回帰分析I

16. 回帰分析の評価

イントロダクション

- これまで線形回帰モデルについていろいろ説明してきました。
  - 推定の仕方（モーメント法、最小二乗法）
  - 統計的検定（*t*検定・*F*検定）
  - 関数形（level-level, log-level, log-log, 二乗項など）
  - ダミー変数
  - 相互作用項
  - 不均一分散を考慮に入れた標準誤差の計算
- この講義で扱う（ほぼ）最後のトピックは、線形回帰モデルを使った分析において、
  - どういうときに結果は信頼でき、どうときに信頼に値しない？
  - 結果は一般化できる？
  - それらをどのように評価する？

です。

内部の正当性・外部の正当性

- 回帰分析の評価に用いるフレームワーク:

「内部の正当性」と「外部の正当性」

- 「統計分析が内部に正当である（internally valid）」とは、、、、

因果効果に関する統計的な推論が、分析対象の母集団や設定にとって正当である状況のこと

- 「統計分析が外部に正当である（externally valid）」とは、、、、

因果効果に関する統計的な推論が、より一般に他の母集団や設定にも妥当なもので、結果の一般化が可能である状況のこと

「分析対象の母集団」と「関心のある母集団」

- 「分析対象の母集団」とは、そこから標本データが抽出される主体（例：個人、企業、市町村）の母集団のこと。
- 「分析対象の母集団（population studied）」は「関心のある母集団（population of interest）」とは必ずしも同じではありません。
- 例えば、日本の大学生の喫煙行動について知りたいとしましょう。
- このとき「関心のある母集団」は「日本の大学生」ですね。
- 日本の大学生を母集団として、そこから無作為抽出し、それに基づき分析すればいいのですが、、、
  - 日本の大学生を母集団にして無作為抽出できるほどの予算が無い、、、
  - 早大の学生なら比較的簡単にデータを集めることができる、、、
  - たまたま早大の学生の喫煙行動に関するデータがあった、、、

などの理由から、「早大の学生」が「分析対象の母集団」になるかもしれません。

設定

- 「設定」とは、制度的、法的、社会的、そして経済的な環境のことを意味します。
- 先ほどの例であれば、20歳未満の喫煙はダメ、これは全国共通ですが、大学によって「設定」はかなり違うかもしれません。
  - 早大はキャンパス内で喫煙可ですが、東北大学はキャンパス内全面禁煙です。
  - 早大の喫煙所の数は他大学と違うかもしれません。
  - 早大学周辺は路上喫煙禁止ですが、そうではない大学もあります。
  - 禁煙教育の実施の有無、、、
  - 禁煙教育の質、、、

  などなど。

5

内部の正当性

- 「統計分析が内部に正当である」とは

  因果効果に関する統計的な推論が、分析対象の母集団や設定にとって正当である状況のこと

- この内部の正当性は二つの部分からなります。

(1) 係数の推定量が不偏性・一致性をもつこと

  - この講義では「線形回帰モデルを最小二乗法で推定」を取り扱ったので、不偏性・一致性が内部の正当性に必要になります。
  - より一般的には、不偏性が必要ではないこともあります。

(2) 仮説検定が設定された通りの有意水準を持つこと、また信頼区間も設定された通りの信頼水準を持つこと

6

内部の正当性を危うくする要因

(1)「係数の推定量が不偏性・一致性をもつこと」を危うくする要因は以下の通りです。

A. 関数形の特定化の誤りによるバイアス

B. 欠落変数バイアス

C. 説明変数のの計測誤差によるバイアス

D.同時双方向の因果関係によるバイアス

E.標本セレクションバイアス

(2)「仮説検定が設定された通りの有意水準を持つこと、また信頼区間も設定された通りの信頼水準を持つこと」を危うくする要因は以下の通りです。

F. 不均一分散

G. 誤差項の観測値間の相関

7

最小二乗推定量にバイアスを生じさせる要因

- 「係数の推定量が不偏性・一致性をもつこと」を危うくする要因とは、別の言い方をすれば、

  「どういうときに係数の推定量にバイアスを生じてしまうか」

の「どういうとき」ですね。

- さまざまな要因により、最小二乗推定量はバイアスします。
- 本講義で強調してきた「欠落変数バイアス」は、数ある要因の中の一つに過ぎません。
- 以下、一つ一つ見ていきましょう。

