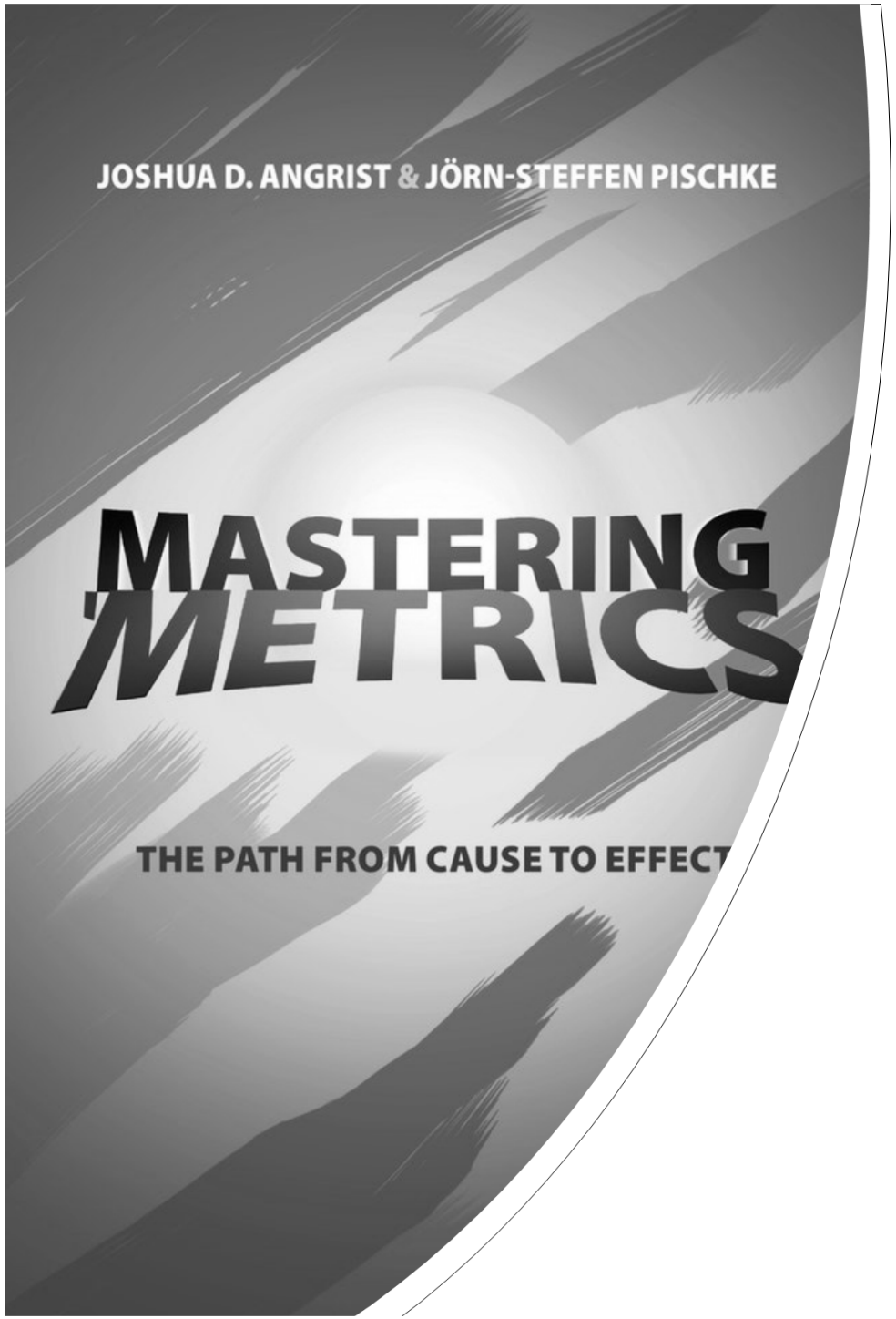


Ch 2. Regression

3.2. Ceteris Paribus?

前回の結果

私立大学への進学が収入に与える効果は Matchmaker を含めると消えてしまう。

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

再考：Harvey と Uma の例

Harvey

・SATの成績は1400点
・Harvard（私立大学）に入学

Uma

・SATの成績は1400点
・U-Mass（州立大学）に入学

⇒ 両者はほんとうに Ceteris Paribus か？

再考：Harvey と Uma の例

- Umaが州立大学に進学したのは、州立大学の入学料・授業料にしか利用できない奨学金を獲得したからかもしれない。
- 仮にそのことが影響していたなら、UmaはHarveyよりも**“貧しい”**ということになる。

- もちろん、前回の分析には、既に**両親の収入**変数が含まれていた。
- その一方で、**兄弟姉妹の数**変数は含まれていない。
- たとえ両親の収入が高くても兄弟姉妹の数が多いなら、子ども一人あたりの教育投資額は低くなる。

