線形回帰（１）

正田 備也

masada@rikkyo.ac.jp

統計モデルの3つの使い途

https://www.stat.berkeley.edu/~aldous/157/Papers/shmueli.pdf

Explanatory Model (例: 因果推論)

test/quantify causal effect between constructs for “average” unit in population

Descriptive Model (例: 線形回帰における検定)

test/quantify distribution or correlation structure for measured “average” unit in population

Predictive Model (例: 機械学習による予測)

predict values for new/future individual units

統計モデルをdescriptive modelとして使う

線形回帰モデルの場合。

線形回帰における検定

- 説明変数の目的変数に対する影響が有意か調べたい
- 帰無仮説：特定のcoefficientまたはinterceptがゼロ
- 上記の帰無仮説が棄却できるか？という検定
  - 今日の参考資料：大阪大学「計量経済基礎」（谷崎先生）の講義資料
    - http://www2.econ.osaka-u.ac.jp/~tanizaki/class/2018/basic_econome/02.pdf

モデルの仮定（単回帰の場合）

$$y_i = b + ax_i + u_i$$

1. x_iは固定された値をとると仮定
2. すべてのiについて、誤差項u_iの期待値は0と仮定
3. すべてのiについて、誤差項u_iの分散はσ^2と仮定
4. すべてのi, jについて、誤差項u_iとu_jが無相関（$E[u_i u_j] = 0$）と仮定
5. すべてのiについて、誤差項u_iは平均0、分散σ^2の正規分布に従うと仮定
6. $N \to \infty$のとき、$\sum_{i=1}^{N}(x_i - \bar{x})^2 \to \infty$と仮定(ただし$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$)

多変量正規分布では、無相関⇔独立

- 例えば下記Webページを参照

  https://mathtrain.jp/uncorrelated

- 多変量正規分布では、無相関であることと、独立であることとは、同値
- よって、前のスライドの仮定から、誤差項 u_i は、すべて独立に、平均0、分散σ^2の正規分布に従うことが言える

補足：誤差項の正規性の仮定は必要か

- 検定をしたいなら、誤差が正規分布に従うという仮定は必要
- そうでないなら、不要
- 「線形回帰の仮定の誤解について」
  - https://communities.sas.com/t5/Blog/%E7%B7%9A%E5%BD%A2%E5%9B%9E%E5%B8%B0%E3%81%AE%E4%BB%AE%E5%AE%9A%E3%81%AE%E8%AA%A4%E8%A7%A3%E3%81%AB%E3%81%A4%E3%81%84%E3%81%A6/ba-p/495164

補足：回帰診断(regression diagnostic)

- statsmodels
  - https://www.statsmodels.org/dev/examples/notebooks/generated/regression_diagnostics.html

単回帰の正規方程式

- 以下の連立方程式を解けば、傾きaの最小二乗推定量$\hat{a}$と、切片bの最小二乗推定量$\hat{b}$が求まる

$$\begin{bmatrix} \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} x_i y_i \end{bmatrix} = \begin{bmatrix} N & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} {x_i}^2 \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{a} \end{bmatrix}$$

正規方程式を解くと…

$\hat{a}$も$\hat{b}$も、
正規分布に従う
独立な確率変数u_iの
線形和になっている

$$\hat{a} = a + \sum_{i=1}^{N} \omega_i u_i$$

$$\hat{b} = b - (\hat{a} - a)\bar{x} + \frac{1}{N}\sum_{i=1}^{N} u_i$$

ただし、 $\bar{x}$は$x_1, \ldots, x_N$の平均。 ω_iは下記のとおり。

$$\omega_i = \frac{(x_i - \bar{x})}{\sum_{j=1}^{N}(x_j - \bar{x})^2}$$

単回帰における検定(1/4)

- モデルの仮定より、以下のことが証明できる
  - 傾きaの最小二乗推定量$\hat{a}$は不偏推定量、つまり$E(\hat{a}) = a$
  - $\hat{a}$の分散は、$V(\hat{a}) = \frac{\sigma^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$
  - 切片bの最小二乗推定量$\hat{b}$は不偏推定量、つまり$E(\hat{b}) = b$
  - $\hat{b}$の分散は、$V(\hat{b}) = \frac{\sigma^2 \sum_{i=1}^{N} {x_i}^2}{N \sum_{i=1}^{N}(x_i - \bar{x})^2}$
  - $\hat{a}$と$\hat{b}$の共分散は、$\mathrm{Cov}(\hat{a}, \hat{b}) = -\frac{\sigma^2 \sum_{i=1}^{N} \bar{x}}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$
- ここまでは、誤差の正規性の仮定は使っていない

