

回帰分析 I

1. イントロダクション

1

イントロダクション

「小学校一クラスの人数を減らす」

この提案どう思います？

賛成？なぜ？

反対？なぜ？

2

- 「小学校一クラスの人数を減らす」
- 生徒少ない → 先生が一人一人により多くの注意を払える
- 学力向上。あるかも。
- いじめの減少。あるかも。
- 良いかもしれないですね。

3

- でも、それだけかな？
- 小学校一クラスの人数を減らすにはコストがかかる。
- より多くの教師を雇う必要あり。
- 教室がすでに一杯なら、新しい校舎も必要に。
- 提案の導入の検討のためにはコスト・ベネフィットの比較が必要。
- コストは比較的簡単に計算できそう。
- ではベネフィットは？

4

- 先ほど挙げたベネフィットは二つ：「学力向上」と「いじめの減少」
 - もちろん他にもベネフィットはあるかもしれません。
 - 考えてみて下さい。
 - そんな感じで思いついたものが研究のトピックになります。
- ここでは「学力向上」について考えてみましょう。
- 生徒少ない → 先生が一人一人により多くの注意を払える → 学力向上
- 常識的には確かにありそう。
 - 常識や身近な経験から「生徒少ない → 学力向上」はなんとなく想像できる。
- でも、この効果がどの程度の大きさなのかは、常識や身近な経験では分からない。

5

- 効果がどの程度の大きさなのか知る必要はある？
- 次の例を考えてみて下さい。

「毎年3000億円ほど余計にかかりますが、平均点が1点伸びます。」

 - これに対する反応はもちろん人それぞれですよ。
 - その反応に正解とか不正解はないです、、、
- それでは

「毎年3000億円ほど余計にかかりますが、平均点が50点伸びます。」

だとどうですか？
- 効果を定量的に把握するためには、クラス人数と学力の関係についての実証的な分析（データを使った分析）が必要。
- やってみましょう！

6

初めての实証分析

- リサーチエクステションの設定

クラス人数と学力の間に関係はあるのか？
あるとすればどの程度のものなのか？

- 関連した研究を調べる
 - 関連した研究をした論文が一本でも見つければ、あとはその論文に載っている参考文献を見て、、、またその参考文献を見て、という感じで。
 - とっかかりの一本の見つけ方ですが、私の場合は、キーワードいれてググります。
- データを手にする
 - ここでは以下のデータを手に入れたとしましょう。

7

初めての实証分析

- 1998年カリフォルニア州の420の学区で集められたデータ。
 - 参考文献に挙げてある
Stock and Watson(2016) *Introduction to Econometrics*, Pearson Education, Inc.
からのデータです。日本語版は共立出版から出ています。
- 1学級の大きさと基礎学力の二変数からなるデータ。
- 小規模学級の学区の生徒の方が、大規模学級の学区の生徒よりも、共通テストの成績が良い傾向あり。
- 見てみましょう(lecture1_ex1.dta)。
 - STATAフォーマットのデータ。
 - STATAの使い方はこの講義でおいおい説明していきます。

8

まずデータをよく見る

- 420の観測値(420学区) = サンプルサイズは420
- ID: 学区のID
- *score*: 共通テストにおけるその学区の生徒(5年生)の平均点
- *stratio*: Student-Teacher Ratio = その学区における先生一人に対しての生徒数の平均

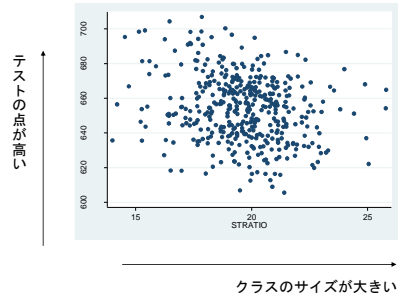
- sum → 記述統計量を与えてくれるコマンド

標本平均(mean)、標本標準偏差(std.dev)、最小値(min)、最大値(max)