回帰分析と欠落変数バイアス

- **回帰分析**は、推定モデルに含まれる制御変数の観点で条件を一定にする方法である。

- 一方で、必要十分な制御変数を含み損ねれば、その欠落した変数の観点において条件が一定でなくなり、セレクション・バイアスを生む。

- 回帰分析におけるこのセレクション・バイアスを「**欠落変数バイアス**」と呼ぶ。
  - Omitted variables bias (OVB)

欠落変数バイアスの例示

- ５名のデータを再使用して、解説する。
- グループA所属ダミー変数が欠落している場合を考える。

		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

セットアップ

- グループA所属ダミー変数を含むモデル（Long）

$$Y_i = \alpha^l + \beta^l P_i + \gamma A_i + e_i^l. \qquad (2.3)$$

- そのダミー変数を含まないモデル（Short）

$$Y_i = \alpha^s + \beta^s P_i + e_i^s.$$

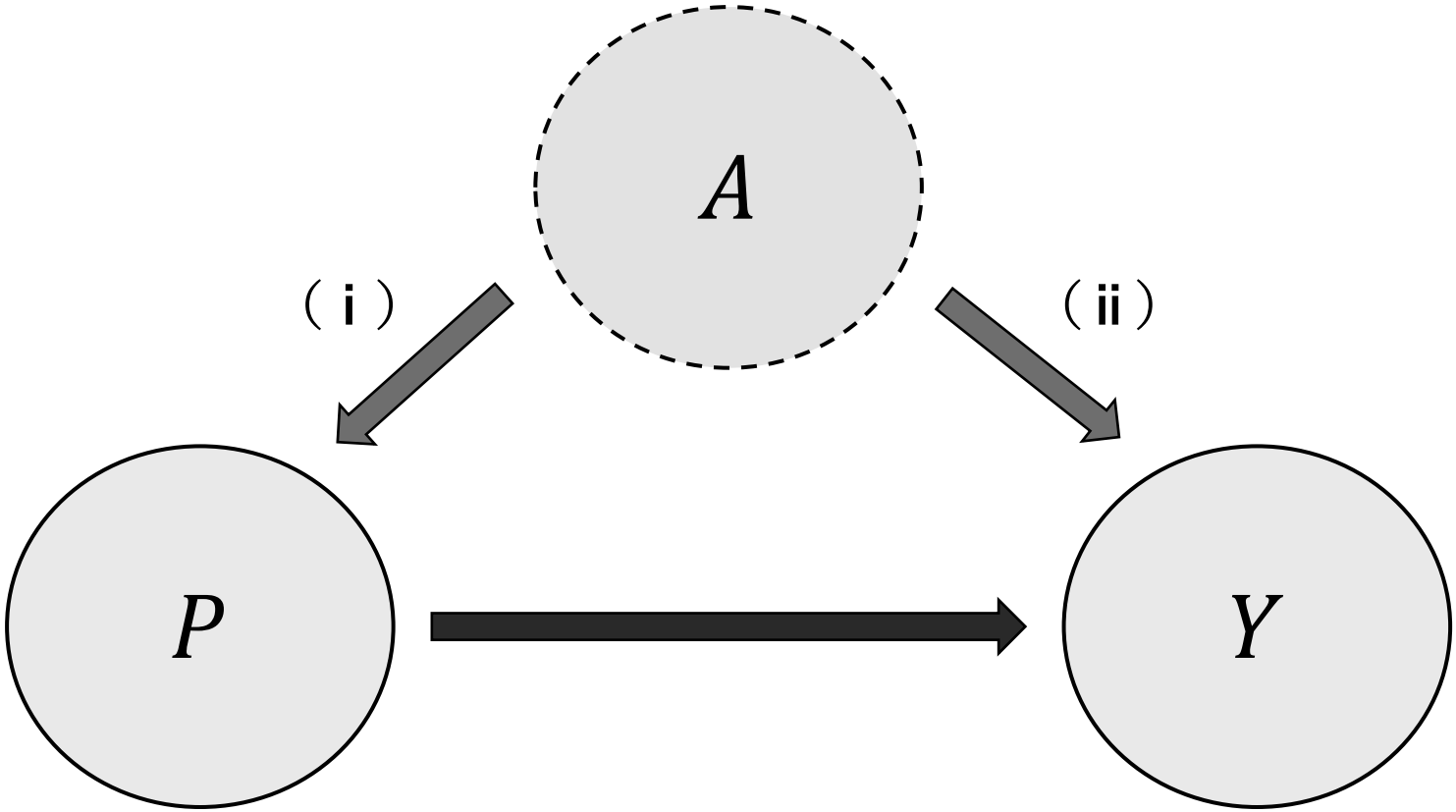
欠落変数バイアスの大きさ

- すでに2.1節の結果を確認したように、$\beta^s = 20{,}000$・ $\beta^l = 10{,}000$。
- よって、この場合の**欠落変数バイアス**は、

$$
\begin{aligned}
OVB &= Short - Long \\
&= \beta^s - \beta^l \\
&= 20{,}000 - 10{,}000 = 10{,}000
\end{aligned}
$$

