（ほぼ）図（だけ）で説明する回帰分析

筒井淳也

2011.7

1　決定モデルの場合

計量分析では、何のために回帰分析をするのでしょうか？

たとえば、「学歴が収入に与える影響」について知りたいとしましょう。このとき、ランダム抽出されたデータを使って、学歴ごとの収入の平均値を求めて、その差をみるだけではダメです。なぜなら、学歴と収入の両者に影響する要因（たとえば性別）があって、その効果のために学歴の収入に対する効果が過大（あるいは過小）に見積もられる可能性があるからです。学歴によって性別が変わるわけではない以上、性別の効果は学歴と切り離して考えるべきなのは当然でしょう。

ここで、もし収入に影響する要因が性別（男女）と学歴（高卒と大卒のみ、という世界を考える）だけであるとしよう。このときは、男女の高卒・大卒、計 4 グループからひとりずつ誰でもよいので抽出して、その 4 点から回帰係数を計算すればいいです。図 1 左は、性別と学歴に**交互作用**がない場合です。交互作用がある場合は右のようになります。つまり、学歴の効果が性別ごとに異なる（男性において大卒の効果がより大きい）場合です。

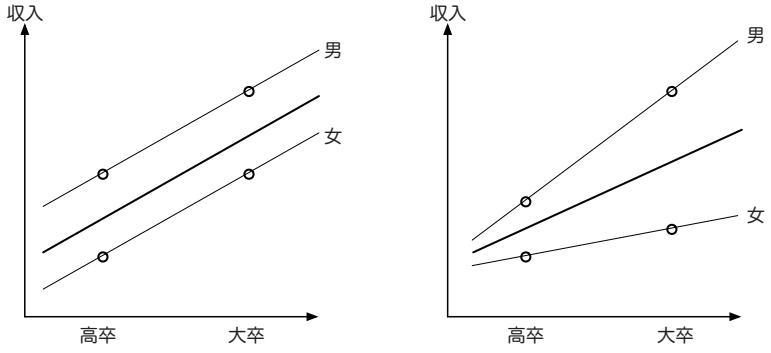

図 1　回帰直線（右は交互作用がある場合）

このとき、性別変数を投入せずに学歴のみで回帰したら、図の太線の回帰直線が推定されます。ただし、それは性別と学歴が相関しない場合です。女性で高卒が多く、男性で大卒が多いというふうに、説明要因どうしが相関しているとしましょう。このとき、性別と学歴の両方をモデルに投入していれば、図 2 左の 2 本の細い線が推定され、性別と学歴のそれぞれの正しい効果が推定されます。しかし学歴のみ投入して性別を投入して

いないと、太い点線のような間違った（バイアスのかかった）推定がなされてしまいます。

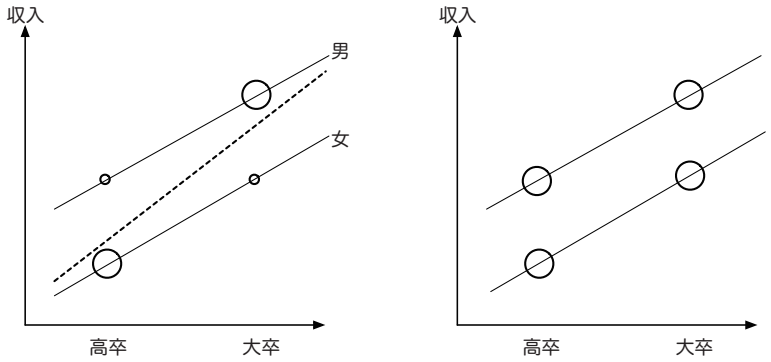

図 2　交絡がある場合の回帰直線のバイアス

ここで注意すべきなのは、以下の点です。**もし収入が性別と学歴のみで説明される場合、標本における性別と学歴の構成割合は、母集団を反映している必要がない**ということです。たとえば母集団におけるこの 4 グループの構成割合が表 1 の左、そこから抽出したのが右のとおりだったとしましょう。構成割合は母集団と標本で全く異なっています。