- 生徒の平均点はだいたい654点
- 平均すると、先生一人に対して生徒はだいたい19.6人

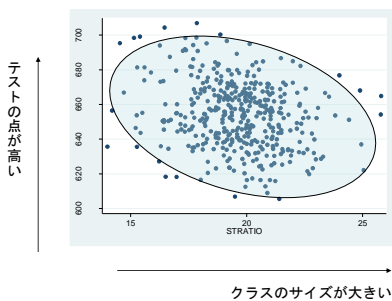
9

- histogram *score* → *score*のヒストグラム
- histogram *stratio* → *stratio*のヒストグラム
- scatter *score stratio* → *score*と*stratio*の散布図



10

ざっくりいうと右下がりの関係にあるかな？



11

- corr *score stratio* → *score*と*stratio*の標本相関係数
 - -0.23 (*score*と*stratio*は負に相関している)
 - すなわち、先生一人当たりの生徒数が多いところでは、生徒のテストの点が低い傾向にある。
- =====
- 重要: データを手にしたら、まずこのように「よく見てみる」こと。
- 記述統計からデータの大まかな傾向をつかむ。
 - 最大値、最小値があり得ない値をとっている場合、入力ミスの可能性あり。
 - そんなことも記述統計は教えてくれる。
 - ヒストグラムを使って分布の形状を知る(特に従属変数)。
 - 散布図を使って二変数がどのように散らばっているか見る。
 - 他の観測値からボツンと遠くに離れている「外れ値」があるかどうか知ることができる。

12

回帰分析

- それでは、先生一人当たりの生徒が一人増えると、生徒のテストの点は（平均すると）どれだけ下がる？
- この答えを与える道具が「**回帰分析**」。
- 「回帰分析」とは統計的分析において最もよく使われる手法の一つ。
- ほぼすべての社会科学（経済学、ファイナンス、経営学、マーケティング、会計学、社会学などなど）で使われている。
- 政府機関でも。そしてビジネスでも。

「これと字面の「回帰分析」の由来の関連性数は...

13

回帰分析

- 「回帰分析」とは、データを用いて変数間（ここでは $score$ と $stratio$ が変数）の関係性を計測する手法の一つ。
- 少し難しく言うと...

回帰分析は、従属変数と呼ばれる一つの変数を、（一つまたは複数の）独立変数と呼ばれる変数の関数として定式化し、それをデータを使って推定するもの。

- **従属変数（被説明変数ともいう）**とは、その動きを説明したいもの。
➢ この例では、 $score$ が従属変数。
- **独立変数（説明変数ともいう）**とは、従属変数の動きを説明するもの。
➢ 分析前に、分析者が、説明すると考えるもの、と言った方がより正確かもしれない。
➢ 分析してみて、「関係なし」という結果を得ることはしばしばある。
➢ この例では、 $stratio$ が独立変数。

14

回帰分析：実証モデルの設定

- 難しいことは追って学習するとして、とりあえず回帰分析してみましょう。

回帰分析は、従属変数と呼ばれる一つの変数を、（一つまたは複数の）独立変数と呼ばれる変数の関数として定式化し、それをデータを使って推定するもの。

- ここでは $score$ と $stratio$ の関係を以下のように定式化しよう。

$$score = \alpha + \beta \cdot stratio$$

- α は「**定数項**」と言う。
- β は「**傾きを表す係数**」。
➢ $stratio$ が増えたときに、 $score$ がどれだけ増えるか（ β が正の時）、またはどれだけ減るか（ β が負の時）、あるいは変化しないか（ β がゼロの時）を表す。
➢ β の値が特にここで知りたいこと。
- α と β はこの式の「**パラメータ**」とも呼ばれます。

15

$$score = \alpha + \beta \cdot stratio$$


- この式どう思います？
- 試験の点って、先生一人当たりの生徒数だけでは決まらないんじゃない、
- そう。なので、回帰分析ではこうします。

$$score = \alpha + \beta \cdot stratio + u$$

- u は「**誤差項**」と呼ばれるもの。
- $score$ に影響を与える $stratio$ 以外のさまざまな要因の影響を捉える。
- 分析者は u を観測することはできない。