8

A. 関数形の特定化の誤りによるバイアス

- 関数形の特定化を誤れば、最小二乗推定量にはバイアスが発生します。
- 例えば、真のモデルが

$$lnY = \beta_0 + \beta_1 X + U$$

のときに

$$Y = \beta_0 + \beta_1 X + U$$

を推定すれば最小二乗推定量にはバイアスが発生します。

  - もちろん、逆のケースでもバイアスは発生します。

- 真のモデルには二乗項（交互作用項）が入っているのに推定モデルには入れなかった、このときにも最小二乗推定量にはバイアスが発生します。

9

関数形の特定化の誤り：対処法

- まずデータをよく見ましょう。プロットです。
- まずlevel-levelでプロットですね。
  - 二乗項は必要？それとなくチェック。
- 連続的変数で対数を取れるならlog-levelでプロットしたり、log-logでプロットしたり。
  - 二乗項は必要？それとなくチェック。
- 先行研究をよく見ましょう。logを取っているか、二乗項をモデルに入れているかを確認しましょう。
  - 先行研究を真似ておくのが無難です、専門家が試行錯誤の末にたどり着いたものと考えられるので。

10

- （例外はありますが）よく使われている変数の変換は、、、
- 非負の連続的変数には対数を取ることが多い。
- 必ずしも正しくはないが、ゼロ以上（ゼロを含みます）の値を取る連続的変数には1を足して対数をとることがしばしばある。
  - 変数Xは$X \geq 0$である連続的な変数。
  - このとき$\ln(X+1)$と変換してモデルに入れることあり。
- 比率には対数を取らない。
  - 対数を取ると意味的にはパーセントのパーセントに、、、
  - パーセントが何パーセント増えると → 解釈上よく分からなくなる。

11

B. 欠落変数バイアス

- これについては説明の必要は無いですね。
- なので、ここでは対処法のみを説明します。
- このバイアスをどう最小化するかは、除外された変数のデータが利用可能かどうかに依存します。
- 最初に利用可能である場合、次に利用可能ではない場合について説明します。

12

欠落変数バイアス：除外された変数が観察されるとき

- 除外された変数が観察可能（データとして利用可能）なときは、その変数をモデルに入れることで、問題を解決することができます。
- これでこの問題は解決です。ですが、、、
- このアドバイスはあまり意味のあるものではありません。
- というのも、そもそも「この変数をモデルに入れないと欠落変数バイアスが生じる」と分かっているなら、その変数をモデルの中に最初から入れてますよね。
- 従って、考えるべきことは、変数をモデルに入れるか・入れないかを、どのように決めればいいかということになります。

13

- 一つのアプローチは、「除外変数が無いように努めます、そのためにとにかく手あたり次第変数をモデルの中にぶち込みます」です。
- 確かにそうすれば欠落変数バイアスが生じるリスクは小さくなりますね。
- しかし、「手あたり次第」だと、数多くの不必要な変数をモデルの中に入れてしまいそうですね。
- 不必要な変数をモデルの中に入れると、推定量の分散が大きくなります（講義ノート10の84ページ参照）。
- 推定量の精度が落ちるということですね。
- また推定量の標準誤差が大きくなりますから、本当は関係ある変数なのにもかかわらず、統計的検定で「有意では無い」ということになってしまうかもしれません。
- ということで、「手あたり次第」はやめた方が良いです。

14

- ここでは、研究のタイプを「BがAに与える影響」を分析するものとして、変数選択の方法についてアドバイスをしておきます。
  - ここでBは「分析の興味の対象である説明変数」です。
- 手あたり次第に変数をモデルに入れるのではなく、まずB以外でAに影響を与えそうな変数（コントロール変数）を考えます。
  - 分析では、その他の要因の影響を一定にしたときの（またはコントロールしたときの）「BがAに与える影響」に興味があるわけです。
  - その意味でコントロール変数と呼ばれます。
- コントロール変数の選択の際、Aについて分析している先行研究を調べることが重要になります。
  - 「BがAに与える影響」を分析している先行研究があるなら、その先行研究がモデルに入れている変数はできる限りモデルに入れるようにします。
  - 「BがAに与える影響」を分析している先行研究がないなら、「CがAに与える影響」や「DがAに与える影響」を分析している先行研究を見つけましょう。
  - そしてそれらがモデルに入れている変数をできる限りモデルに入れるようにします。