表 1　被説明変数の構成割合

母集団	女	男	標本	女	男
大卒	10	20	大卒	1	1
高卒	20	10	高卒	1	1

このときでも、もし性別と学歴の両方を観察して重回帰モデルに投入していれさえすれば、性別と学歴の効果の両者とも正しく推定されます。

もし性別変数を投入しない場合はどうでしょう？　このときに正しい学歴の効果を推定するには、図 2 の右のように、学歴ごとに性別構成を同じ割合にしてやる必要があります。繰り返し注意点。母集団を反映させた割合にすると、学歴の効果に間違って性別の効果が含まれてしまうのでダメです。男女の構成割合は、学歴で同じでさえあれば、あとはどうでもよいのです（たとえば「女性だけ」でも大丈夫）。

要するに、**Y ← X という回帰分析にある変数 W を追加投入する**ということは、**X における W の構成割合が同じであるとしたときの X の Y に対する効果**を推定する、ということです。「男でも女でも学歴構成が同じ」だとすれば、男女の効果を測定するときに学歴の効果は入ってきません。こういった状態を計算上で作り出すことができるのが、回帰分析の優れた点です。

2　撹乱項がある場合

さて、上の例では収入は性別と学歴だけで決まっていると想定していましたが、実際にはそのようなことはないでしょう。収入を決定する要因には、年齢、勤続年数、個人の能力など、様々な要因が絡んできます。

$$\text{収入} = f(\text{性別、学歴、年齢、勤続年数、...}) \tag{1}$$

すべての要因を観察できるわけはないので、計量分析では観察されていない要因をまとめて「撹乱項」にぶち込みます。

$$\text{収入} = f(\text{学歴}) + \epsilon \tag{2}$$

ϵ（イプシロン）が撹乱項です。収入の決定要因のうち、学歴以外の要因です。このような場合、観察値は図 3 の左側のようになります。つまり、同じ高卒（大卒）でも様々な要因によって収入にバラつきがでるわけです。収入 $= \beta$学歴 $+ \epsilon$ というモデルで回帰分析をする場合、高卒の収入の平均値と大卒の収入の平均値を通る直線が回帰直線として推定されます。この直線から個々の観察値までの距離が、撹乱項の値です（撹乱項は個々のケースによって異なった値を取ります)。

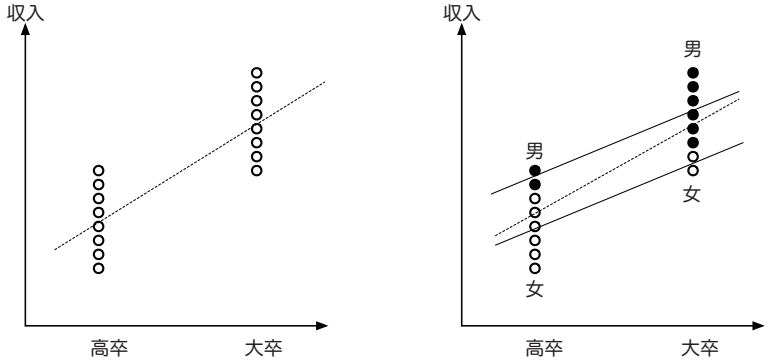

図 3　撹乱項

さて、このように求めた回帰直線の傾きは、正しく学歴の収入に対する効果を表しているといえるでしょうか？　たぶん言えません。図 3 の右側は、男を黒丸で、女を白丸で表現しています。学歴の正しい効果は、高卒と大卒の男女それぞれの平均値を通る 2 本の実線になります。点線の傾きよりも緩やかになっていることがわかります。これは、点線の効果（傾き）が、本来学歴の効果ではない効果（ここでは性別の効果）を間違って含みこんでしまっていたからです。

このように、撹乱項（それぞれの説明変数における収入の平均値からの距離）のなかに、説明変数と相関する要因が含まれていたとき、当の説明変数だけで回帰分析しても正しい効果を得ることはできません。要するに、**撹乱項（未投入の要因）の中にもし説明変数と相関するものがあれば、観察して投入しなければならない**というのが、回帰分析の原則です。「撹乱項と説明変数の相関」というと難しく聞こえてしまうのですが、忘れてはならないのは、撹乱項とは未観察要因の集積である、ということです。「撹乱項と説明変数の相関」とはしたがって、「モデルに投入した説明変数と、撹乱項の中に隠れた説明要因との相関」ということです。説明変数と被説明変数の両者に影響するような要因のことを、**交絡要因**といいます。

