Chapter 2

Regression

p47~55

Chapter 1では…

- “Randomized Trials”

→因果効果を測定するのに関して（実現可能な範囲で）最も理想的な状況を、“実験”することによって作り出す

→もう少しだけ具体的に言えば、

無作為に、十分大きなサンプルサイズを持つように、適切な、グループ分けを行い、それぞれで得られたデータを使って因果効果を測定

（詳細はChapter 1参照）

利点

- 理論上、（実現可能な範囲で）最も理想的と考えられる状況を作り出し、そのもとで因果効果を測定できるので、最良の方法と言える。

しかし…

この“最良の方法”は“万能な方法”ではない。

何故か？

問題点

- （特に経済学のような分野では、）倫理観、労力等、様々な制約によって、“実験”を行うことが往々にして困難

→実際には、既に手元にあるデータを上手く利用して、理想的な状況を近似的に作り出すことが多い

→Chapter 2では、その方法の一つとして Regression を学ぶ

Chapter2(p47~55)の概要

- 導入
- 学費の差＝教育の質の差？

→より高い学費が、将来の収入にどれほど影響を与えているかを測定したい

→性別、成績、両親の収入等、将来の収入に影響を与えうる要素が多数存在するので、何か工夫が必要

→“Matchmaker(仲介人？)”の考え方

導入

- 先にも述べたように、Randomized Trialsは様々な理由によって適用できない場合がある
- しかし、計量的な技術を適切に使うことで、実験を行うのと同等に（あるいはそれ以上に）因果効果を測定できる
- そのような計量的な技術の最も基本となるのが、本章で紹介する“regression”

Regressionの基本的な考え方

- 『鍵となるobserved variables（観測変数）が、介入群と比較群の間で等しくなっているのであれば、直接観測することが出来ない変数（潜在変数）から生じるselection biasは、大抵取り除かれている』

→どういうことか？

具体例を用いて考えていく

学費の差＝教育の質の差？

- 2012年度における米国大学の学費

→私立大学：約29000ドル/年

公立(州立)大学（in their home state）：9000ドル未満/年

→20000ドル/年にも及ぶ学費の差は、果たしてそれに見合うだけの価値の差をうみだすのか？

補足：米国の大学について

- 大学区分

→主に、私立大学と州立大学に大別される

→国立大学は、軍事学校などの特殊な学校のみ

- 大学のレベル

→一般に、私立大学のほうがレベルは高い

→本文に出てくる、Harvard, Princeton, Duke, Stanford, MITなどは全て私立大学

（Ivy Leagueは米国有名私立大学グループを指す）

補足（続き）

- 学費（2018年度平均）

→私立大学：35830ドル/年

州立大学：10230ドル/年（州内出身者）

26290ドル/年（州外出身者）

（参考資料）

https://www.ryugaku.com/kind/

https://www.ryugaku.ne.jp/knowledge/flow/select/match/budget.html

高額な学費の価値を何で測るか？

- 大学を選ぶ際には、より高額な将来収入の見込み、一生の友人や将来の配偶者との出会いなど様々なことを考慮しうる
- とはいえ、余分に高額な学費を払おうとする時、そこにはより高額な将来収入への期待が含まれている

→（金が全てではないが、）ここでは、大学選択が将来収入にどれほどの影響を与えたのかを測定していく

適切な測定方法は？

- 例えば、Harvardを卒業した人が、Harvardの代わりにthe University of Massachusetts(U-Mass)を卒業していた場合に、いくら収入を得るかを測定できれば、学費と将来収入の関係を適切に測定できそう

→しかし、これは非現実的

→では、私立大学卒業生と州立大学卒業生の二つのグループ間で収入を比較してみたらどうか？

比較結果

- この方法で実際に比較してみると、私立大学卒業生の方が相当高額な収入を得ていることが分かる

→しかし、これでは学費と将来収入の関係を適切に測定できていないのは明白

→Chapter 1でも学んだように、ここには“selection bias”が存在

Selection biasの存在

- 念のために、ここではどのようなselection biasが存在しているのかを確認する

（例）HarvardとU-Massの比較

・Harvardの学生は、一般的に見て、学力水準はより高く、より意欲的であり、更には何か他の技術や才能を有している可能性もある

・Harvardに入学するのは相当難しいのに対して、U-Massでは、ある程度の成績を持つほぼ全てのMassachusetts出身出願者を受け入れ、更には奨学金まで提供している