15

- 先行研究で使われているコントロール変数をモデルに入れた結果、そのコントロール変数が重要では無かった（統計的に有意では無かった）ということがしばしばあります。
- その場合でも、そのようなコントロール変数はモデルの中に入れたままにするのが一般的です。
  - そのようなコントロール変数は、先行研究の積み重ねによって選ばれたものであることが多いため、、、
  - 抜いてしまうと、読み手に「この変数が入っていない、除外変数バイアスが生じているのではないか？」と疑われてしまう可能性があります。
- 確かに、これにより推定量の分散は大きくなってしまうかもしれません。
- それでも、先行研究に基づきコントロール変数を選択した方が、なんでもかんでもとりあえずモデルに入れるよりは、「無駄な変数」をモデルに入れてしまう確率は低くなるでしょう。

16

- 欠落変数バイアスに関して、研究上で「おいしい」ケースを挙げておきます。
- 先行研究のメインの仮説が「BはAに与える影響を与える」というものであり、回帰分析の結果が仮説と整合的であったとします。
- 確かに自分の手持ちのデータでもそのような結果がでます。
- ただし、その先行研究では「CがAに与える影響」をコントロールしていませんでした。
- Cを分析に加えてみると、、、、、
  - Bの係数のサイズが大きく変わった、、、
  - Bの係数の符号が変わった、、、
  - 「Bの係数はゼロ」という帰無仮説を棄却できない

となるだけでなく、実は重要なのはBではなくCだった、ということが分かったとします。

17

- このようなときは、論文に、、、
- 先行研究と同じようなモデルの結果を報告。
- そして、Cを加えたモデルの結果を報告。
- そして先行研究で生じていたと思われる欠落変数バイアスを指摘します。
- これにより、「先行研究は効果を過大（または過小）に推定していた」や「先行研究の仮説は実はデータと整合的では無いかもしれない」などと結論づけます。
- 勝ちゲームです。

18

- 別の「おいしい」ケースは、先行研究が、
  - 「分析の興味の対象である説明変数」の二乗項を考慮し忘れた
  - 「分析の興味の対象である説明変数」と何らかの変数の相互作用項を考慮し忘れた

ようなときです。

- 理論的に考えると（妄想とも言いますが）、「分析の興味の対象である説明変数」に関してそのような可能性があるなら、そのようなモデルも推定してみましょう。
- あとは、先ほどと同じです。
- 先行研究と同様に二乗項（相互作用項）が入っていないモデルを推定。
- 入れたモデルを推定。本当に思った通りなら、、、
- 「BとAの関係は実は非線形で、、、」や「BとAの関係は実はCに依存していて、、、」という新しい知見を与えられます。

19

欠落変数バイアス：除外された変数が観察されないとき

- 除外された変数のデータが手元になければ、その変数を回帰分析に入れることはできません。
- 実際の分析では、何らかの重要な変数が除外されていることがほとんどでしょう。
- 個人レベルの分析では、個人が生まれもっている能力（遺伝子）のような変数はデータとして存在しないのが普通です。
- 企業レベルの分析では、経営者の性格みたいな変数もデータとして存在しないのが普通です。
- 県レベルの分析では、県民性みたいな変数もデータとして無いかもしれません。
- そのため、一定水準を超えた研究は、最小二乗法による結果を使って「これは因果関係です」と主張することをほとんどしません。

20

欠落変数バイアス：除外された変数が観察されないとき

- 欠落変数バイアス、、、詰んだ、、、
- 詰んでません、対処法は少なくとも二つあります！

(1) パネルデータがある ＋ 「固定効果推定法」

もしパネルデータが利用可能で、除外された変数が時間を通じて一定の場合、「固定効果推定法」と呼ばれる方法で欠落変数バイアスを回避できます。

(2) 操作変数法

「操作変数」と呼ばれるある条件を満たす変数があるなら、その変数と操作変数法と呼ばれる推定方法を使うことで欠落変数バイアスを回避できます。

これらの方法は来年春学期に開講予定の「回帰分析・中級」で取り扱います。興味のある方はぜひ受講して下さい（宣伝）。

21

C. 説明変数の計測誤差によるバイアス

- 変数が計測誤差を伴っている場合があります。
- 例えば、データがアンケート調査によるものだとしましょう。
- 回答した人は誤った答えを与えているかもしれません。
- また、アンケート調査で「昨年の年収は？」という質問があったとします。
- 回答者によっては、正確な年収を覚えていないため、ざっくりした値を記入するかもしれません。
- 説明変数が不正確に計測され計測誤差を伴う場合、最小二乗推定量はバイアスします。
- 見てみましょう。

