Problem Set 3. 重回帰分析と統計的検定の実践と解釈

PUBLISHED
July 9, 2024

提出期限

- 2024年 7月 24日（水）17:00

提出形式

- LMSにアップロードされている"problemset3_base"を使用して、(1)レポートと(2)コードをフォルダー構造にしたがって格納し、提出すること
- **提出の際には、フォルダー名を"problemset3_学生番号"に変更して提出する**
  - 例: 学籍番号が"9999999"の学生の場合は以下のフォルダーを提出する
    - "problemset3_9999999"
- レポートはpdf形式で作成し、02_reportに"report3_学籍番号.pdf"という名前で保存すること
- 作成したグラフはレポートに加えること。（**コードを実行しても図は自動的に保存されない**）

フォルダ構造

- problemset3_base
  - 01_code
  - 02_report
  - 99_docs

課題

1. 子の所得の決定要因の回帰分析

目的

1. Rの回帰分析のコマンドに触れ、今後自分自身の研究で使えるように慣れること（基礎として学ぶデフォルトのコマンド lm() では、「正しい」推定をできないことが多い!）
2. 具体例を通じて、重回帰分析における『ガウス・マルコフの定理』の仮定の役割を理解すること

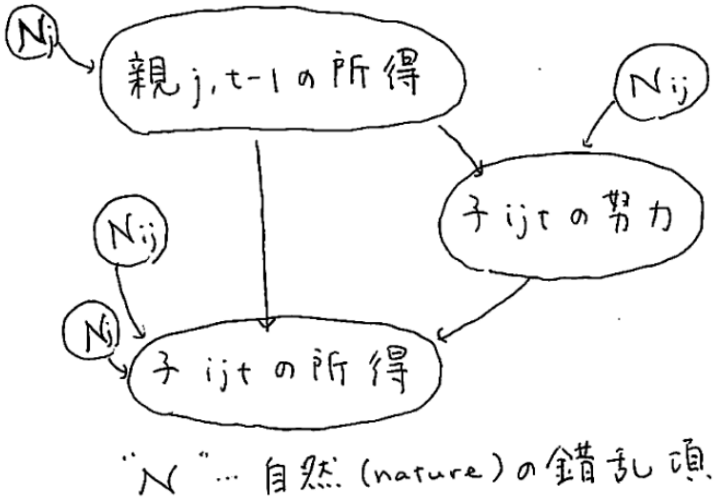
問題設定

『親の所得と子本人の努力が、子の所得にどのような影響があるか』について重回帰分析を用いて考えていきます。累進税や教育などの政策、および社会の公平性を考える上で、とても重要な問題です。

本課題では、500世帯を考えます。それぞれの世帯jには、一人の世帯主(親)と二人の子$i = 1, 2$がおり、子は大学を卒業したばかりですでに働いていて所得を得ているとします。親の変数を$t - 1$でインデックスし、子の変数をtでインデックスします。

Step 1. モンテ・カルロ・シミュレーションによるデータの生成

この問題では、まず所得の決定過程を定め、モンテ・カルロ・シミュレーションでデータをつくります。そして、このデータについて回帰分析を行い、その結果によってどれぐらい、もともと定めた決定過程を正確に知ることができるかを考えます。ここでは、真の所得の決定過程を、以下の因果グラフで表現できるとします。



この因果モデルを、以下の二つの回帰式で考えられるとします。(ただし、ここでlogは常用対数。)

- 子の努力(勉強時間など)が、親の所得に依存する

$$\text{effort}_{ijt} = \alpha_0 + \alpha_1 \log(\text{income}_{j,t-1}) + \text{nature1}_{ijt}$$

なお、$\alpha_0 = 1, \alpha_1 = 0.2$である。

- 子の所得ijtが、親j, t-1 と本人ijtの努力の所得に依存する

$$\log(\text{income}_{ijt}) = \beta_0 + \beta_1 \text{effort}_{ijt} + \beta_2 \log(\text{income}_{ijt-1}) + \text{nature2}_{ijt}$$

なお、$\beta_0 = 0.7, \beta_1 = 1, \beta_2 = 0.15$ である。

ここで、誤差項$\text{nature2}_{ijt} = \nu_{ijt} + \eta_{jt}$とし、$\text{nature1}_{ijt}, \nu_{ijt}, \eta_{jt}$ がすべて正規分布に従うものとする。(因果グラフの図においてN_iとN_ijが"nature"という意味で2回出てきますが、これらは別々の錯乱項を表していると解釈してください。)

備考. 回帰分析を考える前に、この所得決定過程について解釈をしてみましょう。

i. 「勉強していない友達が成績がよくない」ところを見て、ときとして「勉強していないヤツが悪い」と考えてしまいがちです。しかし、その友達の家庭環境などを省みず、どうして「勉強していない」ことがその原因だと言いきれるでしょうか。

ii. ここでは、「親の所得」を「生まれ与えられた環境」として考えることが解釈の上で分かりやすいかもしれません。「親の所得」そのものが「子の所得」に影響を与えるというより、「親の所得」に反映されているIQ(遺伝)や住環境などだと考えてください。「努力」について、「勉強時間」や「集中度」などだと解釈してください。

iii. このように所得の対数を取ると、誤差項に正規分布を仮定できること(すなわち、所得の成長率はもとの所得レベルから独立していること)は、Gibrat's Lawと言い(1931年の発見)、場所と時期を問わずに近似的に成り立つ法則として、所得決定過程に関する経済学研究の基盤となっています。

データの説明

上の所得決定過程に従いシミュレーションされたデータが、 `df_simulation` です。

`df_simulation` には、次の8つの変数が含まれています。

- `household_id`, `sibling_id`
  - 世帯および兄弟姉妹を特定するインデックス番号
- `income_child`, `income_parent`, `effort`
  - 子の所得、親の所得、努力の変数
- `income_child_noisy`, `income_parent_noisy`, `effort_noisy`
  - 測定誤差を含むそれぞれの変数(以下、*をつけて記述する)