16

- イメージするなら、、、

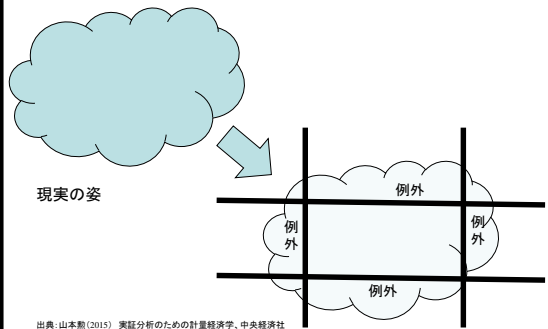


現実の姿

出典: 山本龍(2015) 実証分析のための計量経済学, 中央経済社

17

- イメージするなら、、、



現実の姿

例外

例外

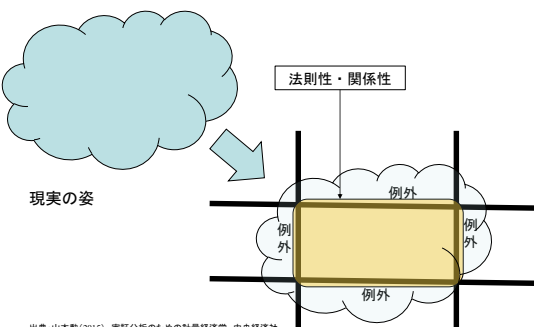
例外

例外

出典: 山本龍(2015) 実証分析のための計量経済学, 中央経済社

18

- やろうとしていることを、イメージするなら、、、



現実の姿

法則性・関係性

例外

例外

例外

例外

出典: 山本龍(2015) 実証分析のための計量経済学, 中央経済社

19

- 回帰分析は、現実から本質的な法則性・関係性・因果性を浮き彫りにする道具。
- 本質的な部分とノイズ(例外)を分ける。

$$score = \alpha + \beta \cdot stratio + u$$

- 例外という言葉は、あまり正しくないかもしれないけど、こんな感じでイメージするのは悪くない。

出典: 山本龍(2015) 実証分析のための計量経済学, 中央経済社

20

回帰分析: パラメータの推定

$$\text{score} = \alpha + \beta \cdot \text{stratio} + u$$

- それではパラメータを推定しましょう。
- かなりざっくり言うと、
- パラメータの推定とは、二変数からなる回帰分析の場合、散布図のプロットの中心付近に一本の直線を引くことを意味します。
- 引かれた直線を「**回帰直線**」と言います。
- 引き方は、「各プロットにできるだけ近くなるように」。
- 「**最小二乗法**」と呼ばれる方法です。

21

- reg score stratio

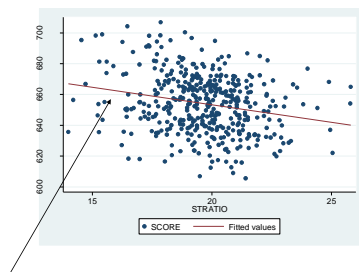
Source	SS	df	MS	Number of obs =
Model	7789.39296	1	7789.39296	420
Residual	144312.057	418	345.244156	PC = 1, (418) = 22.56
Total	152101.45	419	363.010621	Prob > F = 0.0000
				R-squared = 0.0512
				Adj R-squared = 0.0489
				Root MSE = 18.581

score	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
stratio	-2.279063	.4798082	-4.75	0.000	-3.2222 -1.335925
_cons	698.9222	9.467187	73.83	0.000	680.3129 717.5314

- β (傾きを表す係数)の**推定値**は-2.27。
- この推定結果は「stratio が1増えたときに、平均するとscore が2.27点下がる」ことを意味しています。
- すなわち、
「先生の受け持つ生徒が一人増えると、生徒のテストの点がだいたい2点下がる」
ということです。

22

- twoway (scatter score stratio) (lfit score stratio)