学歴の効果に関する交絡要因が性別しかない場合、その他の撹乱項の中に残された要因は観察しなくても、学歴の効果はきちんと（バイアスなく）推定されます。

$$収入 = f(学歴、性別) + \epsilon \tag{3}$$

ただし交絡要因ではなくても、収入に大きく影響する要因は観察してモデルに投入するべきです。たとえば図 4 は性別が学歴の交絡要因になっていない場合です。このとき、ここから標本をいくつか抽出してそこから回帰直線を計算するとき、どの点（人）を抽出するかによって回帰直線は変わってきます。図 4 左側の A は高卒・大卒ともまたまたプラスの撹乱項を持つ人を抽出してしまった場合です。係数には誤差が生じませんが、切片にプラスの誤差が生じます。B は、高卒でたまたまマイナスの撹乱項を、大卒でたまたまプラスの撹乱項を引き当ててしまった場合です。このときは係数にプラスの誤差が発生します。

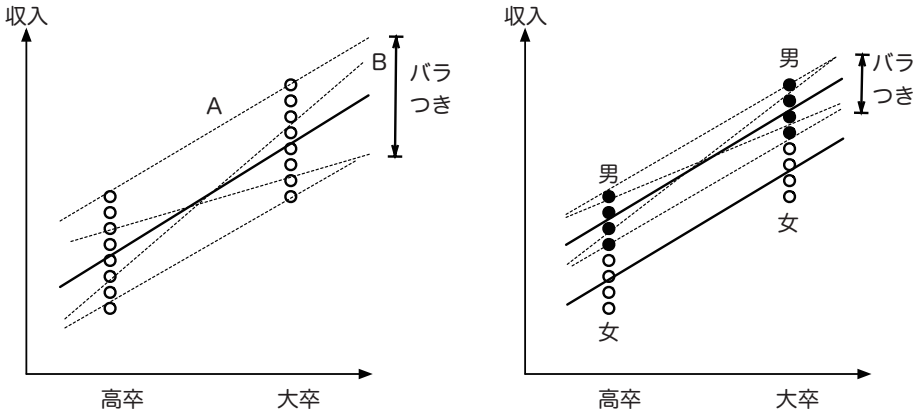

図 4　誤差の大きさ

誤差の範囲は、撹乱項のちらばりの大きさによって決まります。このとき、性別を観察してモデルに投入すれば、図 4 右側では性別を観察しているので、男女それぞれで回帰直線がばらつく余地が小さくなっています。その分、学歴の効果の誤差が小さくなり、より精度の高い推定が可能になるわけです。

3　信頼区間

信頼区間とはなんでしょうか。たとえば「95% 信頼区間」の場合、その範囲の外に真の値が入る確率が 5% しかないよ、ということです。このあたりの説明は省略しますが、一度に観察する標本サイズが大きければ大きいほど信頼区間が狭くなって、精度の高い推定が可能になります。

信頼区間は、説明変数の個々の値をとる部分標本サイズによって決まります。図 5 左側の点線は、高卒と大卒の標本サイズが同じ場合の信頼区間です。信頼区間の様々な点上を通るいろんな直線を引いてみてください。図の点線の範囲に入るように回帰直線を描くことができることがわかります。右側は高卒の標本が大卒の標本よりも多い場合の信頼区間です。高卒の方が標本サイズが大きい分推定精度が上がり、信頼区間（真の値が 95% の確率でこのあいだに入っている範囲）が狭くなります。

さて、図 5 左側の場合、実は「学歴は収入に有意な効果をもたらさない」と判断されます。というのは、信頼区間が重なっているからです。ということは、高卒と大卒で真の年収の平均値が同じである確率もそこそこある、ということになるからです。これに対して右側だと、高卒と大卒の信頼区間は重なっていないので、「高卒と大卒の平均収入が同じである」ということは非常に考えにくい、ということになります。