ただし、回帰分析に用いることのできるデータには、変数しかなく、パラメータの値は含まれません。これから、シミュレーションをしたデータに基づいて回帰分析をして、パラメータを推定します。 このシミュレーションは、functionである `generate_simulate_df()` に書いてあります。今回は、シミュレーションについてコードを変更する必要はありません。

Step 2. 回帰分析

このdf_simulationについて、以下の6つの回帰式を考えます。変数が欠落したり測定誤差があることは、シミュレーションで自然には起こりませんが、実際の世帯調査において起こらないことはほぼありません。これらのデータの不完全性の影響を考えるために、意図的にシミュレーションでこのような問題を発生させています。

A. 真のモデル. (モデル1)

- モデル1:

$$\log(\text{income}_{ijt}) = \beta_0 + \beta_1 \text{effort}_{ijt} + \beta_2 \log(\text{income}_{ijt-1}) + \varepsilon_{ijt}$$

B1. 欠落変数があるモデル

- モデル2: 「努力」が欠落したモデル

$$\log(\text{income}_{ijt}) = \gamma_0 + \gamma_1 \log(\text{income}_{ijt-1}) + \varepsilon_{ijt}$$

- モデル3: 「親の所得」が欠落したモデル

$$\log(\text{income}_{ijt}) = \delta_0 + \delta_1 \text{effort}_{ijt} + \varepsilon_{ijt}$$

B2. 変数に測定誤差があるモデル

- モデル4: 「子の所得」に測定誤差があるモデル

$$\log(\text{income}^*_{ijt}) = \tilde{\beta}_0 + \tilde{\beta}_1 \text{effort}_{ijt} + \tilde{\beta}_2 \log(\text{income}_{ijt-1}) + \varepsilon_{ijt}$$

- モデル5: 「努力」に測定誤差があるモデル

$$\log(\text{income}_{ijt}) = \tilde{\gamma}_0 + \tilde{\gamma}_1 \text{effort}^*_{ijt} + \tilde{\gamma}_2 \log(\text{income}_{ijt-1}) + \varepsilon_{ijt}$$

- モデル6: 「親の所得」に測定誤差があるモデル

$$\log(\text{income}_{ijt}) = \tilde{\delta}_0 + \tilde{\delta}_1 \text{effort}_{ijt} + \tilde{\delta}_2 \log(\text{income}^*_{ijt-1}) + \varepsilon_{ijt}$$

問題

以下の(a) - (i)の問いに答えてください。

1. コードの実行

a. write_regression_modelsのfunctionの中を、モデル1~6の回帰分析に書き換えてください。(現時点では、すべてモデル1の内容が書かれています)

b. main()を実行し、回帰分析表をアウトプットしてください。

2. 回帰分析結果の解釈

回帰分析の結果をもとに、重回帰分析を解釈します。(なお、でエラーが発生してしまった場合などは、宿題に付随しているファイルにすでにあるアウトプットをもとに答えてもよい。)

A. 真のモデルと仮説検定

まず、データの不完全性などはなく、真のモデルで回帰分析をできるとき(モデル1)のアウトプットを 考えます。

c. 推定値$\hat{\beta}_1$ と$\hat{\beta}_0$の95%信頼区間は何ですか

d. 帰無仮説:$\beta_1 = 0$ と帰無仮説 :$\beta_2 = 0$ を別々に考えたとき、それぞれを棄却できますか。

e. (Adv.) 帰無仮説H_0を$\beta_1 = 1$に設定して考えたとき、これを棄却できますか。有意水準の両側検定を行ったとき、この帰無仮説を棄却してしまう確率は何ですか。

B. 不完全なモデルと推定値の解釈

データが不完全なときは、推定値にバイアスが生じる可能性があります。

f. B1のモデル2,3のいずれか、そしてB2のモデル4,5,6のいずれかを選び、回帰表から読み取れる情報を用いて、解釈・説明してください。その際に、真のパラメータに比べてどのようなバイアスがあるかを講義の内容を踏まえて説明してください。

- B1. モデル2 ... 「努力」が欠落したモデルの$\hat{\gamma}_2$
- B1. モデル3 ... 「親の所得」が欠落したモデルの$\hat{\delta}_1$
- B2. モデル4 ... 「子の所得」に測定誤差があるモデルの$\hat{\beta}_1$
- B2. モデル5 ... 「努力」に測定誤差があるモデルの$\hat{\gamma}_1$
- B2. モデル6 ... 「親の所得」に測定誤差があるモデルの$\hat{\delta}_1$

g. 上記のモデルの内、モデル2は必ずしも問題ではない場合があります。それはどのような場合でしょうか。（分析目的を考えてみてください。）

h. (Adv.) 自分が選んだモデル以外についても解釈・説明してください。

C. 標準誤差の正しい計算

この推定には、estimatrパッケージのlm_robust()という、分散不均一性やクラスター構造に対して頑健な標準誤差を計算できる回帰分析のコマンドを使っています。これに対し、真のモデル1について、デフォルトで使うlm()による回帰分析のアウトプットを表示しています。これらの推定結果がどのように異なるのか、なぜ異なるのか説明してください。

i. (Adv.) 真のモデル1では、親の所得を制御した上で、子本人の努力が所得に与える影響を推定しています。これは、すなわち、データセットの中の誰と誰を比べて、この影響を推定しているということでしょうか。

参考資料

R

私たちのR
R for Data Science (2e)

cheat sheet

Posit Cheatsheets

レポート関連

Quarto
Overleaf