいつでも受け付けます。
- １コマ目：10時から11時半まで
  - 統計分析の復習
- ２コマ目：12時半から14時まで
  - 相関と因果の考え方
  - 効果量の測定方法
    - DIDの考え方
- ３コマ目：14時半から16時まで
  - バイアスの除去方法
  - 傾向スコアマッチング
  - IPV法
  - 効果量の測定方法
- ４コマ目：16時半から18時まで
  - 検証実験の設計
  - 検出力解析
  - 効果検証データの解析
  - 検証結果の妥当性

データの可視化

データの可視化

■ 因果推論に関わる統計解析とデータ操作の復習を行う。

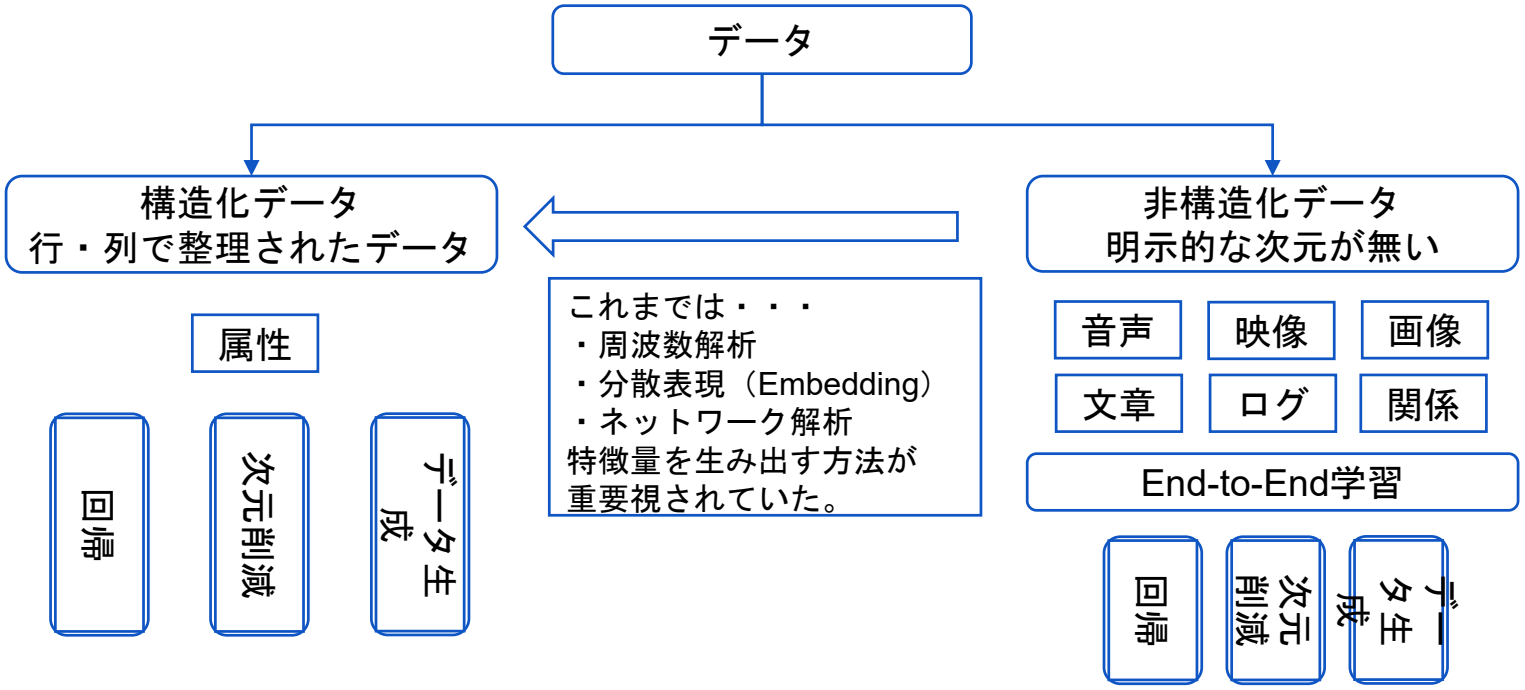
- あくまで因果推論に関わる部分をピンポイントで抜き出したので、個々で補完してください。
- データの可視化
- 回帰分析の復習
- 相関関係
- 仮説検定

■ Pandasを主に利用し、データの取り扱い・統計の基礎をおさらい。

データの種類

■ 主には構造データを利用するが、非構造データにおける因果推論方法も多くある。

- 本講義では、表形式のデータ。単純なテーブルデータやパネルデータ、時系列。
- 統計で利用される回帰分析の多くは、構造化データ＋時系列データを対象とする。



データの取り扱い

■ Pandasでのデータを取り扱いを行う。
- JupyterNotebookを用意できる方は、準備お願いします。

■ データの所在
- https://github.com/kabigon-ds/uec_de

■ JupyterNoteBookの所在
- GitHub、ドライブにデータをアップロード。

利用するデータの説明

■ Stataが配布するサンプルデータセットから選定

- Kielmc.csv　　：データ本体
- Kielmc.txt　　：データ定義

■ ゴミ焼却場の建設が住宅価格に与える影響

- 1978年と1981年の2時点のデータ同一の住宅ではない。

■ ゴミ焼却場の建設の噂が立ち、住宅価格に影響があったかどうかを調べられる

- 1978年時点　→うわさはなかった
- 1981年時点　→うわさがあった

■ 焼却場の近隣の住宅かどうかのフラグがあり、影響を受けたか受けなかったかの比較ができる。

利用するデータの説明

■ 1978年と1981年の321のデータ（179と142）がある。

1. year	1978 or 1981
2. age	age of house
3. agesq	age^2
4. nbh	neighborhood #, 1 to 6
5. cbd	dist. to central bus. dstrct, feet
6. intst	dist. to interstate, feet
7. lintst	log(intst)
8. price	selling price
9. rooms	# rooms in house
10. area	square footage of house
11. land	square footage lot
12. baths	# bathrooms
13. dist	dist. from house to incinerator, feet
14. ldist	log(dist)
15. wind	perc. time wind incin. to house
16. lprice	log(price)
17. y81	=1 if year == 1981
18. larea	log(area)
19. lland	log(land)
20. y81ldist	y81*ldist
21. lintstsq	lintst^2
22. nearinc	=1 if dist <= 15840
23. y81nrinc	y81*nearinc
24. rprice	price, 1978 dollars
25. lrprice	log(rprice)

