回帰分析 I

2. イントロダクションの続き

1

イントロダクションの続き

- 前回のおさらい。
- 1998年カリフォルニア州の420の学区で集められたデータを使って、クラスサイズと生徒の学力の関係について実証分析した。
- 回帰分析の結果、クラスサイズと生徒の学力の間には統計的に有意な関係があることが分かった。
- 具体的には、先生の受け持つ生徒が一人増える(減る)と、生徒のテストの点が平均すると2点程下がる(上がる)ことが推定結果より明らかになった。
- この結果は、クラスサイズが現状の半分になると学力は3.4%上昇、一方、現状の倍にすると学力は6.8%低下することを示唆するものである。

2

イントロダクションの続き

- 確かにそれっぽいですよね。
- ここで、分析の結果をより正確に理解しておきましょう。
- この分析において、観測値は学区ごとのものです。
  - その学区の生徒(5年生)の共通テストでの平均得点
  - その学区にある学校の5年生のクラスサイズの平均
- なので、分析結果は、

「小規模クラスの学区の生徒の方が、クラス規模が大きい学区の生徒よりも共通テストの成績が良い」

と言った方がより正確かもしれません。

3

イントロダクションの続き

- 「小規模クラスの学区の生徒の方が、クラス規模が大きい学区の生徒よりも共通テストの成績が良い」
- この結果は予想通りかもしれません。
- が、次のような可能性を考えてみましょう。
- 小規模クラスの学区の生徒の方が、クラス規模が大きい学区の生徒よりも他の条件に恵まれていたとしたら、、、
- 例えば、小規模クラスの学区にはより豊かな住民が多いとしましょう。その場合、
  - 子供により多くの本を買い与えているかもしれない。
  - 子供をより多く塾に行かせているかもしれない。
  - 子供をより質の良い塾に行かせているかもしれない。

4

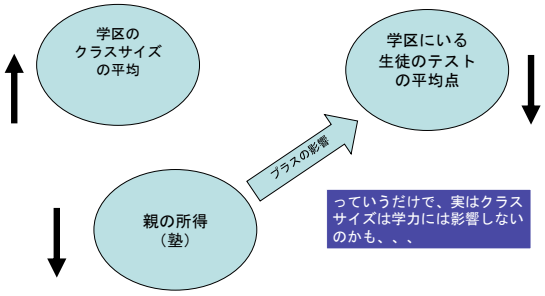
イントロダクションの続き

- この可能性が示すことは、分析結果は、、、

を示すと思っていたけど、実は、、、

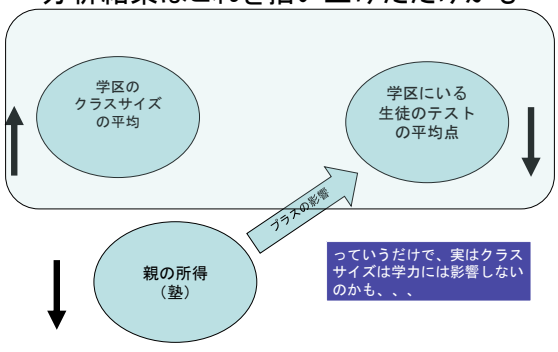
5

イントロダクションの続き



6

分析結果はこれを拾い上げただけかも



7

- 自ら計量分析をする際、またほかの人の分析結果を見る際に、このような可能性を考えることは非常に重要になります。
- このような可能性をすぐに思いつけるようになること、これは本講義の目的の一つです。
- 練習しましょう（今後何度も練習していきます）。
- 次の例を考えましょう。米国の住宅金融機関の話です。
- 「住宅ローン」の承認を判断する際、人種を考慮してはならない」
- こういう法律があります。

