MASTERING ‘METRICS
Chapter 2 Regression

pp.56~68

前回の復習

- 因果効果を測定するにはランダム化比較試験を行うのが良い
- しかし、現実にはさまざまな制約（コスト、倫理的問題…）がある
  →いつもランダム化比較試験が行えるわけではない

- そこで、理想に近い状況を作り出す方法の一つが回帰分析
- 回帰分析はランダム化比較試験に比べてカジュアルな方法。回帰分析は、より精密な分析を行う足がかりとなる。

前回の復習（回帰分析の性質）

- 回帰分析の正当性の根拠：「効果を調べたい変数（ここでは私立 or 公立）以外の条件」を揃えれば、セレクションバイアスがほとんど排除される

- 「効果を調べたい変数以外の条件」は現実には無限にあるが、そのうち重要なもの(matchmaker)に着目することで分析は近似できる
- （Matchmakerがそろっているデータは*ceteris paribus*に近い）

前回の復習（C&B data set）

・14,000人以上の人の大学受験結果、進学先、収入のデータ

・TABLE 2.1は9人の学生のデータを表にまとめたもの

・受験結果が同じ人たちをグループA~Dに分類

・全体での比較、グループ内での比較、グループとグループの比較をおこなった
→正確な結果を求めるには、データをどう工夫して用いるかが重要

TABLE 2.1
The college matching matrix

		Private			Public			
Applicant group	Student	Ivy	Leafy	Smart	All State	Tall State	Altered State	1996 earnings
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.

Make Me a Match, Run Me a Regression

- 回帰分析は、より実証的な分析を行うためのベンチマークとなる

- 回帰分析の流れ

Matching matrixを作る（例：TABLE2.1)

↓

様々な要素の比較を行う

↓

それに従って重みのついた平均をとる

回帰分析の式の要素

- 従属変数(*dependent variable*)（被説明変数、結果変数ともいう）
  ここでは（学生）iの将来の収入。Y_i

- 説明変数(*treatment variable*)
  ここではダミー変数であり、学生が私立大に進学したかを表す。P_i

- 制御変数(*control variable*)
  ここでは、学生が受験した大学と合格した大学の組み合わせ、つまりグループを表す。

ここでの制御変数について

- 先ほどのTABLE2.1では、グループAとBは有用なデータで、グループCとDは役に立たないデータだった（なぜなら、AとBは同一グループ内に私立進学者と公立進学者が両方いたのでグループ内での比較ができたが、CとDはグループ内の進学先が同じで、比較ができなかった）

- ということは、グループAとBだけを考えるとすると、どのグループに属しているかという情報はダミー変数で表すことができる

  →そのダミー変数をA_iとする

ダミー変数(dummy variable)

- ここで、P_iとA_iはダミー変数
- ダミー変数は「ある状態にあてはまるか、否か」を表すために使う

- $P_i = \begin{cases} 1 & (\text{学生}i\text{が私立進学者のとき}) \\ 0 & (\text{学生}i\text{が公立進学者のとき}) \end{cases}$
- $A_i = \begin{cases} 1 & (\text{学生}i\text{がグループAに属するとき}) \\ 0 & (\text{学生}i\text{がグループBに属するとき}) \end{cases}$

回帰分析の式の要素

- 従属変数(*dependent variable*)（被説明変数、結果変数ともいう）
  ここでは学生iの将来の収入。Y_i

- 説明変数(*treatment variable*)
  ここではダミー変数であり、学生が私立大に進学したかを表す。P_i

- 制御変数(*control variable*)
  ここでは、学生が受験した大学と合格した大学の組み合わせ、つまりグループを表す。

回帰式 (2.1)

- $Y_i = \alpha + \beta P_i + \gamma A_i + e_i$ (2.1)

- α：切片
- β：説明変数の効果（ここでは、私立大に行くことによる効果）
- γ：グループAに属することによる効果
- e_i：残差(residual)(または誤差項(error term))
  fitted valueとの差