データの取り扱い

■ データ操作する上で必要なこと

- 読み込み・前処理
- 可視化（ヒストグラム、グラフ）
- 分析のための前処理
- 基礎集計

データの読み込み

■ 読み込み・前処理

- Pandasでの操作
- データの型
  - カテゴリ変数、連続変数 ：　　astype()

データの表示・可視化

■ 可視化（ヒストグラム、グラフ）

- Seabornによる可視化
- ペアプロットの表示
- 探索的データ分析によるデータの概観をつかむ
  - 関係性の「たかそう」なところを知る。

データの前処理

- 前処理
  - 欠損など

データの前処理

■ 基礎集計

- 説明変数ごとの平均・標準偏差などの要約統計量 ： describe()

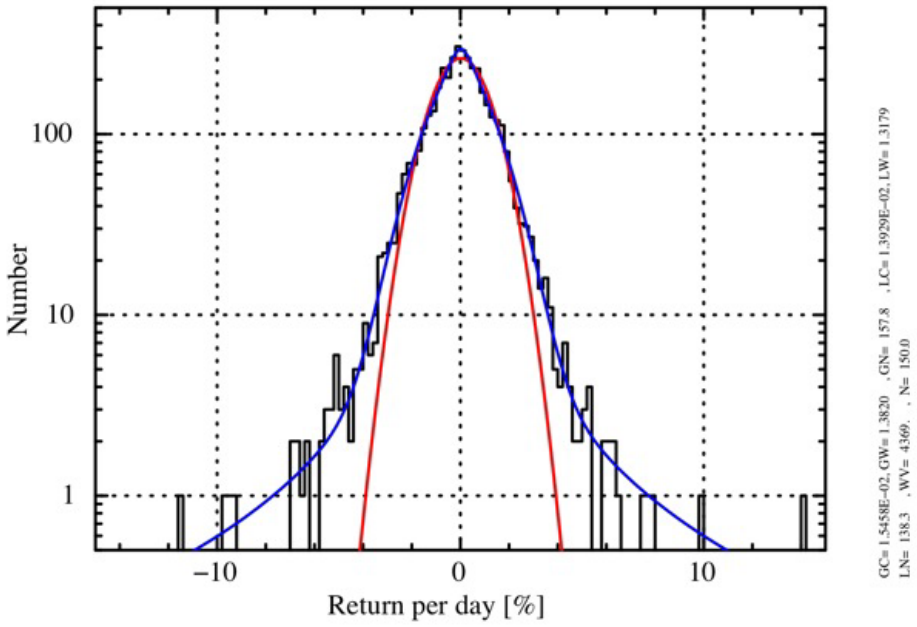
データの基礎集計

■ 散布図の概要図。

データの基礎集計

■ 分布・散布図や平均・標準偏差を確認する意義。

- 不偏量（サンプルサイズによらない集団を特徴づける統計量）となるか。
- べき関数分布であるかどうかは、後の相関や回帰分析の正確性にもかかわるので確認する。
  - 為替価格差の分布、個人所得の分布、文章中の単語の頻度分布など



- コーシー分布 (安定分布のひとつ)
- 安定分布
- 逆ガンマ分布
- べき分布 (パレート分布)
- ユール・サイモン分布
- スチューデントの t 分布
- ゼータ分布
- 逆正規分布の一部
- レートディストリビューション (比分布) の一部
- タリス分布 (t 分布)
- レヴィ分布 (安定分布のひとつ)

データの基礎集計

■ 分布・散布図や平均・標準偏差を確認する意義。

- 不偏量（サンプルサイズによらない集団を特徴づける統計量）となるか。
- べき関数分布であるかどうかは、後の相関や回帰分析の正確性にもかかわるので確認する。
  - 為替価格差の分布、個人所得の分布、文章中の単語の頻度分布など

たとえば、これらのべき関数分布は、次の関数に近似される。

$$f(x) = \alpha x^{-(\alpha+1)} \quad (x > 1)$$

平均や分散を求めると

$$\mu = \int_1^\infty x \cdot \alpha x^{-(\alpha+1)}\, dx = \frac{\alpha}{\alpha-1} \quad (\alpha > 1)$$

$$\sigma^2 = \int_1^\infty x^2 \cdot \alpha x^{-(\alpha+1)}\, dx = \frac{\alpha}{(\alpha-1)^2(\alpha-2)} \quad (\alpha > 2)$$

コーシー分布は、$\alpha = 2$のときに相当。平均や分散は存在しない

$$f(x) = \frac{1}{2\pi} \cdot \frac{1}{1+x^2} \sim |x|^{-2} \quad (x \gg 1)$$

- コーシー分布 (安定分布のひとつ)
- 安定分布
- 逆ガンマ分布
- べき分布 (パレート分布)
- ユール・サイモン分布
- スチューデントの t 分布
- ゼータ分布
- 逆正規分布の一部
- レートディストリビューション (比分布) の一部
- タリス分布 (t 分布)
- レヴィ分布 (安定分布のひとつ)