8

- すなわち、ローンの応募者が二人いて、
  - 一人は白人、一人は黒人
  - ただし他の条件（例えば所得、職種、勤続年数など）はすべて同じ

なら、この二人には等しくローンが認められなければならないということ。

- 人種差別はだめ、ってこと。
- ここで次の調査報告があります。1990年代初期、、、
  - 黒人の応募者の28%が住宅ローンを承認されなかった。
  - 一方、白人の応募者で承認されなかったのは9%程度。
- この調査報告は、人種差別があったことを示している？

9

イントロダクションの続き

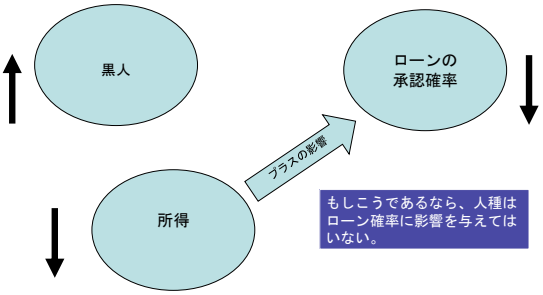
- 分析結果は、、、

を必ずしも示すものではない。例えば、、、

10

イントロダクションの続き

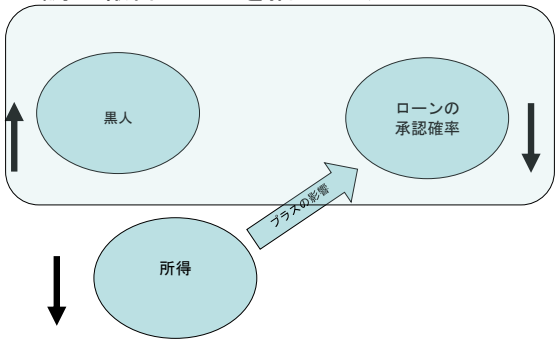
- 分析結果は、、、



もしこうであるなら、人種はローン確率に影響を与えてはいない。

11

調査報告はこれを拾い上げただけかも



12

- 今あげた二つの例。
- 最初の例では、「クラスサイズが学力に与える影響」を知りたかった。
- しかし、クラスサイズの影響とそれ以外の要因の影響（例：親の所得の影響）を区別できていないかもしれない。
- 次の例では、「ローンの承認において人種差別があるかどうか」が知りたかった。
- しかし、報告の結果は、人種差別が存在するという強いエビデンスにはなりえない。
- なぜなら、人種の影響とそれ以外の要因の影響（例：所得の影響）を区別できていない可能性があるから。

13

- 分析で明らかにしたいことは、、、

他の要因は一定の下で、ある変数が変化したときに別の変数に与える影響

- 最初の例では、

クラスサイズが変わると、他の要因（例えば親の所得）を一定としたとき、学力はどの程度変化するのか（またはしないのか）？

- 次の例では、

住宅ローンの承認確率は、ローン返済能力などの他の要因を一定としたとき、人種からの影響をどの程度受けるのか（またはうけないのか）？

14

他の要因は一定の下で、ある変数が変化したときに別の変数に与える影響

を定量的に計測する方法が、

「重回帰分析」

です。

- ちなみに前回やった分析は「**単回帰分析**」と呼ばれます。
  - 式は「**単回帰モデル**」といいます。
- 重回帰分析は、説明変数が複数ある場合です。
- 従属変数をy、（k個の）独立変数を$x_1, x_2, \ldots, x_k$とすると、「**重回帰モデル**」は、