22

- 母集団の回帰式が

$$Y_i = \beta_0 + \beta_1 X_i + U_i, \qquad E(U_i|X_i) = 0$$

だとします。そして母集団からのランダムサンプルがあるとします。

- $E(U_i|X_i) = 0 \rightarrow COV(X_i, U_i) = 0$ですから、この式を最小二乗法で推定して何の問題もありません： 最小二乗推定量$(\hat{\beta}_0, \hat{\beta}_1)$は不偏性・一致性を持ちます。
- ここでX_iが誤って$\tilde{X}_i$と計測されているとしましょう。
- 計測誤差が生じるメカニズムによって計測誤差の性質が決まってきます。
- ここでは最もシンプルなものを考えます。

23

$$\tilde{X}_i = X_i + W_i$$

- W_iは計測誤差です。
- この計測誤差は次の性質を持つとします。

$E(W_i) = 0$（計測誤差は平均するとゼロ）
$VAR(W_i) = \sigma_W^2$（計測誤差の分散は一定）
$COV(W_i, U_i) = 0$（計測誤差と回帰式の誤差項に相関無し）

- 母集団の回帰式は$Y_i = \beta_0 + \beta_1 X_i + U_i$で、$X_i = \tilde{X}_i - W_i$を代入すると

$$Y_i = \beta_0 + \beta_1 \tilde{X}_i + [-\beta_1 W_i + U_i] \rightarrow Y_i = \beta_0 + \beta_1 \tilde{X}_i + V_i$$

- 分析者がY_iを$\tilde{X}_i$に回帰させ最小二乗法で推定するとどうなるでしょうか？

24

$$\tilde{X}_i = X_i + W_i$$

- W_iは計測誤差です。
- この計測誤差は次の性質を持つとします。

  $E(W_i) = 0$（計測誤差は平均するとゼロ）
  $VAR(W_i) = \sigma_W^2$（計測誤差の分散は一定）
  $COV(W_i, U_i) = 0$（計測誤差と回帰式の誤差項に相関無し）

- 母集団の回帰式は$Y_i = \beta_0 + \beta_1 X_i + U_i$で、$X_i = \tilde{X}_i - W_i$を代入すると

$$Y_i = \beta_0 + \beta_1 \tilde{X}_i + [-\beta_1 W_i + U_i] \rightarrow \boxed{Y_i = \beta_0 + \beta_1 \tilde{X}_i + V_i}$$

- 分析者がY_iを$\tilde{X}_i$に回帰させ最小二乗法で推定するとどうなるでしょうか？
- V_iはW_iを含みます、そしてW_iは$\tilde{X}_i$の一部です→ $\tilde{X}_i$とV_iは相関します。
- 最小二乗推定量$\hat{\beta}_1$はバイアスします。

25

$$Y_i = \beta_0 + \beta_1 \tilde{X}_i + [-\beta_1 W_i + U_i], \quad \tilde{X}_i = X_i + W_i$$

ケース1: $\beta_1 > 0$ の時

- $\tilde{X}_i$とW_iは正に相関、だから$\tilde{X}_i$と$-\beta_1 W_i$は負に相関。
- 最小二乗推定量$\hat{\beta}_1$は下にバイアスする。

ケース2: $\beta_1 < 0$ の時

- $\tilde{X}_i$とW_iは正に相関、だから$\tilde{X}_i$と$-\beta_1 W_i$は正に相関。
- 最小二乗推定量$\hat{\beta}_1$は上にバイアスする。

==========

$plim$を使って計算するともう少し言える。

$$plim\hat{\beta}_1 = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_W^2}\beta_1$$