関係を調べる

関係を調べる

■ これまでは、ひとつの変数に注目した。変数間の関係を記述することで、データの理解が広がる。

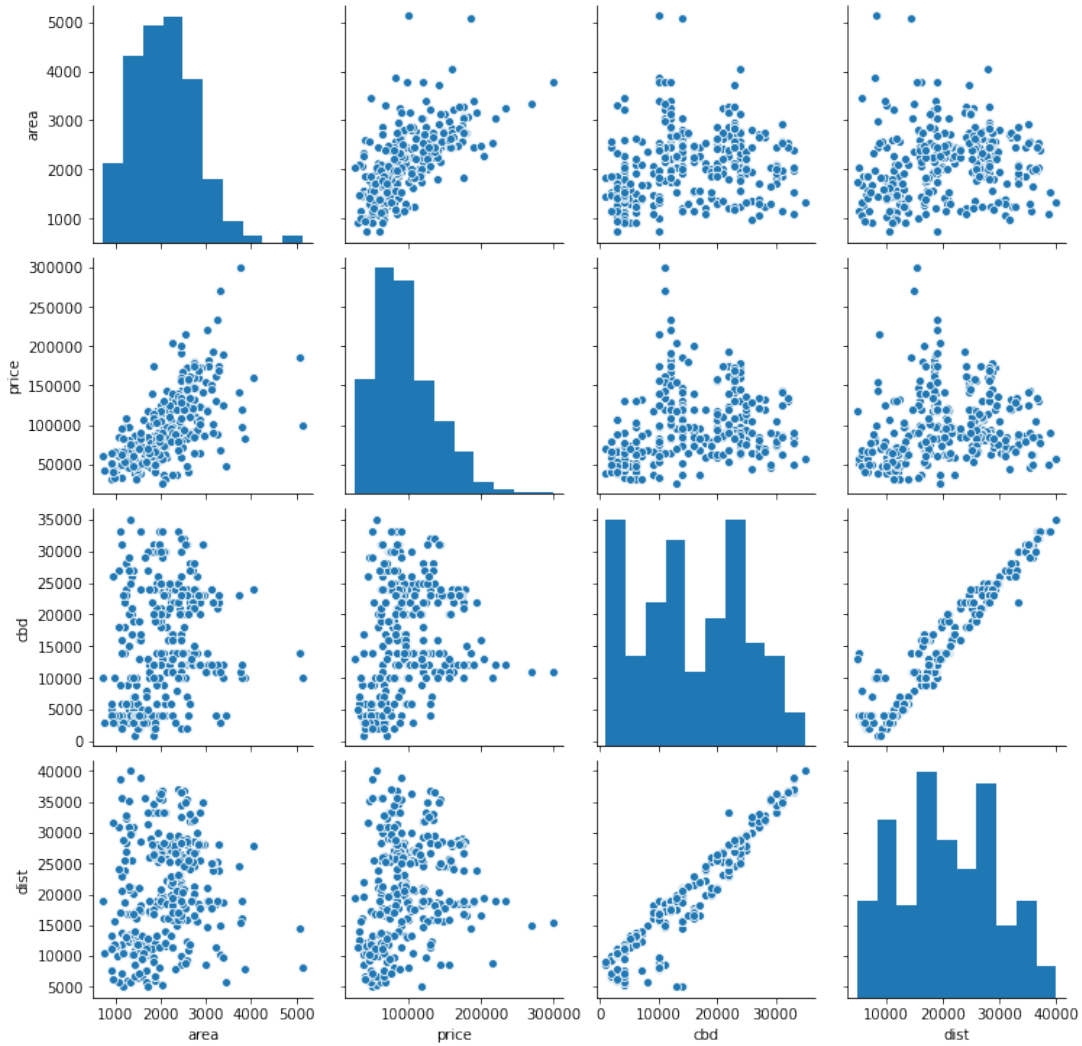
■ そのような関係の定量化に、代表的なものに次のようなものがある。

- 相関関係
- 偏相関関係

■ これらは因果関係につながる重要な概念。

相関関係

■ 変数間の散布図を作ると、似たような傾向を示すものが散見される。



相関関係

■ピアソンの積率相関係数

- もっともよく利用される相関係数。
  - $r > 0$のとき正の相関：一方が増加すると、もう一方の変数も増加する傾向がある。
  - $r < 0$のとき負の相関：一方が増加すると、もう一方の変数は減少する傾向がある。
  - $r < 0$のとき無相関。
- ただし、直線関係でないものや、平均・分散が存在しない変数はこの相関係数では測ることができないという特徴がある.

  この場合は...
  - 確率変数の変換（主に対数変換）後に，積率相関係数を用いる.
  - 順位相関係数を用いる.

$$r = \frac{\sum_{i=1}^{n}(x_i - E[x])(y_i - E[y])}{\sqrt{\sum_{i=1}^{n}(y_i - E[y])^2}\sqrt{\sum_{i=1}^{n}(x_i - E[x])^2}}$$

- Pandasでは、データフレームに以下のメソッドを呼ぶと、変数間の相関係数のペアが出力される。
  - > df.corr()

相関関係

■ スピアマンの順位相関

- ケンドールの順位相関係数もあるが、ここでは割愛
- データに正規性がない場合や非線形の関係があることが自明な時に利用。

数値データ

X	Y
10	15
33	28
54	21
12	18
28	45

順位化データ

X	Y
5	5
2	2
1	3
4	4
3	1

順位化したデータで前述の相関係数を計算する

$$r = \frac{\sum_{i=1}^{n}(x_i - E[x])(y_i - E[y])}{\sqrt{\sum_{i=1}^{n}(y_i - E[y])^2}\sqrt{\sum_{i=1}^{n}(x_i - E[x])^2}}$$