Fitted value

- Fitted value→あてはめ値、予測値

$$\hat{Y}_i = \alpha + \beta P_i + \gamma A_i$$

- 残差　$e_i = Y_i - \hat{Y}_i = Y_i - (\alpha + \beta P_i + \gamma A_i)$
- 回帰分析では、$\hat{Y}_i$ をY_i に近づけたい
- つまりパラメータ(α, β, γ)を適切に決めることで残差（の二乗和）を最小化したい

  →最小二乗法 ordinary least squares (OLS)

具体例

- TABLE2.1のグループA,Bに回帰分析を行うと、次の予測が得られる
- $\alpha = 40{,}000$　（切片）
- $\beta = 10{,}000$　（私立大に行くことによる増分）
- $\gamma = 60{,}000$　（グループAに属することによる増分）
- βの値は前回求めた結果と比べてどうか？

β（私立効果）の検討

- 10,000という値も、各グループの値に重みをつけた平均だといえる（前回：Aでの効果＝-5,000、Bでの効果=30,000）

  →A:4/7　B:3/7　という重み付け

- βは、前回の方法で求めた結果と一致してはいないが、それほど遠くはない

  （前回：単純な平均=12,500、グループのサイズで重みをつけた平均=9,000）

- いずれにしても、何も工夫せずに平均をとったとき(19,500)よりは私立効果の値は小さく出ている。

Public-Private Face-Off

- C&Bデータセットについて：

  全部で14,000人を超える人のデータが収められている

  （各人は、進学先と別に検討した大学を三校以上回答している）

  →それぞれの人の受験経歴は多種多様

  もちろん、私立しか受験していない人、公立しか受験していない人もいる（ここではそのようなデータは有用でない）

より有用な比較を行うためには

- 前提：よりよい比較を行うためには、調べたい要素（私 or 公）以外の要素（個人の性質）がそろっているデータが多くあるほどよい

- ここでは個人の性質を「どの大学を受け、どこに合格し、どこに進学したか」という情報で表現できるのではないか

  →その情報がそろったデータのセットがたくさん欲しい

- 大学の分類方法→Barron’s selectivity categories

Barron’s selectivity categories

- 大学を、入試の合格率や入学者のレベルに基づいて
  - Most Competitive
  - Highly Competitive
  - Very Competitive
  - Competitive
  - Less Competitive
  - Noncompetitive

  に分類（ただし、ここで使われるのはMost Competitive, Highly Competitive, Competitiveの３種類のみ）

Barron’s selectivity categories

TABLE2.1での分類（上段）と、Barronの分類（下段）の対応は以下のとおり。
分類が６種類から３種類と、簡単になっている。

私立			公立		
Ivy	Leafy	Smart	All State	Tall State	Altered State
Most Competitive	Most Competitive	Highly Competitive	Competitive	Competitive	Highly Competitive

Barron’s selectivity categories

- よって、Barronの分類法で考えると、たとえば、
  ・Tall State(=Competitive), Smart(=Highly), Leafy(=Most)
  を受験して、Tall State と Smart に合格した人
  と
  ・All State(=Competitive), Smart(=Highly), Ivy(=Most)
  を受験して、All State と Smart に合格した人
  を、同じグループに属するとして比較することができる
  つまり、比較に使えるデータのセットが増える

Barron’s selectivity categories

- C&Bデータ全体のうち、9,202人がBarronの方法でマッチングできた。
- しかしここで問題：Barronの分類法では、私立しか受験していない人、公立しか受験していない人のパターンが出てくる。
- 私立と公立の差を測ることが目的なのでそのようなデータは省く

  →残ったのは5,583人
- このデータを、似ている人同士でグループにしたところ151のグループに落とし込むことができた。

回帰モデルの作成

- Barronの分類法に基づいて回帰モデル(結果は式(2.2))を作る。
- しかし、式(2.1)すなわちTABLE2.1に基づいて作った回帰モデルとは異なる点がある。

- $Y_i = \alpha + \beta P_i + \gamma A_i + e_i$ (2.1)

- $\ln Y_i = \alpha + \beta P_i + \sum_{j=1}^{150} \gamma_j GROUP_{ji} + \delta_1 SAT_i + \delta_2 \ln PI_i + e_i$ (2.2)

モデルの違い1

- 左辺
- Y_i (2.1) に対し、 $\ln Y_i$ (2.2)