この線が推定された「回帰直線」です。

23

- reg score stratio

Source	SS	df	MS	Number of obs =
Model	7789.39296	1	7789.39296	420
Residual	144312.057	418	345.244156	PC = 1, (418) = 22.56
Total	152101.45	419	363.010621	Prob > F = 0.0000
				R-squared = 0.0512
				Adj R-squared = 0.0489
				Root MSE = 18.581

score	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
stratio	-2.279063	.4798082	-4.75	0.000	-3.2222 -1.335925
_cons	698.9222	9.467187	73.83	0.000	680.3129 717.5314

- α (定数項)の**推定値**は698。
- そのまま解釈するなら、先生が一人も生徒を受け持っていないとき、生徒のテストの平均点は698点ということになります。
 - これは何を言っているのかちょっと分からない感じがですね、
 - 式の定式化に少し工夫が必要かも。

24

- reg score stratio

Source	SS	df	MS	Number of obs = 420
Model	7789.39296	1	7789.39296	F(1, 418) = 22.56
Residual	144312.057	418	345.244156	Prob > F = 0.0000
Total	152101.45	419	363.010621	R-squared = 0.0512
				Adj R-squared = 0.0489
				Root MSE = 18.581

score	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
stratio	-2.279063	.4798082	-4.75	0.000	-3.2222 -1.335925
_cons	698.9222	9.467187	73.83	0.000	680.3129 717.5314

- これらは「標準誤差」と呼ばれるもの。
- ざっくりいうと、パラメータの推定値がどれだけ正確かを教えてくれるもの。
- 小さければ小さいほど正確。

25

- reg score stratio

Source	SS	df	MS	Number of obs = 420
Model	7789.39296	1	7789.39296	F(1, 418) = 22.56
Residual	144312.057	418	345.244156	Prob > F = 0.0000
Total	152101.45	419	363.010621	R-squared = 0.0512
				Adj R-squared = 0.0489
				Root MSE = 18.581

score	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
stratio	-2.279063	.4798082	-4.75	0.000	-3.2222 -1.335925
_cons	698.9222	9.467187	73.83	0.000	680.3129 717.5314

- これは「決定係数」とよばれるもの。
- モデルのデータに対する当てはまり(フィット)の良さを示す指標。
- この分析では、決定係数が0.05。
- これは、scoreの変動(ばらつき)の5%がモデルによって説明されることを意味します。

26

回帰分析: 仮説検定

- 「先生の受け持つ生徒が一人増えると、生徒のテストの点がだいたい2点下がる」という結果が得られたけれど、この結果信頼していい?
- データを使った推定だから推定誤差はあるかもしれない。
 - 2点下がるっていう結果は、推定誤差によるものかも?
 - 本当は生徒が増えても点数に変化はない($\beta = 0$) のでは?
- そこで、回帰分析では、 $\beta = 0$ かどうかパラメータの推定値を使って「仮説検定」します。
- この仮説検定では「t値」と呼ばれるものを使います。

27

回帰分析: 仮説検定

- t値はパラメータの値を標準誤差で割ったものです。

Source	SS	df	MS	Number of obs = 420
Model	7789.39296	1	7789.39296	F(1, 418) = 22.56
Residual	144312.057	418	345.244156	Prob > F = 0.0000
Total	152101.45	419	363.010621	R-squared = 0.0512
				Adj R-squared = 0.0489
				Root MSE = 18.581

score	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
stratio	-2.279063	.4798082	-4.75	0.000	-3.2222 -1.335925
_cons	698.9222	9.467187	73.83	0.000	680.3129 717.5314

- ざっくりいって、この値が絶対値で2より大きければ、、、

$$\beta = 0 \text{ である可能性は低い}$$

と判断します(この場合「 β は統計的に有意にゼロと異なる」と言います)。

- 今の例では、 β は統計的に有意にゼロと異なります。

28

シミュレーション

- 推定結果を使って、被説明変数の予測値を出すこともできます。
- 予測値は、誤差項を除いた式の右辺に、パラメータとデータを代入して算出します。

$$score = 698.9 - 2.23 \cdot stratio$$

- $score$ の上のハットは予測値であることを強調するためのものです。
- この $stratio$ に具体的な数を入れれば、そのときの $score$ の予測値が得られます。
- それではシミュレーションしてみましょう。