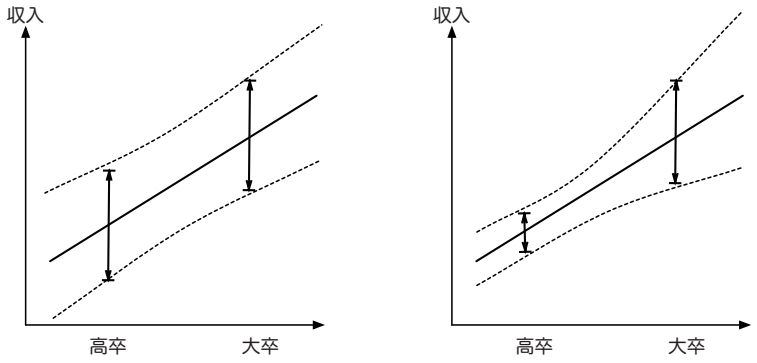

図 5　信頼区間

説明変数が複数あって、ひとつが連続量でもうひとつがカテゴリーの時は、信頼区間は次の図 6 のようになります。左側では、男女とも若年層において標本サイズが小さいため、信頼区間が広がり差が検出できていませんが、高年齢層だと差が有意になります。右側は年齢の効果が男女で異なるという交互作用効果があるモデルです。標本サイズは年齢ごとに同じなので信頼区間の広さも同じですが、交互作用効果のせいで若年層において差が検出されません。

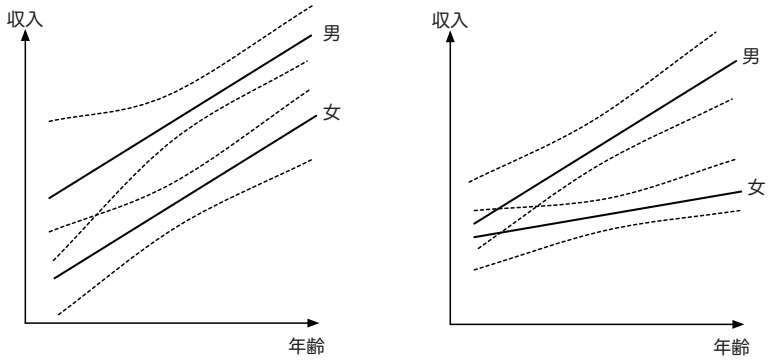

図 6　信頼区間 (2)

回帰分析をしたら、ぜひ関心のある説明変数についての、被説明変数の予測値の信頼区間をプロットしてみてください。（ソフトウェアに慣れることが必要になりますが...。）そうすることで、より豊富な情報を回帰分析の結果から引き出すことができます。特にダミー変数を多く含んだモデルについては、デフォルトの結果の表から各ダミー変数の差の有意性を読み取ることは難しいですから、各ダミー変数ごとの予測値の信頼区間をプロットしたほうが親切だと言えます。たとえば図 7 をみると、どのグループ間に有意な差があるか一目して分かります。

さて。

以上の推定では、説明変数のどの部分でも被説明変数のバラつきが同じである、という前提（いわゆる**分散均一の仮定**）で話をしてきました。信頼区間を決めるのは、撹乱項全体の分散と、説明変数の個々の値における標本サイズでした。しかし実際には、何らかの要因で分散不均一になることが多いので、その際には普通に回帰分析をした場合に算出される信頼区間はあまりあてになりません。

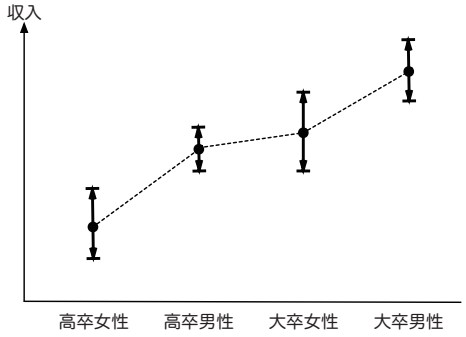

図 7　ダミー変数ごとの信頼区間

そこで活躍するのが、**ロバスト推定**です。ロバスト推定は、説明変数の値に応じて異なった誤差を計算するやり方です。図 8 の右側は、高卒と大卒の標本サイズが同じデータで、通常の回帰分析をしたときの信頼区間です。たとえ標本において大卒の方が撹乱項のバラつきが大きくても、回帰分析は「分散が同じである」という前提で信頼区間を計算します。このときロバスト推定をすれば、右図のように大卒の信頼区間が大きく推定されます。