- ここでlnは自然対数
- 対数をとると、値の変化が「何%変化したか」ということになる
  例えば、β（説明変数P_i の係数）が0.05と予測されるとき、私立出身者の収入は公立出身者の収入より約5%高いといえる。

モデルの違い2

- (2.1)は制御変数がA_iひとつだけ（A,Bの2グループのみだったから）
- (2.2)は多くの制御変数がある
- 主要なもの→どのグループに属しているか($\sum_{j=1}^{150} \gamma_j GROUP_{ji}$)

- $GROUP_{ji}$はダミー変数
- 例えば、学生1がグループ2に属しているとすると、

  $GROUP_{1,1} = 0, GROUP_{2,1} = 1, GROUP_{3,1} = 0, \ldots, GROUP_{150,1} = 0$

  （グループ151を表したいときはすべて0とすればよい）

モデルの違い3

- 式(2.2)では式(2.1)から制御変数が追加されている
- SAT_i：学生i のSATスコア
- $\ln PI_i$：学生i の親の収入（対数）
- 他にも、式(2.2)には書かれていないが、複数の制御変数（ダミー変数）が追加されている

  （性別・人種・運動経験・高校のランク）

式 (2.2)

- $\ln Y_i = \alpha + \beta P_i + \sum_{j=1}^{150} \gamma_j GROUP_{ji} + \delta_1 SAT_i + \delta_2 \ln PI_i + e_i$ (2.2)

- α, βは式(2.1)と同じ意味。
- $\sum_{j=1}^{150} \gamma_j GROUP_{ji}$：グループの制御関数を適切に設定することが、*ceteris paribus*な比較への大きな一歩となる。

Regressions Run

実際に、制御変数の有無で様々に回帰分析を行った結果がTABLE2.2

右半分(4)(5)(6)は受験した大学が同じ人たちをグループにまとめて分析をした結果
(Selectivity-group dummies=Yes)

左半分(1)(2)(3)はグループ化せず、様々な受験経歴の人が混ざった状態で分析をした結果
(Selectivity-group dummies=No)

TABLE 2.2
Private school effects: Barron's matches

	No selection controls			Selection controls		
	(1)	(2)	(3)	(4)	(5)	(6)
Private school	.135 (.055)	.095 (.052)	.086 (.034)	.007 (.038)	.003 (.039)	.013 (.025)
Own SAT score ÷ 100		.048 (.009)	.016 (.007)		.033 (.007)	.001 (.007)
Log parental income			.219 (.022)			.190 (.023)
Female			−.403 (.018)			−.395 (.021)
Black			.005 (.041)			−.040 (.042)
Hispanic			.062 (.072)			.032 (.070)
Asian			.170 (.074)			.145 (.068)
Other/missing race			−.074 (.157)			−.079 (.156)
High school top 10%			.095 (.027)			.082 (.028)
High school rank missing			.019 (.033)			.015 (.037)
Athlete			.123 (.025)			.115 (.027)
Selectivity-group dummies	No	No	No	Yes	Yes	Yes

Notes: This table reports estimates of the effect of attending a private college or university on earnings. Each column reports coefficients from a regression of log earnings on a dummy for attending a private institution and controls. The results in columns (4)–(6) are from models that include applicant selectivity-group dummies. The sample size is 5,583. Standard errors are reported in parentheses.

TABLE2.2 (1)

- (1)は、log earnings（収入の値の対数をとったもの）と説明変数（私立に行ったか否か）のみを用い、制御変数が全くない状態で回帰分析を行った結果
- .135 (.055)
- つまり私立進学者は公立進学者に比べ約14%高い収入（？）
- カッコ内の数字（.055）は標準誤差
- .135 は .055 の二倍以上の値なので、この結果は単なる偶然ではない→統計的に有意な結果

しかし……

- この約14%という差の中には、他の要因によるものもあるだろう
  →セレクションバイアスの存在

- そこで、他に様々な制御変数を入れて分析してみる
  （SATのスコア、親の収入、性別、人種など……）

- その結果がTABLE2.2 (2)(3)

TABLE2.2 (2)