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

と表されます。

15

重回帰分析（重回帰モデル）

- 「**重回帰モデル**」

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

において、独立変数の傾きを表す係数（以下「独立変数の係数」）は

他の要因は一定として、その変数の一単位の変化が従属変数に与える（平均的な）影響

を表します。

- 例えば、β_1は

「$x_2, \ldots, x_k$を一定として、x_1が一単位の変化すると、yが平均するとどれだけ変化するか」

を表します（偏微分の考え方です）。

16

重回帰分析やってみる

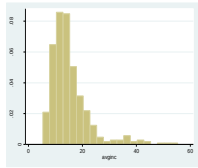
- それでは難しいことは後回しにして、重回帰分析してみましょう。
- 今回使うデータは caschool.dta です。
  - Stock and Watson (2007) *Introduction to Econometrics* (2nd edition), Pearson Education, Inc からのものです。
- 前回使ったデータは今回のデータの一部です。
  - このデータには数多くの変数が含まれています。
- 今回はこのデータの中の三つの変数を使います。
  - testscr (テストスコア：前回のscoreと同一のものです)
  - str （先生一人当たりの生徒の数：前回のstratioと同一のものです）
  - avginc（その地区の平均所得、単位 $1,000）

重回帰分析やってみる

- 前回の手順、覚えていますか？まず記述統計です。
- sum testscr str avginc （sumのみの場合は、変数全部について計算します）

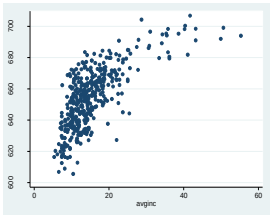
Variable	Obs	Mean	Std. Dev.	Min	Max
testscr	420	654.1565	19.05335	605.55	706.75
str	420	19.64043	1.891812	14	25.8
avginc	420	15.31659	7.22589	5.335	55.328

- 平均所得（正確には平均「平均所得」ですね）は$15,300、最大値は$55,000です。
- hist avginc
- 所得分布はこんな感じ。



重回帰分析やってみる

- テストスコアと所得の散布図を見てみましょう。
- scatter testscr avginc
- 平均所得が高い地域ほど、テストのスコアが高い傾向がある。
- 次に相関行列を計算してみましょう。
- corr testscr str avginc



	testscr	str	avginc
testscr	1.0000		
str	-0.2264	1.0000	
avginc	0.7124	-0.2322	1.0000

重回帰分析やってみる

- それでは、まず前回の単書きモデルの推定。
- reg testscr str

Source	SS	df	MS	Number of obs = 420
Model	7794.11004	1	7794.11004	F(1, 418) = 22.58
Residual	144315.484	418	345.252353	Prob > F = 0.0000
Total	152109.594	419	363.030056	R-squared = 0.0512
				Adj R-squared = 0.0490
				Root MSE = 18.581

testscr	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
str	-2.279808	.4798256	-4.75	0.000	-3.22298	-1.336637
_cons	698.933	9.467491	73.82	0.000	680.3231	717.5428

- 今度は重回帰モデル
- reg testscr str avginc

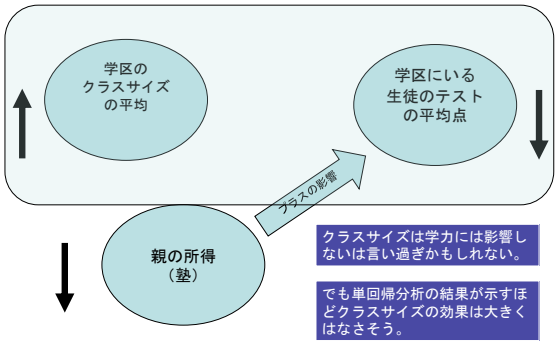
生徒教師比率の影響は、単回帰分析が示すほど大きくはない。

単回帰分析では、生徒教師比率と所得の影響を区別できてなかったのかも

Source	SS	df	MS	Number of obs = 420
Model	77801.487	2	38900.7435	F(2, 417) = 217.29
Residual	74308.1067	417	178.196898	Prob > F = 0.0000
Total	152109.594	419	363.030056	R-squared = 0.5115
				Adj R-squared = 0.5091
				Root MSE = 13.349

testscr	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
str	-.6487401	.354405	-1.83	0.068	-1.345383	.0479028
avginc	1.839117	.0927868	19.82	0.000	1.656724	2.0215
_cons	638.7292	7.449077	85.75	0.000	624.0867	653.3716