→学力や能力が高く意欲的な学生ほど、Harvardを好んで選ぶ？

ではどうするか？

- Chapter 1で見た通り、Randomized Trialsによってこの種のselection biasは取り除くことが出来る
- しかし、この場合Randomized Trialsは実行できない

→様々な大学と学生の出願、選考等に関するこれまでのデータを用いて、Randomized Trialsに近いことが出来る

Serendipitous variation

- このとき、多くの意思決定や選択には、財政的な考慮やその人を取り巻く環境、タイミングなどから生じる“Serendipitous variation”が含まれるということが重要

- serendipity（形 serendipitous）：掘り出し物を偶然見つける才能、予期することなく大きな発見をすること、掘り出し上手（出典　ジーニアス英和辞典）

→つまり…

serendipitous variation：一見恵まれていないと思える環境で成功を収める才能の差異（？）

二人の友人の例

①Nancy

・Harvardとthe University of Texas(UT)の両方から入学を許可される

・多額の学資援助が受けられることを理由に、Harvardよりもレベルの低いthe UTを選択

・その後、彼女は経済学の教授として別のIvy League schoolに勤めた

二人の友人の例（続き）

②Mandy

・Duke, Harvard, Princeton, Stanfordからの誘いを断り、出身州の州立大学であるthe University of Virginiaにて学士を取得

・その後、Harvardの教員になる

→より多くのサンプルを用意できれば、より一般的なことが言えそうではあるが、それほど多くのサンプルを用意するのは困難

→ではどうするか？

ceteris paribus

- Serendipitous variationを切り離すためには、私立大学と州立大学のそれぞれを選ぶ学生間の、最も明白で重要な差異を一定に保つことが大切

→この方法で、『その他の事柄は全て等しい状態』（ceteris paribus）にしようとしている

→ceteris paribusを例を用いて見ていく

UmaとHarveyの例

- 仮定：人生において（少なくとも収入に関して）重要なのは、SATの成績とどの学校に行くかだけである
- Harvey→・SATの成績は1400点
  ・Harvardに入学
- Uma→・SATの成績は1400点
  ・U-Massに入学

→このときはceteris paribusになっているので、二人の収入を比較すれば大学の選択と収入の関係を調べられる

UmaとHarveyの例（続き）

- 当然、現実ではより複雑になる

例えば…

・Umaは若い女性、Harveyは若い男性

・差別、産休、育休などの影響によって、女性は同等の教育を受けた男性に比べて収入が少ない

→この場合、HarveyがUmaよりも20％多く収入を得ていることは、Harvardの優れた教育の効果かもしれないし、他の要因から生まれる男女間の賃金格差を反映しただけかもしれない

matching estimator

- 先の例で、SATの成績と学歴以外に重要な要素が性別だけであるならば、Harveyの代わりに、SATの成績が1400点でありHarvardに入学した女性学生HannahをUmaと比較すれば良い
- 同様のことを個人レベルではなく、HarvardグループとU-Massグループを用いて行い、平均的な収入格差を計算すれば、より一般的な解が求まる

→このとき、ここで求まった平均的な収入格差を、『性別とSATの成績を固定したmatching estimator』といい、Harvardの教育が収入に対して持つ平均的な因果効果を表す

Matchmaker(仲介人？)

- 現実には当然、性別や学歴、SATの成績以外にも収入に影響を与える要因は多数存在する
- これらの要因は、数が膨大な上に、学生の性格など数値化しにくいものを含んでいることもあり、その全てをコントロールするのはほぼ不可能

→Stacy Berg Dale＆Alan Krueger：

『全てを確認する代わりに、学生が出願し入学許可を受けた大学の特徴を、様々な要因についての情報を与える“仲介人”として考えられるのでは？』

UmaとHarveyの例

- 二人とも、U-MassとHarvardの両校に出願し入学許可も受けた

→出願状況から、二人にはU-MassやHarvardに行くモチベーションがあり、その両校から入学許可を受けたことから、二人ともそれらの大学で成功を収めるだけの能力があったのだと言える