- (1)のデータに加え、SAT（大学進学適性テスト）のスコアも考慮に入れる

- すると、SATのスコアが100点上がるごとに、収入は約5%(.048)増加した（この当時のSATは1600点満点）
- 一方で、私立効果は約10%(.095)に減少した

TABLE2.2 (3)

- 今度は、更に制御変数を増やす：
  親の収入、性別、人種、高校のランク、Athlete（運動部？）

- 親の収入高、高校のランク高、運動経験有→収入に正の効果
- 一方で、女性→収入の面で不利
- 人種→一概に言えない？（標準誤差が大きい）

- 私立効果は .086 に減少したが、やはり統計的に有意な結果

TABLE2.2 (4)

- グループに分類して計算していない(1)~(3)の結果はおそらく大きすぎる　←　正のセレクションバイアスが働いている

- (4)はSATスコア、親の収入などの制御変数は入っていない
- しかし、(1)~(3)との大きな違い：$\sum_{j=1}^{150} \gamma_j GROUP_{ji}$　の項の存在
  （TABLE2.2のSelectivity-group dummies = Yes）

TABLE2.2(4)(5)(6)

- (4)の私立効果はほとんど0。さらに標準誤差が0.038なので、統計的に有意な結果ではない。
- さらに、(5)(6)を見ると、制御変数を増やしても、私立効果の結果には改善が見られなかった。

- ここから言えること：グループ分け（受験経歴と進学先）という方法は、正当な回帰を行うための比較にはほど遠い
- 改善策は？　→次ページへ

TABLE2.2の考察、改善

- TABLE2.2で用いたサンプルの数は、Barronの分類法でマッチングできた5,583人のみ（元々のデータは14,000人以上）
- この限られたサンプルには何か特別なものがある

  (一般的な結果ではない？)

- この問題を解決するには、制御変数の縛りを緩めるのがよい
- 150個のダミー変数（ $\sum_{j=1}^{150} \gamma_j GROUP_{ji}$ ）を使うのではなく、「受験した大学の数」と「受験した大学の平均SATスコア」を使う

新しいモデルの作成

- 「受験した大学の数」：正確には、n校受験したか否かというダミー変数

  （たとえば、iさんが受験した大学の数が1校であるときには1の値をとり、そうでないときは0の値をとる、というダミー変数）
- この変数を使うと、C&Bデータセット14,000人以上全員のデータを利用することができる
- 以上の変数を使ったモデル

  →”self-revelation model” （自己発見モデル）

self-revelation model

- このモデルの根拠：参加者は、自分の能力や、自分はどこの学校なら受かりそうか、ということについて十分な知識を持っているはずだ
- この自己評価が「受験する大学の数」や「受験する大学の平均的なselectivity（＝SATスコア）」に反映される
- 原則として、弱い志願者は強い志願者に比べてselectivityの低い大学に出願し、受験する大学の数も少ない

self-revelation modelの結果

TABLE2.3（右図）：
14,238人の学生のデータを計算した結果
（カッコ内の数字は標準誤差）

・self-revelation modelの結果は、Barronの分類法による結果と類似したものになった

TABLE 2.3
Private school effects: Average SAT score controls

	No selection controls			Selection controls		
	(1)	(2)	(3)	(4)	(5)	(6)
Private school	.212 (.060)	.152 (.057)	.139 (.043)	.034 (.062)	.031 (.062)	.037 (.039)
Own SAT score ÷ 100		.051 (.008)	.024 (.006)		.036 (.006)	.009 (.006)
Log parental income			.181 (.026)			.159 (.025)
Female			−.398 (.012)			−.396 (.014)
Black			−.003 (.031)			−.037 (.035)
Hispanic			.027 (.052)			.001 (.054)
Asian			.189 (.035)			.155 (.037)
Other/missing race			−.166 (.118)			−.189 (.117)
High school top 10%			.067 (.020)			.064 (.020)
High school rank missing			.003 (.025)			−.008 (.023)
Athlete			.107 (.027)			.092 (.024)
Average SAT score of schools applied to ÷ 100				.110 (.024)	.082 (.022)	.077 (.012)
Sent two applications				.071 (.013)	.062 (.011)	.058 (.010)
Sent three applications				.093 (.021)	.079 (.019)	.066 (.017)
Sent four or more applications				.139 (.024)	.127 (.023)	.098 (.020)

Notes: This table reports estimates of the effect of attending a private college or university on earnings. Each column shows coefficients from a regression of log earnings on a dummy for attending a private institution and controls. The sample size is 14,238. Standard errors are reported in parentheses.