29

シミュレーション

$stratio$	$score$
5	687.8
10	676.6
15	665.5
20	654.3
25	643.2
30	632.0
35	620.9
40	609.7
45	598.6
50	587.4

- $score$ だとちょっとピンときませんね。
- パーセント変化で出した方が良いかも。
- 現状では $stratio$ は平均19.6人ですから、ざっくり20人として、その時にくらべて $score$ が何パーセント変化するかも出してみましょう。

30

シミュレーション

$stratio$	$score$	%変化
5	687.8	5.1%
10	676.6	3.4%
15	665.5	1.7%
20	654.3	Baseline
25	643.2	-1.7%
30	632.0	-3.4%
35	620.9	-5.1%
40	609.7	-6.8%
45	598.6	-8.5%
50	587.4	-10.2%

- クラスサイズが今の半分(10人)になると、(共通テストの点で測った)学力は3.4%上昇。
- クラスサイズが今の倍(40人)になると、(共通テストの点で測った)学力は6.8%低下。

31

結論

- 本研究は、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。

32

おまけ

- ここでは行いませんでしたが、、、
- 先生の受け持つ生徒を一人減らすことのコストがどの程度なのか分かれれば、
 - 先生の受け持つ生徒をA人減らすのにB円かかる。
 - ただしそれにより学力がC%上がる。
 - 何人ぐらい減らすのがいい、、、(またはコストが高すぎるから減らさなくていい、、、)

などと言った政策提言もできるかもしれません。

33

まとめ:実証分析の流れ

- 今やったように実証分析を行います。
- ここでは実証分析の典型的な流れをまとめておきます。
- **ステップ1:リサーチ・クエスションの設定**
- いくつかの**パターン**があります。
 - 自分で仮説を構築し設定する。そしてそれを実証する。
 - すでにあるよく知られている仮説を実証する。
 - すでにある対立する仮説について、どちらが正しいか実証する。
 - 重要な「効果」(例えば政策の効果)をできる限り正確に推定する。
- などなど。
- ここら辺のことについては、ゼミに入っている場合には、指導教授に聞いてみるのがいいでしょうね。

34

まとめ:実証分析の流れ

- **ステップ2: 実証モデルを考える**
- **ステップ3: データを探す、手に入れる、集める、、、**
 - データがなくて実証分析できないこともあります。
 - その場合は、ステップ1に戻らなくてはいけないかもですね。
 - また先にデータがあつて、そこからリサーチクエスションを考える場合もあります。
- **ステップ4: データをよく見る(これ大事)**
 - 記述統計、ヒストグラム、散布図、相関、、、
 - とにかく自分が扱っているデータの特性を理解しましょう。
 - このステップを経ないですぐに回帰分析する人が結構います。
 - そういう人の分析はたいてい、、、
- **ステップ5: 回帰分析(推定、仮説検定)**
- **ステップ6: シミュレーション**
 - これはやっぱりやらなかったりです。研究の目的によります。

35

終わりに

- 今日は細かいことは気にせずに実証分析をしました。
- 今学期の終わりには、こんな感じのことが、もろもろのことを理解した上でできるようになると思います。
- ゼミに所属されている方は、高確率で実証分析をやることになると思いますので、ここでしっかり学習しておくことをお勧めします。
 - 回帰分析を独学するのは結構困難です。かりにできたとしても効率が良くないです。
- 今後の講義では、細かいんだけど知らなくてはいけないことについても説明していきます。
- 「細かいんだけど知らなくてはいけないこと」の理解のためには、確率・統計の知識がある程度必要になります。

36

終わりに

- 必要となる確率・統計については、講義で説明します。
- が、丁寧に説明する時間はありません。
- 従って、学部「基礎統計学」を履修済みの場合はぜひ復習しておいてください。
- 履修済みでない場合は、学部「基礎統計学」を並履修してください。

37

宿題

- <https://hbr.org/2015/11/a-refresher-on-regression-analysis>
- この記事を読んでおいてください。

38