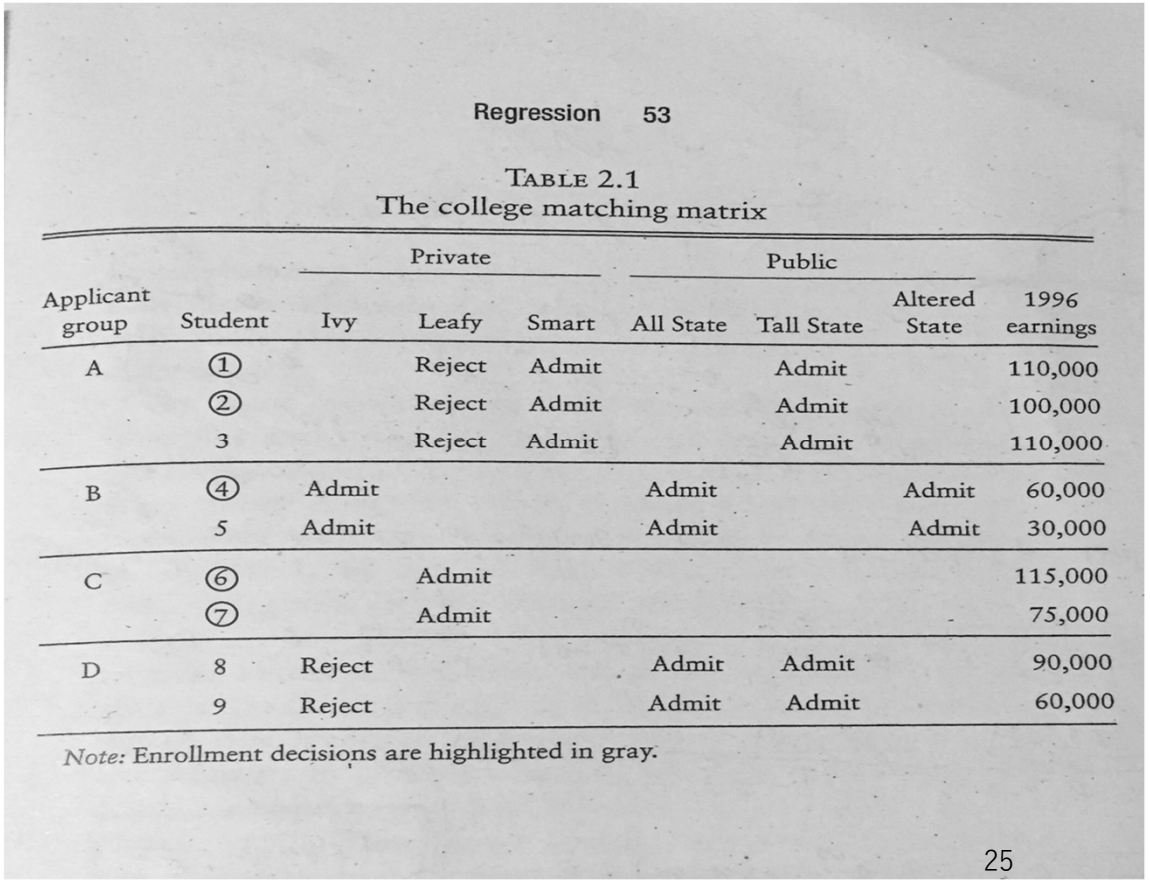
- しかし、UmaだけがU-Massを選んだ
- これは、成功を収めた叔父がU-Mass出身だったとか、Umaの能力とはあまり関係ない要因によるかもしれない

→こうした要因がUmaとHarveyにとって決定的なものだったのなら、この二人は（比較するのに）良い組み合わせである

college matching matrix

- DaleとKruegerは、様々な大学の学生について、彼らのSATの成績や卒業後の収入を調べた膨大なデータを、良い組み合わせを探しそれらを用いることで分析した
- 右に示したTABLE 2.1はその分析の簡略化版

→“college matching matrix”



Regression 53

TABLE 2.1
The college matching matrix

Applicant group	Student	Private			Public			1996 earnings
		Ivy	Leafy	Smart	All State	Tall State	Altered State	
A	①		Reject	Admit		Admit		110,000
	②		Reject	Admit		Admit		100,000
	3		Reject	Admit		Admit		110,000
B	④	Admit			Admit		Admit	60,000
	5	Admit			Admit		Admit	30,000
C	⑥		Admit					115,000
	⑦		Admit					75,000
D	8	Reject			Admit	Admit		90,000
	9	Reject			Admit	Admit		60,000

Note: Enrollment decisions are highlighted in gray.