となることが示せる（証明は省略）。$0 < \frac{\sigma_X^2}{\sigma_X^2 + \sigma_W^2} < 1$だから

最小二乗推定量$\hat{\beta}_1$はゼロの方向にバイアスする
（XがYに与える影響を過小に評価する）

26

説明変数の計測誤差によるバイアス：対処法

- できる限りXに関する正確な指標を得てください。
- また計量的手法で説明変数の計測誤差によるバイアスを回避することが可能です。
- そのような方法の一つが、欠落変数バイアスのところでも登場した「操作変数法」です。
- また最近開発された方法（本講義のレベルをはるかに超えますが）としては、

Lewbel, A. (2012) "Using Heteoskedasticity to Identify and Estimate Mismeasured and Endogenous Regressor Models," *Journal of Business and Economic Statistics*, 30, 67-80.

27

D.同時双方向の因果関係によるバイアス

- ここまでは、説明変数から従属変数への方向の因果関係を前提としてきました。
- Xが原因、Yが結果ですね。

- しかし、もし逆の関係も同時に存在するとすればどうなるでしょうか？

- このような双方向の因果関係が存在する場合、最小二乗推定量はバイアスします。

28

- 式を使って見てみましょう。

$$Y = \beta_0 + \beta_1 X + U$$
$$X = \gamma_0 + \gamma_1 Y + V$$

- 単純化のために、$COV(U,V) = 0$ とします。
- 一本目の式でUが正の値を取るとします。→ $\uparrow Y$

ケース1：$\gamma_1 > 0$のとき

$\uparrow Y$のとき二本目の式から、$\uparrow Y \rightarrow \uparrow X$

従って、このときには$COV(U,X) > 0$

29

- 式を使って見てみましょう。

$$Y = \beta_0 + \beta_1 X + U$$
$$X = \gamma_0 + \gamma_1 Y + V$$

- 単純化のために、$COV(U,V) = 0$ とします。
- 一本目の式でUが正の値を取るとします。→ $\uparrow Y$

ケース1：$\gamma_1 > 0$のとき

$\uparrow Y$のとき二本目の式から、$\uparrow Y \rightarrow \uparrow X$

従って、このときには$COV(U,X) > 0$

最小二乗推定量$\hat{\beta}_1$は上にバイアス

30

- 式を使って見てみましょう。

$$Y = \beta_0 + \beta_1 X + U$$
$$X = \gamma_0 + \gamma_1 Y + V$$

- 単純化のために、$COV(U,V) = 0$ とします。
- 一本目の式でUが正の値を取るとします。→ $\uparrow Y$

ケース2：$\gamma_1 < 0$のとき

$\uparrow Y$のとき二本目の式から、$\uparrow Y \rightarrow \downarrow X$

従って、このときには$COV(U,X) < 0$

31

- 式を使って見てみましょう。

$$Y = \beta_0 + \beta_1 X + U$$
$$X = \gamma_0 + \gamma_1 Y + V$$

- 単純化のために、$COV(U,V) = 0$ とします。
- 一本目の式でUが正の値を取るとします。→ $\uparrow Y$

ケース2：$\gamma_1 < 0$のとき

$\uparrow Y$のとき二本目の式から、$\uparrow Y \rightarrow \downarrow X$

従って、このときには$COV(U,X) < 0$

最小二乗推定量$\hat{\beta}_1$は下にバイアス

32

- 式を使って見てみましょう。

$$Y = \beta_0 + \beta_1 X + U$$
$$X = \gamma_0 + \gamma_1 Y + V$$

- 単純化のために、$COV(U, V) = 0$ とします。
- 一本目の式でUが正の値を取るとします。→ $\uparrow Y$

ケース1：$\gamma_1 > 0$のとき

$\uparrow Y$のとき二本目の式から、$\uparrow Y \rightarrow \uparrow X$

従って、このときには$COV(U, X) > 0$ 最小二乗推定量$\hat{\beta}_1$は上にバイアス

ケース2：$\gamma_1 < 0$のとき

$\uparrow Y$のとき二本目の式から、$\uparrow Y \rightarrow \downarrow X$

従って、このときには$COV(U, X) < 0$ 最小二乗推定量$\hat{\beta}_1$は下にバイアス

33

- 極端なケースを見てみましょう。

$$Y = \beta_0 + U$$
$$X = \gamma_0 + \gamma_1 Y + V$$

- XはYに全く影響を与えない、逆にYがXに影響を与えるというケースです。
- 簡単に言えば、「因果が逆」のケースです。
- 一本目の式でUが正の値を取るとします。→ $\uparrow Y$
- ケース1：$\gamma_1 > 0$のとき