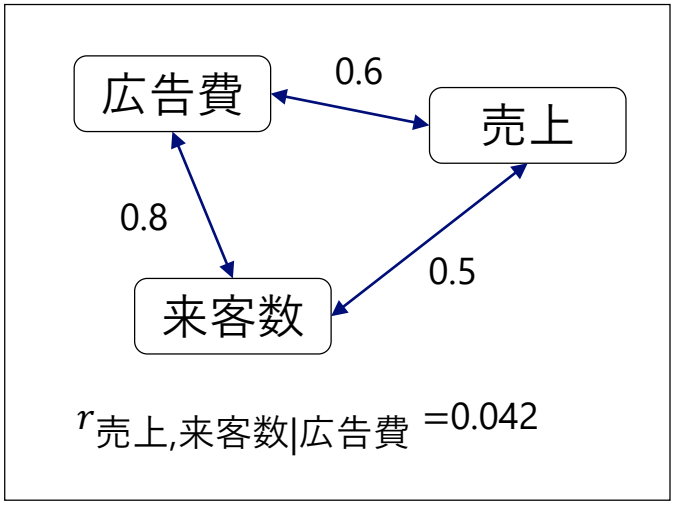
- Pandasでは、データフレームdfに以下のメソッドを呼ぶと、変数間の相関係数のペアが出力される。
  - > df.corr(method = 'spearman)

（参考）偏相関関係

■ 偏相関関係

- ３変数以上のデータにおいて，２変数だけに根差す本質的な相関を取り出す。
- ある２変数に注目して、他の変数の影響を固定するという考え方。
- 例として、A、B、Cの変数に対して、相関係数をr_{AB}、r_{BC}、r_{AC}とするとCの影響を取り除いた偏相関係数$r_{AB|C}$は次の式で求まる。

$$r_{AB|C} = \frac{E[(A_i^c - E[A_i^c])(B_i^c - E[B_i^c])]}{\sqrt{E[(A_i^c - E[A_i^c])^2]}\sqrt{E[(B_i^c - E[B_i^c])^2]}}$$
$$= \frac{r_{AB} - r_{AC}r_{BC}}{\sqrt{1-r_{AC}^2}\sqrt{1-r_{BC}^2}}$$



- Pandas自体には偏相関を計算するメソッドはないようなので、他のライブラリや自作する必要がある。

```
import pingouin as pg
df = pg.read_dataset('partial_corr’)
pg.partial_corr(data=df, x='x', y='y', covar='cv1').round(3)
```

（参考）偏相関関係

■ 偏相関関係

- ある２変数i, j以外の残りの全変数の集合Resを固定した場合の偏相関係数$r_{ij|Res}$を求める式
- r^{ij}は相関行列Rの逆行列の(i, j)成分

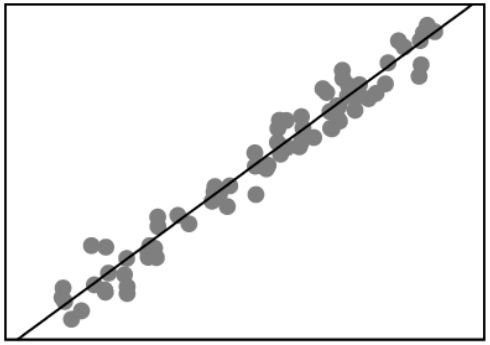
$$r_{ij|Res} = -\frac{r^{ij}}{\sqrt{r^{ii}r^{jj}}}$$

回帰分析

回帰分析

■ これまでは、ある二つの変数間の関係の強さを数値化するもの。
- ここでは、変数間の関係をモデル化する。回帰・フィッティングとも呼ばれる。

直線関係でモデル化



■ 線形回帰と非線形回帰における代表的な方法
- 重回帰（単回帰）
- ロジスティック回帰

■ 回帰(Regression)
- ナノシート収率や株価予測のように，連続する数値データの予測・フィッティング問題

■ 分類(Classify)
- 回帰とは違い具体的な数字を出すのではなく，与えられたクラスに分ける（ラベリングする）ことを目的とした問題.

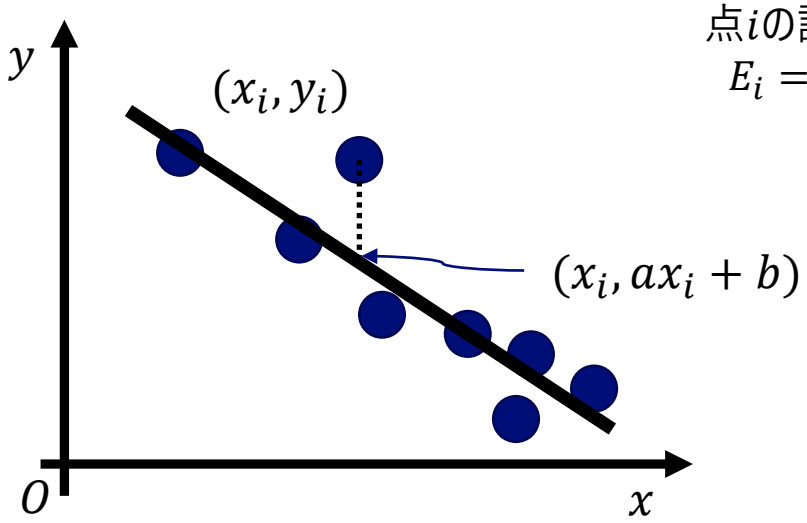
■ あくまで因果推論の基礎概念の中で利用する部分のみ復習する。

回帰分析

■ 単回帰分析

- ある**２つの変数**間の間を、「直線」の関係でモデル化する。
- 仮定した数式とデータ点との誤差が最小になるように、モデルのパラメータを調整する。

①： n個のデータ点(x_i, y_i)を用意する
②： データ点が沿う数式を仮定する　：　$y = f(x) = ax + b$
③： 数式とデータ点の２乗誤差Eを計算する
④： Eを最小化するよう$f(x)$のパラメータを決定する　：　$\frac{\partial E(a,b)}{\partial a} = 0, \frac{\partial E(a,b)}{\partial b} = 0$となる$a, b$