単回帰分析はこれを拾い上げただけかも、は言い過ぎだった。
けれど、、、

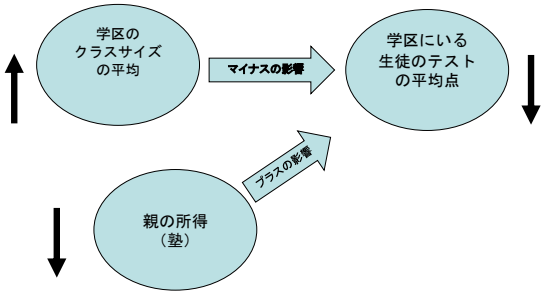


クラスサイズは学力には影響しないは言い過ぎかもしれない。

でも単回帰分析の結果が示すほどクラスサイズの効果は大きくはなさそう。

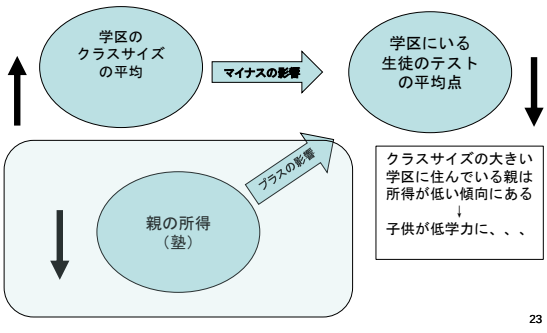
21

なぜ単回帰分析はクラスサイズの効果を過剰に評価した？



22

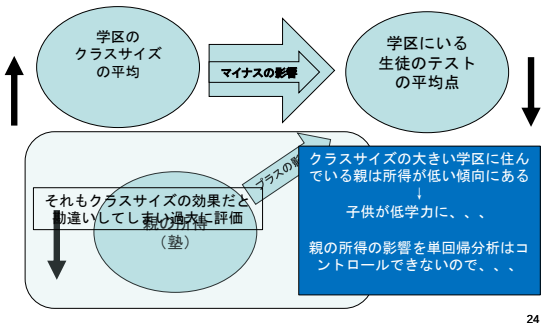
なぜ単回帰分析はクラスサイズの効果を過剰に評価した？



クラスサイズの大きい学区に住んでいる親は所得が低い傾向にある
↓
子供が低学力に、、、

23

なぜ単回帰分析はクラスサイズの効果を過剰に評価した？



それもクラスサイズの効果だと勘違いしてしまい過大に評価

クラスサイズの大きい学区に住んでいる親は所得が低い傾向にある
↓
子供が低学力に、、、

親の所得の影響を単回帰分析はコントロールできないので、、、

24

除外変数バイアス

- これは「**除外変数バイアス**」と呼ばれるものの一例。
- これが生じるメカニズムについて、本講義では数式を使って詳しく解説する予定になっています。
- 慣れると、**バイアス**の方向（過大**評価**または**過小評価**）が瞬時に推理できるようになります。
  - これができると、プロの間でも「こいつやるな」と思われることがあります。
  - ですが、実はそう難しくありません。
  - いったんメカニズムが分かれば、パターンにはめるだけなので。
  - また推理の際に数式は要りません。
- そうなるように、本講義では繰り返し練習します。

25

相関関係と因果関係

- まず言葉を定義します。
- **因果性**（causality）とは、「ある特定の行動や処置が、別の測定可能な結果をもたらすこと」（Stock and Watson, 2007）。
- **因果効果**（causal effect）とは、「ある特定の行動や処置が、結果に及ぼす効果」（Stock and Watson, 2007）。
- **相関関係**と**因果関係**の違いを理解することは極めて重要です。
- 相関関係 $X \leftrightarrow Y$: 例えば、「Xが大きいほど、Yは小さい」
  - Xを先に言っていますが、別に原因というわけではありません。
  - この場合、「Yが小さいほど、Xは大きい」でも、意味は全く同じです。
- 因果関係 $X \rightarrow Y$: Xの変化・差異で、Yの変化・差異が生じた。