TABLE2.3 (1)(2)(3)

- TABLE2.3の左半分の３列を見ると、私立効果(Private School)は、他の制御変数が加わるにつれて減少している（この傾向はTABLE2.2のBarronのときと同じ）

  ただし、いずれも値は有意なものである

TABLE 2.2
Private school effects: Barron's n

	No selection controls		
	(1)	(2)	(3)
Private school	.135 (.055)	.095 (.052)	.086 (.034)
Own SAT score ÷ 100		.048 (.009)	.016 (.007)
Log parental income			.219 (.022)
Female			−.403 (.018)
Black			.005 (.041)
Hispanic			.062 (.072)
Asian			.170 (.074)
Other/missing race			−.074 (.157)
High school top 10%			.095 (.027)
High school rank missing			.019 (.033)
Athlete			.123 (.025)
Selectivity-group dummies	No	No	No

←TABLE2.2

TABLE2.3→

TABLE 2.3
Private school effects: Average SAT sco

	No selection controls		
	(1)	(2)	(3)
Private school	.212 (.060)	.152 (.057)	.139 (.043)
Own SAT score ÷ 100		.051 (.008)	.024 (.006)
Log parental income			.181 (.026)
Female			−.398 (.012)
Black			−.003 (.031)
Hispanic			.027 (.052)
Asian			.189 (.035)
Other/missing race			−.166 (.118)
High school top 10%			.067 (.020)
High school rank missing			.003 (.025)
Athlete			.107 (.027)
Average SAT score of schools applied to ÷ 100			
Sent two applications			
Sent three applications			
Sent four or more applications			

TABLE2.3(4)(5)(6)

- (4)(5)(6)では、受験校数、受験先のselectivityが制御変数に入っている
- (4)の私立効果は0.034と小さく、しかも統計的に有意ではない
- さらに、(4)(5)(6)を見比べると、本人の能力や家庭の状況などの制御変数を加えても、私立効果はほとんど変化していない

  これは、Barronのとき、つまりTABLE2.2(4)(5)(6)（グループ分けのダミー変数を追加したとき）の結果と似ている

私立効果はなかったのか？

- (1)(2)(3)よりもセレクションバイアスが取り除かれているはずの(4)(5)(6)では、私立効果（Private School）はほとんど認められなかった。

  →私立大学に行くことと将来の収入は無関係だったのか？

- →そうではない。比較の際の焦点の当て方を誤っていた。
- 筆者の考え：単純に、Ivyなどの私立大には良いクラスメイトがいる。優秀な学生たちによる相乗効果を受けられるという利点が、私立大の高額な学費を正当化する根拠になっているのではないか？

さらに新たなモデル：友人

- self-revelation modelで用いた私立効果のダミー変数を、peer（友人）の質を示す尺度で置き換えてみよう

- $\ln Y_i = \alpha + \beta P_i + \sum_{j=1}^{150} \gamma_j GROUP_{ji} + \delta_1 SAT_i + \delta_2 \ln PI_i + e_i$　(2.2)
- 式(2.2)のP_iを、「進学した大学のクラスメイトの平均SATスコア」に置き換える

TABLE2.4

TABLE2.4（右図）

項目はほとんどTABLE2.3と同じだが、
一番上の行に注目：
TABLE2.2,2.3では
”Private School”（私立大に行ったか、否か）
だった部分が
”School average SAT score ÷ 100”
（クラスメイトの平均SATスコア）
になっている