点iの誤差
$E_i = (y_i - ax_i - b)^2$

データ点全体と理論点の誤差
$E(a,b) = \sum(y_i - ax_i - b)^2$

$$a = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}$$

$$b = \frac{n\sum x_i^2 \sum y_i - \sum x_i y_i \sum x_i}{n\sum x_i^2 - (\sum x_i)^2}$$

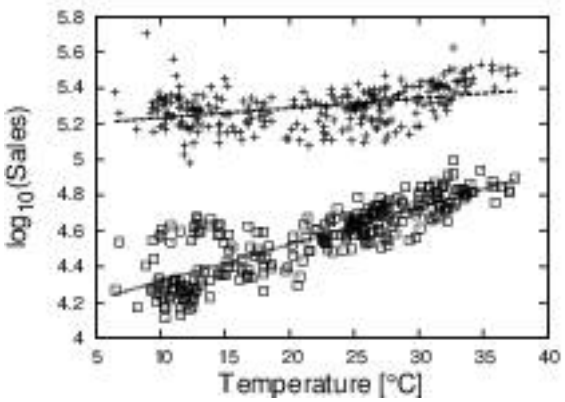
※一般の$f(x)$についても、同様の方法でできる。Eを最小化する際に工夫が必要となる

回帰分析

■ 重回帰分析

- ２つ以上の変数（多変量の説明変数）を用いて目的変数を説明する場合，重回帰分析を用いる。
- 考え方は同じだが、説明変数間が互いに独立であるという仮定が新しく必要となる。
  - そのため、基礎集計を通して、相関関係を見ることが重要となる。

日付	アイス売上 [円]	気温 [℃]	来店者数 [人]
1 月 14 日	197880	9.7	140500
1 月 15 日	154080	11.6	141565
1 月 16 日	188900	12.1	142530
1 月 17 日	147160	8.2	139060
	⋮	⋮	⋮



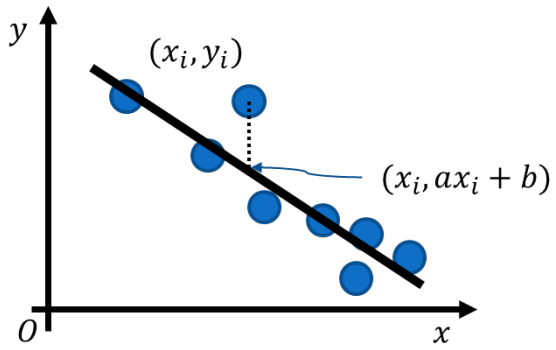
- $\log_{10}\{\text{アイスの売上}\} = \beta_0 + \beta_1 \cdot \text{気温} + \beta_2 \cdot \text{来店者数} + \epsilon$ でモデル化を考える。

回帰分析

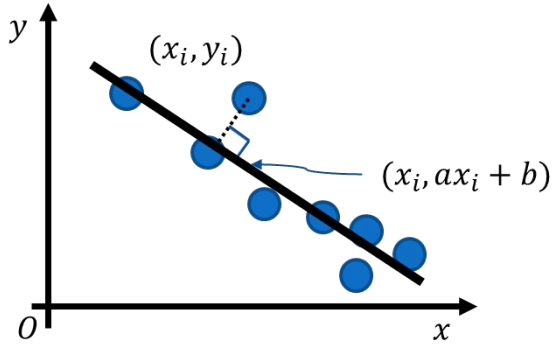
■ 重回帰分析で注意すること

- 変数間に強い相関があると**多重共線性**を起こし、パラメータが求まらなくなってしまう。
- 分散が定義できないベキ分布などには適切でない。
- **片軸を真値だと仮定している解析(片軸を特別視している。因果の方向まで仮定)**

単回帰（x軸は真値、y軸には誤差）



直交回帰（モデル式に対する垂線）



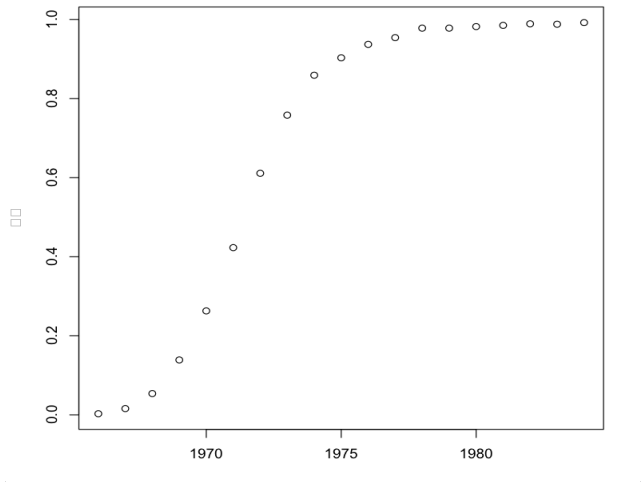
特に、モデル式が強い曲率を持った曲線だと 2 方法のずれは大きい

回帰分析

■ 非線形回帰分析

- 「線形モデル」の「線形」の意味は「目的変数」と「説明変数」の関係性の間に直線の関係があることが仮定されている。
- 非線形の場合では、ひとつひとつ関係を見ていき、適切な関係式を仮定し、フィッティングを行っていく。

非線形な関係



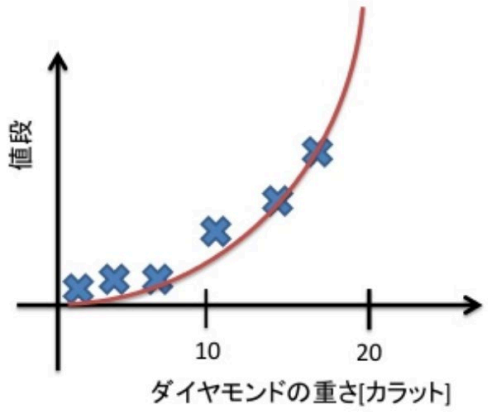
例えば指数・対数関係やべき乗の関係性が入ると,
途端に仮定が崩れて「線形モデル」を使えない。