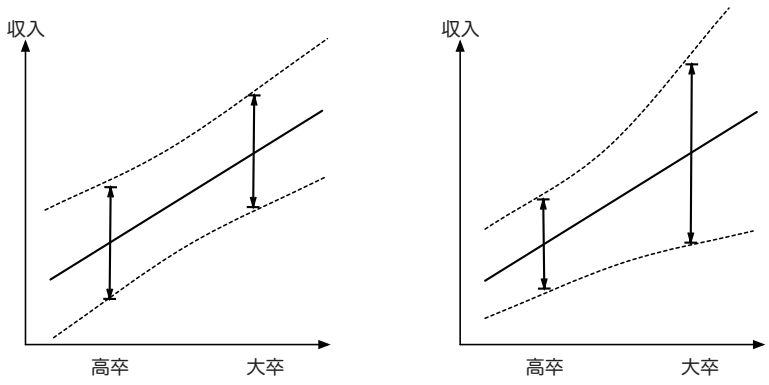

図 8　ロバスト推定

説明変数の値ごとに被説明変数の分散が同じであるという保証は通常ないので、基本的には分散均一の検定を行い、分散不均一が観察された場合にはロバスト標準誤差を計算した上で信頼区間をプロットするようにしたほうがいいでしょう。

分散不均一をもたらす原因はいくつか考えることができます。年収のように被説明変数がゼロ以下をとらない場合、ゼロ付近の分散は当然小さくなります。また、たとえば大卒のみに効く要因があってそれが観察できていない場合、その効果によって大卒のみで撹乱項のバラつきが大きくなる、ということも考えられます。

4　内生変数と外生変数

回帰分析では、変数は大きく分けて内生変数と外生変数に分けることができます。とはいえ内生と外生はそれ自体は相対的な概念で、たとえば図 9 左の場合、性別はいかなる変数からも作用を受けないので外生的（強

外生、とも）です。学歴は収入にとっては外生的ですが[*1]、性別から作用を受けるので、性別からすれば内生的です。

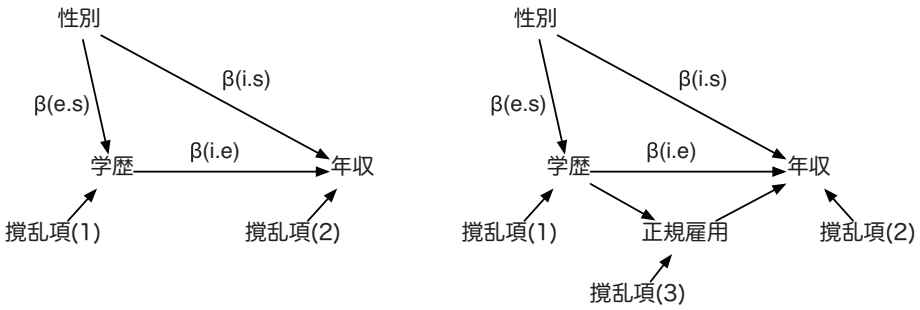

図 9　パス図

図では、学歴と年収に撹乱項からの矢印が付いています。これは、たとえば撹乱項 (1) については、学歴が性別以外の要因（たとえば親の所得）によっても影響を受けていることを表しています。もし撹乱項 (1) がなければ、学歴は完全に性別によって規定されることになりますが、であるならばわざわざ性別と学歴を別々の変数として測定する必要がなくなります。現実にはそのようなことはないでしょう。ここで、撹乱項 (1) は性別と相関しないこと、撹乱項 (2) は性別および学歴と相関しないことが（厳密にはそれぞれの撹乱項のなかにそのような要因が残されていないことが）、それぞれの係数が正しく推定される条件です。