26

相関関係と因果関係

「相関関係があること」は「因果関係があること」を必ずしも意味しない

- いくつか（極端な）例を考えてみましょう。
- サメ被害が多いほど、アイスクリームが良く売れる傾向にある。
- 殺人事件が多いほど、アイスクリームが良く売れる傾向にある。
- これらは統計的に観測されていますが、、、

アイスクリーム会社は売上増大するために、海水浴場にサメをいっぱい放すとよい、、、 とか

アイスクリーム会社は売上増大するために、殺人鬼を雇用して、、、

とはならないですよね。

27

相関関係と因果関係

- 社会科学の仮説の多くは、変数間の「因果関係」に関するもの。
  - 例1：価格が上昇すると、需要が下がる（需要の法則）。
  - 例2：所得水準の向上はその国の環境負荷を大きくする。ただしある水準を超えると環境負荷は小さくなっていく（環境クズネッツ仮説）。
- また「因果効果」を推定したい場合も多い。
  - 例：労働者が職業訓練を受けると、所得はどれだけ変化するか。
- 従って、実証分析の目標は、データを使って変数間の因果関係の有無を明らかにすること、因果効果を推定することとなります。

28

回帰分析と因果関係

- 実証分析の目標は、因果関係の有無を明らかにすること、因果効果を推定すること。
- 回帰分析は、他の要因は一定として、変数の一単位の変化が従属変数に与える(平均的)な影響を明らかにする。
- ってことは、回帰分析は実証分析に役立ちそう。そして、確かにそう。
- ただし、回帰分析をする際(または人の回帰分析の結果を評価する際)に知っておかなければならない重要なこと、あり。
- それは

「回帰分析は因果関係の有無を必ずしも明らかにしない」

- なんか矛盾している感がありますが、こういうことです、、、

29

- ある回帰モデル(従属変数Y、独立変数X)を推定した。
- 結果は、「変数Xの係数が統計的に有意にゼロと異なっている。」
- この結果がXとYの因果関係を示すかどうかは、、、

「回帰モデルの仮定がデータにおいて成り立っているかどうか」

に依存する。

  - 仮定が成り立っているなら、因果関係を示す。
  - 成り立っていないなら、因果関係を示さない。
- 従って、回帰モデルの仮定を理解することが重要になります。
- この仮定については本講義で説明されます。
  - 数式を使っての議論が必要に。
  - ただし、仮定を数式のまま表面的に理解しても、実際の分析では役に立たない。
  - 意味を十分に理解し、仮定がデータにおいて成り立っているかどうか判断できるようになることが大事(本講義ではそうなれるようにトレーニングします)。

30

まとめ

- 単回帰モデル: 独立変数が一つ。
- 重回帰モデル:独立変数が複数ある。

他の要因は一定として、その変数の一単位の変化が従属変数に与える(平均的な)影響

を見ることができる。

- 単回帰モデルでは「他の要因を一定にして、、、」が困難。そのため、独立変数が従属変数に与える(平均的な)影響を過大または過小に評価してしまうことあり。
  - 除外変数バイアス
- 相関関係と因果関係の違い: 相関関係は因果関係を必ずしも意味しない。
- 因果関係と回帰分析
  - 因果関係を示すかどうかは、回帰分析の「仮定」が成り立っているかどうかによる。
  - この仮定が何なのか、なぜそうなのかについては、後日詳しく見ていきます。

31

宿題

- 宿題1: 以下のリンクの記事を読んでおいてください。

https://www.newsweekjapan.jp/stories/us/2013/07/post-2989.php

- 宿題2: 相関関係はあるけど因果関係はない(と考えられる)例を、その理由も添えて考えておいて下さい。
  - 授業で聞きますので、用意しておいた方がいいかと。

32