$\uparrow Y$のとき二本目の式から、$\uparrow Y \rightarrow \uparrow X$

従って、$Y = \beta_0 + \beta_1 X + U$を最小二乗法で推定すると、$COV(U, X) > 0$だから

最小二乗推定量$\hat{\beta}_1$は上にバイアス

XはYに全く影響を与えないのにもかかわらず、XはYに正の影響を与えるという推定結果になりがち

34

- 極端なケースを見てみましょう。

$$Y = \beta_0 + U$$
$$X = \gamma_0 + \gamma_1 Y + V$$

- XはYに全く影響を与えない、逆にYがXに影響を与えるというケースです。
- 簡単に言えば、「因果が逆」のケースです。
- 一本目の式でUが正の値を取るとします。→ $\uparrow Y$
- ケース1：$\gamma_1 < 0$のとき

$\uparrow Y$のとき二本目の式から、$\uparrow Y \rightarrow \downarrow X$

従って、$Y = \beta_0 + \beta_1 X + U$を最小二乗法で推定すると、$COV(U, X) < 0$だから

最小二乗推定量$\hat{\beta}_1$は下にバイアス

XはYに全く影響を与えないのにもかかわらず、XはYに負の影響を与えるという推定結果になりがち

35

同時双方向の因果関係によるバイアス: 対処法

- 計量的手法で同時双方向の因果関係によるバイアスを軽減することが可能です。
- そのような方法の一つが、欠落変数バイアスのところでも登場した「操作変数法」です。
- また、実証分析上よく用いられるプラクティカルな方法が、 XとYの時点をずらす、です。
  - 例えば、Xは一年前のデータ、Yは今年のデータ、といった具合です。
  - こうすることに、よって「逆の因果」の可能性を排除することができます。
- ただし、今年のXと今年のYのモデルと、一年前のXと今年のYのモデルは、厳密にいうと少し違ったものといえます。
- また、XとYの時点をずらすことは、欠落変数バイアスの問題の解決には一切役に立たないことを理解しておいて下さい。

36

E.標本（サンプル）セレクションバイアス

- 次の問題を考えてみましょう。
- ほとんどの大学では、講義評価アンケートがあります。
- 講義Aに対する評価は五段階で、1（とてもだめ）、2、3、4、5（とてもいい）とでもしましょう。
- 学生iの講義Aに対する評価をY_iとします。
- N人の学生が講義Aを履修しています。
- アンケート実施日に授業に出席した学生が評価を調査用紙に記入します。
- インターネットでアンケートに回答する、というのはありません。

37

- このとき授業の評価はどうなるでしょうか？
- 全員授業に出席しているなら、

$$\frac{1}{N}\sum_{i=1}^{N} Y_i$$

でいいですね。

- N人の学生を母集団そのものとするなら、これは母集団の平均（μ）そのものですから、講義への評価そのものといえますね。
- これが知りたいわけなのですが、、、
- 普通は、全員は授業に出席しませんよね。
- 二つのケースを考えましょう。

38

ケース1: 任意の学生が出席するかしないかは完全にランダム

- この場合は、事実上、母集団からのランダムサンプルになります。
- 従って、標本平均を推定すればいいですね。

$$\hat{\mu} = \left(\sum_{i:attend}^{N} Y_i \right) \Big/ \text{出席人数}$$

- この推定量は母集団の平均（μ）の推定量として問題ないです。
- 不偏性・一致性持ちます。

＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝

- 標本が抽出されるプロセスは実際は違いますよね、、、
- リアルなやつ行きましょう。

39

ケース2: 講義内容に興味を持っている学生、講義内容に満足している学生の出席率が高い。

- この場合、出席している学生さんは母集団からのランダムサンプルではないですね。
- サンプルには、高いY_iを持つ学生さんがより多く含まれることになるでしょう。
- 従って、標本平均

$$\hat{\mu} = \left(\sum_{i:attend}^{N} Y_i \right) \Big/ \text{出席人数}$$

上にバイアスします（$E(\hat{\mu}) > \mu$）。

- 本当の評価よりいい評価をもらいがちになりますね。

40

- これは

「標本（サンプル）セレクション・バイアス」

と呼ばれるものの一例です。

- 今の例には、説明変数が入っていませんでしたが、説明変数を入れてもバイアスが生じることになります。
- 例えば、授業の評価Y_iを左辺にして、

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_K X_{Ki} + U_i, \quad E(U_i | X_{1i}, \ldots, X_{Ki}) = 0$$