TABLE 2.1について

- ９人の学生について、彼らの出願状況と入学を許可されたか否か、そして卒業後の収入が書かれている
- 大学は、まず私立大学と州立大学に分けられ、そのそれぞれが３つずつにレベル分けされている
- グループA,B,C,Dは、出願状況と入学を許可されたか否かの情報に基づいて学生を４つのグループに分けたもの
- 赤い丸で囲った学生（１,２,４,６,７）は私立大学に入学
- 他の４人は州立大学に入学

考察

- このデータから私立大学に行くことの収入への効果を調べたい

- 最も短絡的に考えれば…

1,2,4,6,7の平均収入＝92000ドル

3,5,8,9の平均収入＝72500ドル

→これらの差である約19500ドルが解

考察（続き）

- グループA,B,C,Dは、出願校とそのそれぞれから入学許可を受けたか否かが各グループ内で揃えられているので、matchmakerの考え方から、各グループ内での比較は“ceteris paribus”により近い状況を生み出すと言える

→これを用いることで、より適切と思われる解が求まる

グループ内比較

- グループAについて：

・出願校は私立大学２校、州立大学１校→上流中産階級の家庭？

・学生１,２は私立大学に、３は州立大学に入学

・１,２の平均収入＝105000ドル

　３の（平均）収入＝110000ドル

→これらの差である ー5000ドルが解？

グループ内比較（続き）

- 同様にしてグループBに関してもグループ内比較を行えば、計２つのグループについて、各グループ内での、私立大学に行くことの収入への効果が求まる（グループCには私立大学に入学した学生しかいないので、グループ内での比較はこの場合無意味）（グループDも同じような理由でこの場合無意味）

→グループA：－5000ドル

グループB：60000－30000＝30000ドル

考察

- この二つの値を用いれば、求めたい解のより適切な推定値が得られる

→方法1（相加平均）：（－5000＋30000）/ 2 ＝12500ドル

方法2（加重平均）：

（－5000×0.6）＋（30000×0.4）＝9000ドル

→学生数で重み付けをしている方法2の方が、より適切と言える

考察（続き）

- 最初に求めた19500ドルと、より適切と考えられる9000ドルの間の差は何故生じているのか？（ちなみに、グループA,Bの５人について最初に試した方法で効果を測っても、20000ドルとなってやはり9000ドルとは相当差がある）

→“selection bias”が存在している

→ここでは、私立大学に出願し入学許可を受けた学生は最終的にどこに入学しようがより高い収入を得る、つまり私立大学の方が能力やモチベーションが高い学生の割合が大きい、ということ

→グループ内比較をすることでselection biasを排除

考察（続き２）

- また、このselection biasは、グループA,Bについてグループ間比較を行うことで理解できる

→A（出願校の三分の二が私立大学）：平均収入＝約107000ドル
　B（出願校の三分の二が州立大学）：平均収入＝45000ドル
→Aの学生の方が平均収入がかなり高い
　＝能力やモチベーションが高い学生ほど私立大学への出願率が高い？

個人的考察

- AとBの比較では、出願校の内訳に加えて合否状況についても考慮すれば、Aの学生はBの学生に比べて能力というよりモチベーションが高かったと言えそう

- Cの学生はともに、私立大学だけに出願し合格していることから、能力はそれなりに高いと言えそう（実際に、平均収入はAとそこまで変わらない）

- ただ、Cの学生には自信家な傾向もみられるので、その性格ゆえにグループ内での格差がやや大きいか？

個人的考察（続き）

- BとDは、出願した私立大学と州立大学の比が同じで、合格状況を考慮すればBの学生の方が優秀に見えるが、平均収入はDの方が高い

→謎。Dの学生は行きたかったエリート私立大学に落ちて、渋々入学した州立大学で悔しさをバネに頑張ったのに対して、Bの学生は自分の能力を過信し、あぐらをかいていた？？？

まとめ

- 経済分析など、実験を行うことが困難な場合には、手元にあるデータをいかに工夫して分析するかが重要
- その工夫の一つの方法が“Regression”
- 膨大な数の潜在変数を全てコントロールすることは出来ないので、それらを上手く要約していると考えられる要素“matchmaker(仲介人)”を代わりに用いる（本文の例では学生の出願校とその合否）
- “matchmaker”を揃えたうえで比較することで、“ceteris paribus”により近い状況を生み出すことが出来る