単回帰における検定(2/4)

- $\hat{a}$も$\hat{b}$も、正規分布に従うことが示せる
  - 各u_iは独立に正規分布に従う
  - $\hat{a}$も$\hat{b}$も、 u_iの線形和で表される
  - 正規分布にしたがう確率変数の和は、正規分布に従う
    - http://joe.bayesnet.org/?p=4950
  - 以上のことから、 $\hat{a}$も$\hat{b}$も正規分布に従うことが言える

単回帰における検定(3/4)

- $\hat{a}$の分散の式も、$\hat{b}$の分散の式も、未知の値σを含んでいる
  - どうする？
- 以下のs^2が、誤差項の分散σ^2の不偏推定量となることが言える

$$s^2 = \frac{1}{N-2}\sum_{i=1}^{N}\left(y_i - \hat{b} - \hat{a}x_i\right)^2$$

- さらに$\frac{(N-2)s^2}{\sigma^2}$が自由度$N-2$のカイ自乗分布に従うことも言える

t検定のもとになっている命題(1/3)

命題1

母集団が正規分布$N(m, \sigma^2)$に従うとき、

- その標本$X_1, \ldots, X_N$から求めた標本平均$\bar{X}_N$と不偏標本分散$\bar{V}_N$は、独立である。
- $\bar{X}_N$は、正規分布$N(m, \frac{\sigma^2}{N})$に従う。　($\frac{\bar{X}_N - m}{\sigma}\sqrt{N}$は$N(0,1)$に従う。)
- $\frac{N-1}{\sigma^2}\bar{V}_N$は、自由度$N-1$のカイ二乗分布に従う。

t検定のもとになっている命題(2/3)

命題2

2つの確率変数Z, Yが独立で、Zが標準正規分布$N(0,1)$に従い、Yが自由度nのカイ二乗分布に従うならば、$T \equiv \frac{Z}{\sqrt{Y/n}}$は、自由度$n$の$t$分布に従う。

t検定のもとになっている命題(3/3)

系

母集団が正規分布$N(m, \sigma^2)$に従うとき、その標本$X_1, \ldots, X_N$から求めた標本平均$\bar{X}_N$と不偏標本分散$\bar{V}_N$とについて、$(\bar{X}_N - m)\sqrt{\frac{N}{\bar{V}_N}}$は自由度$N-1$のt分布に従う。

- ポイント：標本から求めることのできる値（標本平均$\bar{X}_N$と不偏標本分散$\bar{V}_N$）だけを使って、母集団の平均mという未知の量に関する定量的な推定が可能になっている。

t検定の例

- 例：自由度20のt分布の両側5%点は2.0860

$$-2.0860 \leq (\bar{X}_N - m)\sqrt{\frac{N}{\bar{V}_N}} \leq 2.0860$$

  - 帰無仮説が$m = 0$のとき、 $\bar{X}_N\sqrt{\frac{N}{\bar{V}_N}} < -2.0860$または$\bar{X}_N\sqrt{\frac{N}{\bar{V}_N}} > 2.0860$ならば、有意水準5%において帰無仮説を棄却する。

単回帰における検定(4/4)

- 傾きaの最小二乗推定量$\hat{a}$については、$(\hat{a}-a)/s_{\hat{a}}$が自由度$N-2$のt分布に従うことが言える。ただし

$$s_{\hat{a}} = \frac{s}{\sqrt{\sum_{i=1}^{N}(x_i-\bar{x})^2}}$$

- 切片bの最小二乗推定量$\hat{b}$については、$(\hat{b}-b)/s_{\hat{b}}$が自由度$N-2$のt分布に従うことが言える。ただし

$$s_{\hat{b}} = s\sqrt{\frac{\sum_{i=1}^{N}x_i{}^2}{N\sum_{i=1}^{N}(x_i-\bar{x})^2}}$$

単回帰における検定をどう行うか

- 単回帰における最小二乗推定量$\hat{a}$と$\hat{b}$についても、以上の理屈により、t検定をおこなうことができる

- PythonならstatsmodelsのOLSを使えばよい
  - 実行結果に、ちゃんと検定統計量が表示される。

https://www.statsmodels.org/stable/regression.html

https://www.statsmodels.org/stable/generated/statsmodels.regression.linear_model.OLS.fit.html