回帰分析

■ 非線形回帰分析（ロジスティック回帰）の例

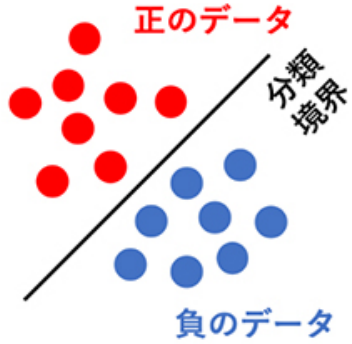
- S字型の曲線で、データをフィッティングする。
- 「買う、買わない」や「契約、解約」といった **2 値を推定**するために行われる。

回帰



分類

正のデータと負のデータからの分類



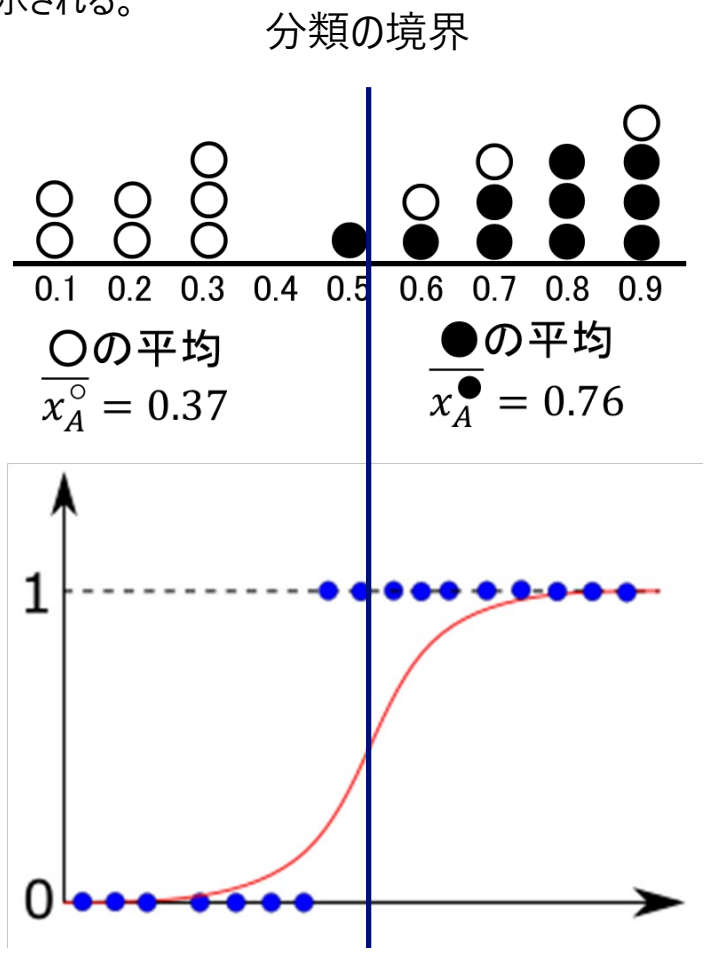
回帰分析

■ロジスティック回帰の理論

- ロジスティック関数でのフィッティングによって、データiが属すクラスの確率が示される。

$$p_i = \frac{1}{1 + e^{-(a_0 + a_1 x_{1,i} + \cdots + a_n x_{n,i})}}$$

- パラメータ求値については割愛。最尤法で行うことがおおい。
- 傾向スコアマッチングの部分で再登場。



回帰分析のまとめ

■ 回帰分析で気を付けること

- 説明変数の正規性
  - べき性が強い場合：対数変換
- 説明変数との非線形関係：20～39歳の年齢層だけ効き、他の年代で効かないなどの傾向が散布図等から分かった場合
  - 階級化、カテゴリ変数

年齢		年齢_0～19	年齢_20～39	年齢_40～59
34歳	⇒	0	1	0

■ 回帰分析自体が間違ってしまうと、誤った因果関係を読み取ってしまうこともある。

■ 簡単な重回帰分析でも、以上のような変数間の関係を緻密に調べ、階級化やカテゴリ化することで精度が出る。

- 売上高と宣伝費の関係が分かっていれば→目標とする売上高に対して宣伝費を決定する（制御）
- 人口と商店数の関係が分かっていれば→ある市の人口からその市の商店数を予測する（予測）