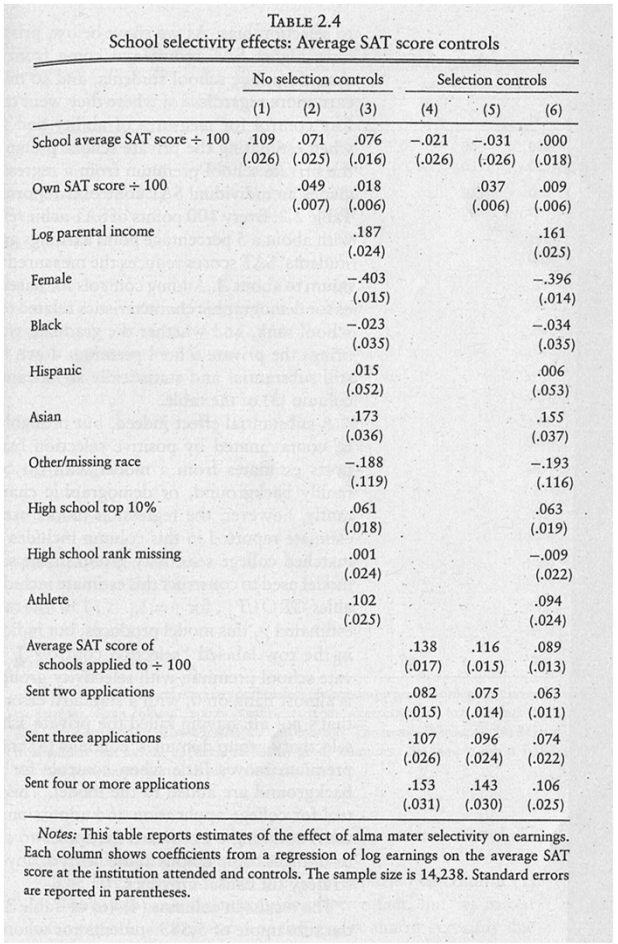

TABLE 2.4
School selectivity effects: Average SAT score controls

	No selection controls			Selection controls		
	(1)	(2)	(3)	(4)	(5)	(6)
School average SAT score ÷ 100	.109 (.026)	.071 (.025)	.076 (.016)	−.021 (.026)	−.031 (.026)	.000 (.018)
Own SAT score ÷ 100		.049 (.007)	.018 (.006)		.037 (.006)	.009 (.006)
Log parental income			.187 (.024)			.161 (.025)
Female			−.403 (.015)			−.396 (.014)
Black			−.023 (.035)			−.034 (.035)
Hispanic			.015 (.052)			.006 (.053)
Asian			.173 (.036)			.155 (.037)
Other/missing race			−.188 (.119)			−.193 (.116)
High school top 10%			.061 (.018)			.063 (.019)
High school rank missing			.001 (.024)			−.009 (.022)
Athlete			.102 (.025)			.094 (.024)
Average SAT score of schools applied to ÷ 100				.138 (.017)	.116 (.015)	.089 (.013)
Sent two applications				.082 (.015)	.075 (.014)	.063 (.011)
Sent three applications				.107 (.026)	.096 (.024)	.074 (.022)
Sent four or more applications				.153 (.031)	.143 (.030)	.106 (.025)

Notes: This table reports estimates of the effect of alma mater selectivity on earnings. Each column shows coefficients from a regression of log earnings on the average SAT score at the institution attended and controls. The sample size is 14,238. Standard errors are reported in parentheses.

TABLE2.4の分析

- TABLE2.4(1)(2)(3)を見ると、よりselectiveな（≒レベルの高い）大学に進学した人は労働市場で成功しているといえる

  （クラスメイトのスコアが100点上がるごとに収入は8%上がる）

- (1)(2)(3)にはセレクションバイアスが含まれているので(4)(5)(6)を見てみる　→　出願した大学の平均SATスコアと収入には本質的な関係は無い
- つまり、本人の能力と受験する大学はあまり関係ない？

  (ダメ元で突っ込む人もいれば、とても慎重な人もいる)

まとめ

- 回帰式に登場する要素を理解する：
  従属変数・説明変数・制御変数　また、「ダミー変数（Yes or No）」
- セレクションバイアスの排除を目指す
- 制御変数をどう設定するか？
  （分析で関係しそうな要素：グループ作成、家庭環境、性別……）
- 説明変数をどう設定するか？
  （最終的に、 私立か公立か：P_i ではなく、大学のクラスメイトの
  質を説明変数においていた）