統計モデルをpredictive modelとして使う

統計モデルを予測に使う（≒機械学習）

- 予測性能が良いのであれば、何が起きても構わない
  - 線形回帰の場合、いくつかの説明変数のp値が大きな値になろうが、予測性能が良いのであれば何の問題もない。
  - 多重共線性も問題にならない。
- モデルが正しいかどうかも、問題にならない
- 特に深層学習の世界では、予測性能の向上だけ考える
  - モデルの解釈性の無さも問題にならない。

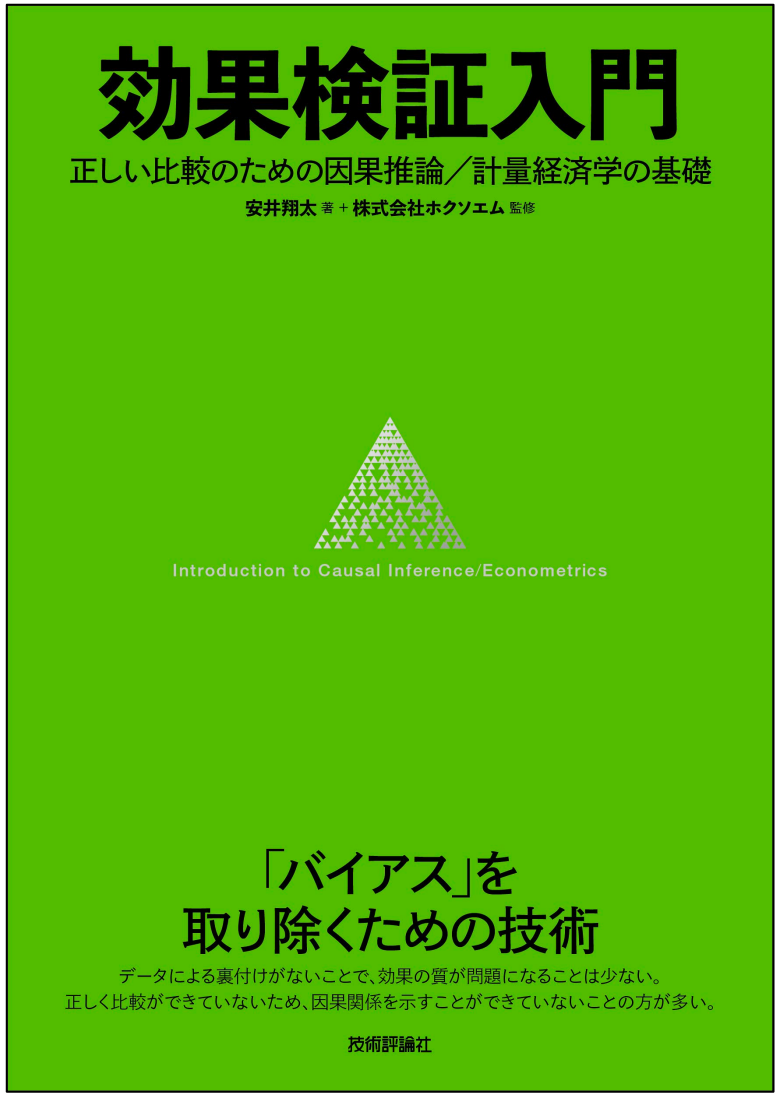
統計モデルをexplanatory modelとして使う

統計モデルを説明に使う（≒因果推論）

「統計的因果効果推定の枠組みを通じて、得られたデータからまずは「ロバストな因果効果の推定」を行い、背後の共変量・交絡要因の影響を排除した「シンプルなメカニズムの理解」を行い”どのような介入が有効か”を示すことが解析実務では有用です。このような一連の解析はシステムに組み込まれていく汎用的機械学習モデルにはできない”データサイエンティストの腕の見せ所”になるのではないでしょうか？」

（星野崇宏「統計的因果効果の基礎」『調査観察データの統計科学-因果推論・選択バイアス・データ融合-』85-86頁）

https://gihyo.jp/dp/ebook/2019/978-4-297-11118-2



https://gihyo.jp/book/2021/978-4-297-12224-9