仮説検定

仮説検定

■ これまでの基礎集計やモデル化の結果が、（統計の意味で）正しいと言えるかを検証すること

- 実際の現場の感覚と適合するかどうかとは別問題

■ 因果推論では、２つのグループの比較を行うことで、効果を検証することが多いため、これらを復習した。

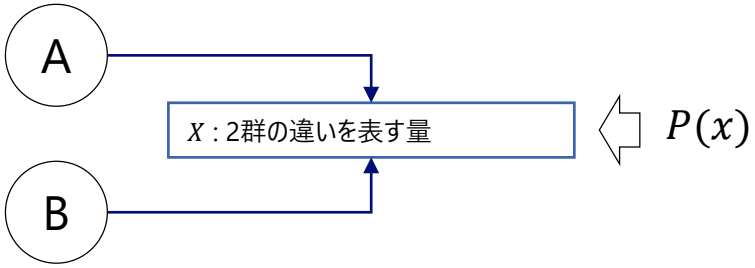
仮説検定の手順

■ 仮説検定の考え方。

- 帰無仮説/対立仮説を設定する。

帰無仮説	対立仮説
A＝B	A≠B

- データから適切な検定統計量Xを定める。
- その検定統計量がある確率分布$P(x)$に従うと仮定。



- 検定統計量Xが、その確率分布$P(x)$上での起こり得る確率を求める。

$$P(x=X) < \alpha \ or \ not \ ?$$

- その確率が小さいと帰無仮説が棄却。対立仮説が採択される。

（参考）仮説検定のチートシート

■ 検証したいものを定めて、どのような条件のデータかをしることで、仮説検定方法を決めることができる。

差／相関	比較データ間の対応性	変数の種類	正規性	比較する群の数	サンプル数	適切な統計手法
差	対応なし	連続変数	正規分布	2	総数 30 以上	スチューデントの t 検定
				＞2	1 群 15 以上	一元配置分散分析
		連続変数／順序変数	非正規分布（連続変数）	2	制限なし	マン・ホイットニーの U 検定* ウィルコクスンの順位和検定*
				＞2	制限なし	クラスカル・ウォリス検定*
		2 値変数		2	総数 20 未満	フィッシャーの正確確率検定*
				≧2	総数 20 以上	ピアソンのカイ 2 乗検定
		打ち切り例のある 2 値変数		≧2	イベント総数 10 以上	ログランク検定
	対応あり	連続変数	正規分布	2	15 組以上	対応のある t 検定
				＞2	15 組以上	反復検定による分散分析
		連続変数／順序変数	非正規分布（連続変数）	2	制限なし	ウィルコクソンの符号順位検定*
				＞2	制限なし	フリードマン検定*
		2 値変数		2	制限なし	マクネマー検定
相関（関連性）		連続変数	正規分布		総数 20 以上	ピアソンの相関係数
		連続変数／順序変数	非正規分布（連続変数）		制限なし	スピアマンの順位相関係数*
		2 値変数			制限なし	ケンドールの順位相関係数* カッパの相関係数（一致性）