ここで、年収 (I) を性別 (S) と学歴 (E) で回帰したとき、

$$\mathrm{I} = \beta_{(i.e)}\mathrm{E} + \beta_{(i.s)}\mathrm{S} + \epsilon$$

という式が成り立ちます（変数を標準化して切片をなくしているとします）。このようにして推定された学歴から年収への作用 $(\beta_{(i.e)})$ からは、性別からの効果が除去されています。では性別から学歴への作用 $(\beta_{(e.s)})$ はどこに言ったのかというと、実はこの回帰式自体からは見えません。重回帰モデルにおいては、複数の説明変数によって同時に説明される部分の情報は捨てられ、個々の変数が独自に説明できる部分のみが推定されます。

重回帰ではなくパス解析の世界では、2 変数間の表面上の相関を、他の変数を入れることによって分解するということが課題になります。その際、関心のある変数が内生的なのか外生的なのかによって分解の仕方が変わってきます。外生変数である性別に効果がある場合、

$$\text{性別と収入の相関} = \text{性別の収入に対する総合効果} = \text{直接効果}\ (\beta_{(i.s)}) + \text{間接効果}\ (\beta_{(e.s)} \times \beta_{(i.e)})$$

となります。つまり相関は直接経路と、学歴を介した間接経路に分解できるわけです。$\beta_{(e.s)}$ は重回帰の結果からはわかりませんが、（以下で述べる擬似相関がない以上）性別の総合効果は要するに何も変数を追加しない状態での性別の効果そのものです。

ある変数 X の Y に対する総合効果とは、X が変動したときに、他の変数を経由する分を含めて Y がどれだけ変動するか、という効果のことです。上の例の場合、「性別が学歴経由で収入に及ぼした影響」は性別の直接効果ではなく間接効果ですが、それでも性別の効果の一部である、ということができる、ということです。

[*1] ただし、収入を得たので大学に通うようになった、というケースがあれば厳密には話は別です。

このように、**外生変数の効果に関心があるときは、「○○の効果の一部を△△が説明した」のように因果関係を細分化・詳細化 (elaborate) していく**ことが分析の焦点になります。ここでは、「性別の効果の一部は学歴によって説明された」となります。

これに対して内生変数である学歴に興味がある場合は、少し事情が異なります。学歴と収入の単純な相関は、以下のように分解されます。

$$\text{相関} = \text{総合効果}\ (= \text{直接効果}\ (\beta_{(i.e)})) + \text{擬似相関}\ (\beta_{(e.s)} \times \beta_{(i.s)})$$

つまり、単純な相関は、本来の学歴の作用である総合効果と、性別によるニセの作用である擬似相関に分解できます。ここでは学歴と収入のあいだに媒介する変数が設定されていませんが、もしそれがある場合（図 9 右のような場合)、総合効果はやはり直接効果と間接効果の合計になります。このように、内生変数に関心があるときは、**単純な相関から擬似相関を除去すること**、「△△の効果から○○の効果を除去した結果、△△の効果はそれでも説明力があることがわかった」といった説明になります。

以上のように、回帰分析においても「ある変数を入れたときにもとの変数の効果が変わる」ということが、どのように説明できるのかを考えていく必要があります。しばしば「諸変数をコントロールする」という言い方でごまかされることもありますが、**ある変数を投入することによる元の変数の効果の変化が、擬似相関の除去によるものなのか、それとも間接効果による媒介によるものなのか**については、常に意識しておいてください。

性別は強外生ですから、性別の効果が変数の追加投入で変化するときは、すべて間接効果による変化です。これに対して学歴の場合は、性別を追加投入して学歴の効果が変化するときは擬似相関の除去による変化ですが、正規雇用（ダミー）変数の投入による変化は間接効果による媒介のせいである、といえるでしょう（図 9 右)。

しばしば「規定要因分析」という名前で、関心のある被説明変数を説明しそうな要因をざーっと並べて、「他の変数が一定だとすればこの変数の効果は...」のように書かれていることがありますが（私もやったことがある)、率直に言ってあまりよい回帰分析の使い方ではありません。関心のある説明変数はなにか、その変数は他の変数に対して外生的か内生的かを考えて、必要ならば追加の変数投入による変数の効果の変化を丁寧に見ていくことが必要です。