みたいなモデルを考えましょう。

- X_{ki}は学生さんの観測可能な属性（例：女性ダミー、年齢、学年、所属学部、GPAなど）。
- ランダムサンプルなら最小二乗法で問題無しですが、、、

41

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_K X_{Ki} + U_i$$

- Y_iが観測された（出席した）人だけをデータとして使うと、、、
- そのデータの中には、U_iが大きい人がより多く含まれることになってしまうかもしれません。
- そのため$E(U_i | X_{1i}, \ldots, X_{Ki}) = 0$ が成り立たなくなってしまうかもしれません。
- もしそうなれば、最小二乗推定量$\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_K$はバイアスします、「**標本（サンプル）セレクション・バイアス**」ですね。
- 標本セレクションバイアスが起こりうる代表的な例としては、賃金のモデルが挙げられます。
  - 働いている人だけ賃金が観測される（出席した人だけに評価が観測される）。

42

標準誤差の推定におけるバイアス

- ここまでは、係数の推定上の問題について説明してきました。
  - どういうときに、最小二乗法は不偏性・一致性を持たないのか？
  - 対処法は？（多くの対処法はこの講義の範囲外が、、、）
- ここからは、

(2)「仮説検定が設定された通りの有意水準を持つこと、また信頼区間も設定された通りの信頼水準を持つこと」を危うくする要因

について説明します。

- 「仮説検定が設定された通りの有意水準を持つ、また信頼区間も設定された通りの信頼水準を持つ」ためには、

推定量の標準誤差が正しく推定されている必要があります。

43

- 言い換えれば、

「仮説検定が設定された通りの有意水準を持つこと、また信頼区間も設定された通りの信頼水準を持つこと」を危うくする要因

とは、

「どういうときに標準誤差が誤ったものになってしまうか」

の「どういうとき」です。

- 以下では、どういうときに標準誤差の推定にバイアスが生じるかについて説明します。

44

F. 不均一分散

- 講義ノート15で取り扱った誤差項の不均一分散です。
- 誤差項の均一分散を仮定して推定した標準誤差は、もしその仮定が間違っていた場合（すなわち不均一分散だった場合）、正しいものではありません。
- 標準誤差にバイアスが生じるということです。
- その場合、仮説検定や信頼区間の計算には使えません。
  - 仮説検定の結果が誤ったものになるかもしれない。
  - 誤った信頼区間になるかもしれない。
- この問題の対処法は、「不均一分散に対して頑健な（ロバストな）標準誤差を計算する」でしたね。

45

G. 誤差項の観測値間の相関

- ある設定の下では、誤差項が観測値間で相関することがあります。
- この問題はこれまでは考えてきませんでした、というより、考える必要がありませんでした。
- というのは、この講義ではランダム・サンプリング（データが母集団からランダムに抽出される）を仮定していたからです。
  - ランダム・サンプリングの下では、ある観測値の誤差項の分布と別の観測値の誤差項の分布は独立であることが保証されます。
- しかし、実際には、サンプリングが部分的にしかランダムでない場合があります。

46

- 例えば、同じ主体についてデータが時間を通じて繰り返し観測される場合、すなわちパネルデータの場合です。
- もし誤差項に含まれる要因が持続的であるならば、回帰式の誤差項には時間を通じた「**系列相関**」と呼ばれるものが発生します。
- 式で表すと

$$Y_{it} = \beta_0 + \beta_1 X_{it} + U_{it}$$

というモデルで、U_{it}（今期の誤差項）とU_{it-1}（一期前の誤差項）が相関する、という具合です。

- これは時系列データでも起こり得ます。

47

- 別の状況でも、誤差項が観測値間で相関することはありえます。
- 例えば、サンプリングが地理的な単位で行われることがあります。
- もし地理的な影響を反映するような要因が誤差項に含まれる場合、同一地域（や隣接した地域）における観測値の間で、誤差項が相関するかもしれません。
- 例えば、式で表すと、

$$y_{ir} = \beta_0 + \beta_1 X_{ir} + U_{ir} \quad （地域r\text{の企業}i）$$
$$y_{jr} = \beta_0 + \beta_1 X_{jr} + U_{jr} \quad （地域r\text{の企業}j）$$