http://www.igaku-shoin.co.jp/paperDetail.do?id=PA02927_03

因果推論で主に利用される検定

■ ２つのグループ間の比較のため、差の比較が中心となる。

- 平均の差を例とした検定
  - **対応のないt検定**、**Wilcoxonの順位和検定**

■ t検定には，さまざまな仮定が伴う

- 各群の標本が，いずれも正規母集団から得られたものである（正規性）
- 各群の母分散が等しいこと（等分散性）

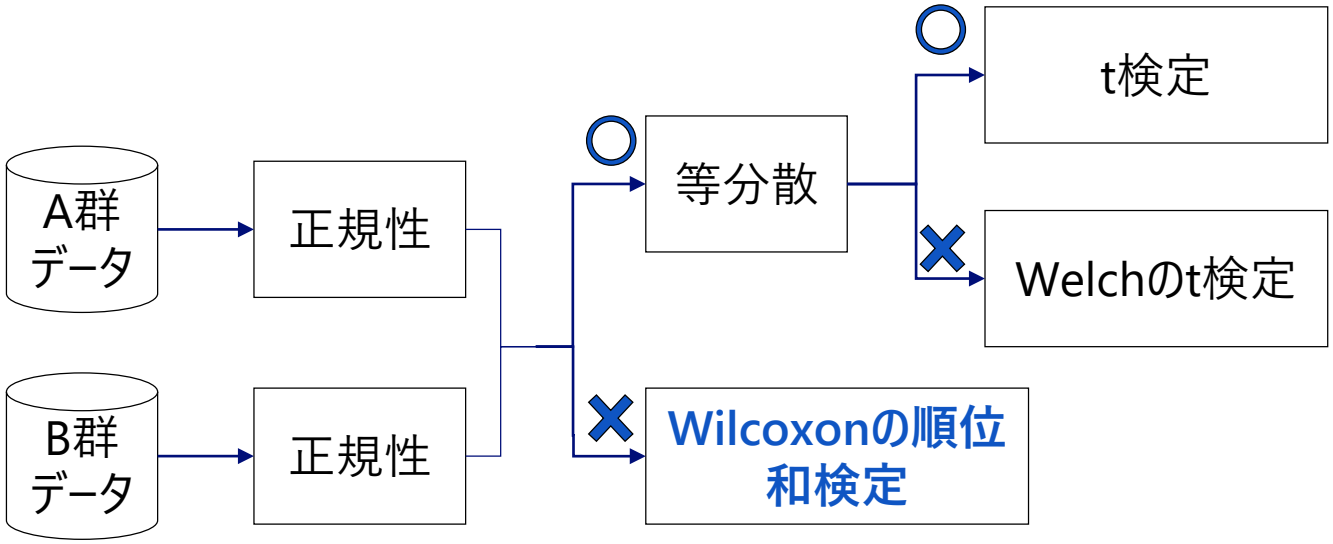
→そのために，正規性と等分散性について，それぞれ異なる検定を行う必要がある

■ 正規性が仮定できないものは、Wilcoxonの順位和検定を行う。

仮説検定

■ ２つのグループの違いを検定するときの確認の流れ

- データ数が十分な数（N＞50）があるとする。
- 2群に対応のある場合とない場合も注視する。
- データの性質を確認し、適切な検定方法を選ぶ。

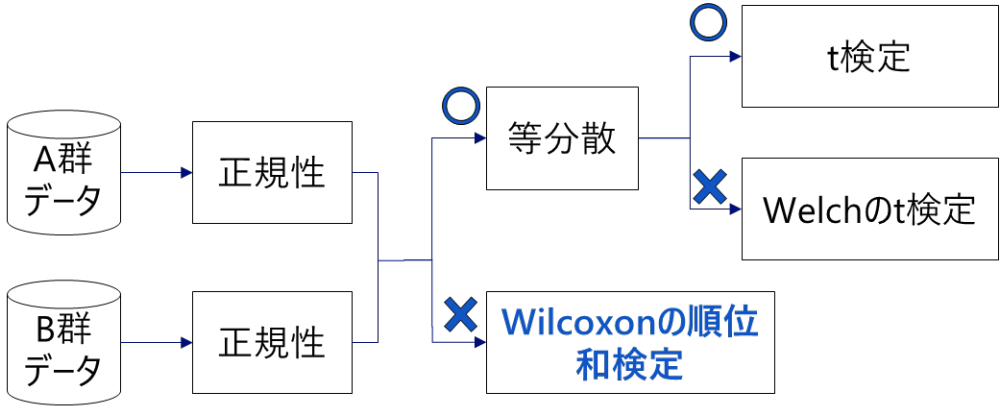


仮説検定

■ 求める検定方法に行きつくために、次の検定を行う。

■ 正規性の検定
- **コルモゴロフ–スミルノフ検定**
- シャピロ-ウィルク検定

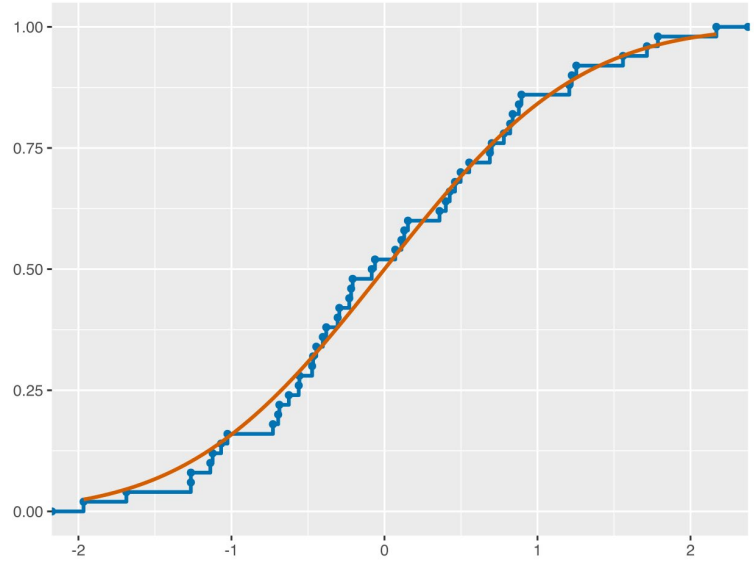
■ 等分散性の検定
- **F検定**



コルモゴロフ–スミルノフ検定

■ 二つの分布が同じかどうかを検定する。

- 1 標本KS 検定は、データから得られた経験分布を帰無仮説において示された累積分布関数と比較する手法
- 経験分布　　$F_N(x) = \frac{\#(X_i > x)}{N}$　：右図データ点
- 帰無仮説の分布　$F(x)$　：標準正規分布の累積分布



- 2 つの分布の差が理論的に求まる
  - $z = \max|F_N(x) - F(x)|$
  - 最大値z はデータ数N の平方根との積$x = z\sqrt{N}$
- $P(> x) = 2\sum_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2x^2}$

- 有意水準を設定し、P(>x)の基づいて、帰無仮説の棄却を判断する

F検定

■ 二つのデータの分散が同じかどうかを検定する。

- 正規分布に従う2つの群の「標準偏差が等しい」という帰無仮説の検定

- 正規性を持つ2つの群のデータの2乗値：$x_1^2, x_2^2, \cdots, x_n^2$ 、 $y_1^2, y_2^2, \cdots, y_m^2$
- 検定統計量　$f = \frac{(\sum_{i=1}^{n} x_i^2)/n}{(\sum_{i=1}^{m} y_i^2)/m}$ が　F分布に従うことが分かっている。（正確には不偏分散）

- このF分布の自由度を（n-1,m-1）として設定した有意水準で検定。

仮説検定

■ t検定

- 等分散性などを事前に検定する必要がある。
- 「2群の代表値（平均）は同じである」という帰無仮説のもと、2群の比較を行う。
- 検定統計量 $t = \frac{|\overline{x_1} - \overline{x_2}|}{\sqrt{s^2(\frac{1}{n}+\frac{1}{m})}}$ 不偏分散$s = \frac{(n-1)s_n+(m-1)s_m}{n+m-2}$
- n+m-2の自由度のt分布に従うことを利用して、検定を行う。

■ Welchのt検定

- 分散が異なる場合でも、「似た」方法で行うことができる。
- 実際には群間の性質が異なることも多いので、Welchを使うことも多い。
- $t = \frac{|\overline{x_1} - \overline{x_2}|}{\sqrt{(\frac{S_n}{n}+\frac{S_m}{m})}}$ として、自由度$\frac{\left(\frac{S_n}{n}+\frac{S_m}{m}\right)^2}{\frac{s_n^2}{n^2(n-1)}+\frac{s_m^2}{m^2(m-1)}}$のt分布を設定して同様に行う。

仮説検定

■ Wilcoxonの順位和検定

- 正規性がないため、順位に変換して行う方法
- 「2群の代表値（中央値）は同じである」という帰無仮説のもと、2群の比較を行う。

- 2群のデータ($n < m$)を混ぜて、昇順の順位をつける。
- X：$x_1, x_2, \cdots, x_n$
- Y：$y_1, y_2, \cdots, y_m$

$$x_1, y_1, y_2, x_2, y_3, \cdots, x_n \quad \rightarrow \quad 1, 2, 3, 4, 5, \cdots, n+m$$

- サンプル数の少ない群の順位和を求める　$T = \sum_i r(x_i)$
- この検定統計量Tと有意水準のもと、順位和表に記載される$W_{n,m}$を比較して、$T < W_{n,m}$であることで検定を行う。

まとめ

講義１のまとめ

■ 因果推論に関わる、以下の内容の要点を復習しました。

■ 講義２以降、これらの考え方を利用していきますが、都度振り返る質問等は歓迎です。

■ データの操作
- Pandasの利用
- データの読み込み、操作

■ 基礎統計
- 相関
- 回帰
- 仮説検定


Share the Next Values!