で$COV(U_{ir}, U_{jr}) \neq 0$

48

重要:

このような誤差項の観測値間の相関は、最小二乗推定量$\hat{\beta}_0, \hat{\beta}_1, \cdots \hat{\beta}_K$にバイアスを生じさせません。

- 問題になるのは標準誤差です。

誤差項の観測値間の相関がある場合、均一分散を仮定した場合の標準誤差、不均一分散に対して頑健な標準誤差ともに不正確なものとなります。

49

誤差項の観測値間の相関：対処法

- 不均一分散、系列相関のどちらの可能性も考慮にいれて標準誤差を計算することもできますし、、、、
- 不均一分散、同一クラスター(例:地域、産業など)の観測値間の相関のどちらの可能性も考慮にいれて標準誤差の計算をすることもできます。
- 「クラスターロバスト標準誤差」で調べてみて下さい。
  - Stataでは、vce(cluster クラスターの名前)がコマンドになります。
  - regコマンドの最後に , robustとつけるのと同じ感じですね。

50

外部の正当性を危うくする要因

- 「統計分析が外部に正当である(externally valid)」とは、、、、

因果効果に関する統計的な推論が、より一般に他の母集団や設定にも妥当なもので、結果の一般化が可能である状況のこと

- 分析される母集団からのランダム・サンプルを使って、因果効果に関する推定結果を得たとします。
- この結果は、他の母集団においても成り立つでしょうか？
- 一般に、因果関係の真の効果は、分析される母集団と関心ある母集団とで同じではないかもしれません。
- なぜなら、母集団の特性が違ったり、地理的に違ったり、時間が違ったり、するからです。

51

- また「設定」の違いにより、分析結果を一般化できないかもしれません。
- 制度的な環境が違ったり、、、
- 法律が違ったり、、、
- 物理的な環境が違ったりすれば、
- 因果関係の真の効果は違ったものになるかもしれません。

52

外部の正当性の評価

- 外部の正当性は、どちらかというと計量的な話ではなく、、、
- 分析された母集団・設定と関心ある母集団・設定の、それぞれに関する知識を使って判断されるものです。
- 重要な違いがあるなら、外部の正当性について疑念が生じます。
- 異なる母集団に関して、複数の分析結果がある場合は、、、
  - 複数の研究で同様の結果が得られているなら、外部正当性は強くなる。
  - 結果が異なるなら、外部正当性は疑わしいと判断します。

53

- 多くの学術論文では、「結論」の章において、外部正当性（この言葉は必ずしも使いませんが）について簡単に議論します。
- 多くのケースでは、
  - 本論文の結果は、これこれこういう理由で、必ずしも一般化できないかもしれない、、、
  - 違う国のデータを使って、本論文の結果が成り立つかどうかを分析してみるのは意義深いと考えられる、、、

なんて感じで書かれています。

54

回帰式を予測に使う際の内部正当性

- 回帰モデルは、因果関係の効果を推定するだけでなく、予測にも使われます。
- もしも目的がYの予測なのであれば、内部の正当性（推定量のバイアスの問題）はあまり重要ではありません。
- これはどういうことでしょうか？
- 単回帰モデル

$$Y = \beta_0 + \beta_1 X + U$$

は、XがYに与える因果効果（他の要因を一定にした時の、XがYに与える効果）の推定が目的なら、ほぼほぼ使い物にはなりません。

- 特殊なケースを除けば、欠落変数バイアスが生じるからです。

55

- しかし、このモデルは予測に有用かもしれません。
- Yの動きを説明できる情報がXに入っているなら、XはYの予測に役立ちます、単回帰モデルだとしてもです。
- 予測はとにかく当てればいい、当てるモデルがえらいです。

56