欠落変数バイアス：10,000は何から生ずるか？

Long, Short の変数同士の関係



Long, Short の変数同士の関係

i. グループA所属ダミー変数（A）と私立大学進学変数（P）の関係

ii. グループA所属ダミー変数（A）と結果変数（Y）の関係

※ 2.3式の γ

《公式》欠落変数バイアス

- 下記の関係が成立する（らしい）

$$\beta^s \qquad\qquad \beta^l$$

$$\textit{Effect of } P_i \textit{ in short} = \textit{Effect of } P_i \textit{ in long} + \big(\{\textit{Relationship between } A_i \textit{ and } P_i\} \times \{\textit{Effect of } A_i \textit{ in long}\}\big).$$

$$\gamma$$

《公式》欠落変数バイアス

- つまり、第２項が欠落変数バイアスに相当する。

$$OVB = \{Relationship\ between\ A_i\ and\ P_i\} \times \{Effect\ of\ A_i\ in\ long\}.$$

《公式》欠落変数バイアス

グループA所属ダミー変数（A）と私立大学進学変数（P）の関係：

$$A_i = \pi_0 + \pi_1 P_i + u_i,$$

結果として、

$$\begin{aligned} OVB &= Effect\ of\ P_i\ in\ short - Effect\ of\ P_i\ in\ long \\ &= \beta^s - \beta^l = \pi_1 \times \gamma, \end{aligned}$$

《公式》欠落変数バイアスの導出

This central formula is worth deriving. The slope coefficient in the short model is

$$\beta^s = \frac{C(Y_i, X_{1i})}{V(X_{1i})}. \qquad (2.11)$$

Substituting the long model for Y_i in equation (2.11) gives

$$\frac{C\left(\alpha^l + \beta_1^l X_{1i} + \gamma X_{2i} + e_i^l, X_{1i}\right)}{V(X_{1i})}$$

$$= \frac{\beta_1^l V(X_{1i}) + \gamma C(X_{2i}, X_{1i}) + C(e_i^l, X_{1i})}{V(X_{1i})}$$

$$= \beta_1^l + \frac{C(X_{2i}, X_{1i})}{V(X_{1i})}\gamma = \beta_1^l + \pi_{21}\gamma.$$

計算確認

i. グループA所属ダミー変数（A）と私立大学進学変数（P）の関係

＝私立大学への進学率がグループAとBでどれだけ異なるか（$\underline{2/3 - 1/2 = 0.1667}$）

ii. グループA所属ダミー変数（A）と結果変数（Y）の関係

※ 2.3式で既に推定済み：γ

＝1996年の年収がグループAとBでどれだけ異なるか（$\underline{\text{2.2節より、}\ 60{,}000}$）

計算確認

$$
\begin{aligned}
OVB &= \{Regression\ of\ omitted\ on\ included\} \\
&\quad \times \{Effect\ of\ omitted\ in\ long\} \\
&= \pi_1 \times \gamma = .1667 \times 60{,}000 = \boxed{10{,}000.}
\end{aligned}
$$

留意点

- 現実のデータ分析ではデータセットに欠落変数が含まれていないことがほとんどである。

- つまり、欠落変数を含めた場合の推定結果と、含まない場合の推定結果を比較したり、両者の推定結果の情報を使って欠落変数バイアスの大きさを計算したりすることはできない。

⇒ では欠落変数バイアスの《公式》を理解することは、どのように役に立つのか？

再々考：Harvey と Uma の例

Harvey

・SATの成績は1400点
・Harvard（私立大学）に入学

Uma

・SATの成績は1400点
・U-Mass（州立大学）に入学

分析モデルには、**両親の年収**変数は含まれるが、
兄弟姉妹の数変数（$Family\ Size$）は欠落。

セットアップ２

- Longモデル

$$\ln Y_i = \alpha^l + \beta^l P_i + \sum_j \gamma_j^l GROUP_{ji} + \delta_1^l SAT_i + \delta_2^l \ln PI_i + \lambda FS_i + e_i^l, \quad (2.5)$$

- 私立大学進学変数と兄弟姉妹の数変数の関係

$$FS_i = \pi_0 + \pi_1 P_i + \sum_j \theta_j GROUP_{ji} + \pi_2 SAT_i + \pi_3 \ln PI_i + u_i. \quad (2.4)$$

《公式》欠落変数バイアス

$$
\begin{aligned}
OVB &= Short - Long \\
&= \{Relationship\ between\ FS_i\ and\ P_i\} \\
&\quad \times \{Effect\ of\ FS_i\ in\ long\}.
\end{aligned}
$$

以上より、

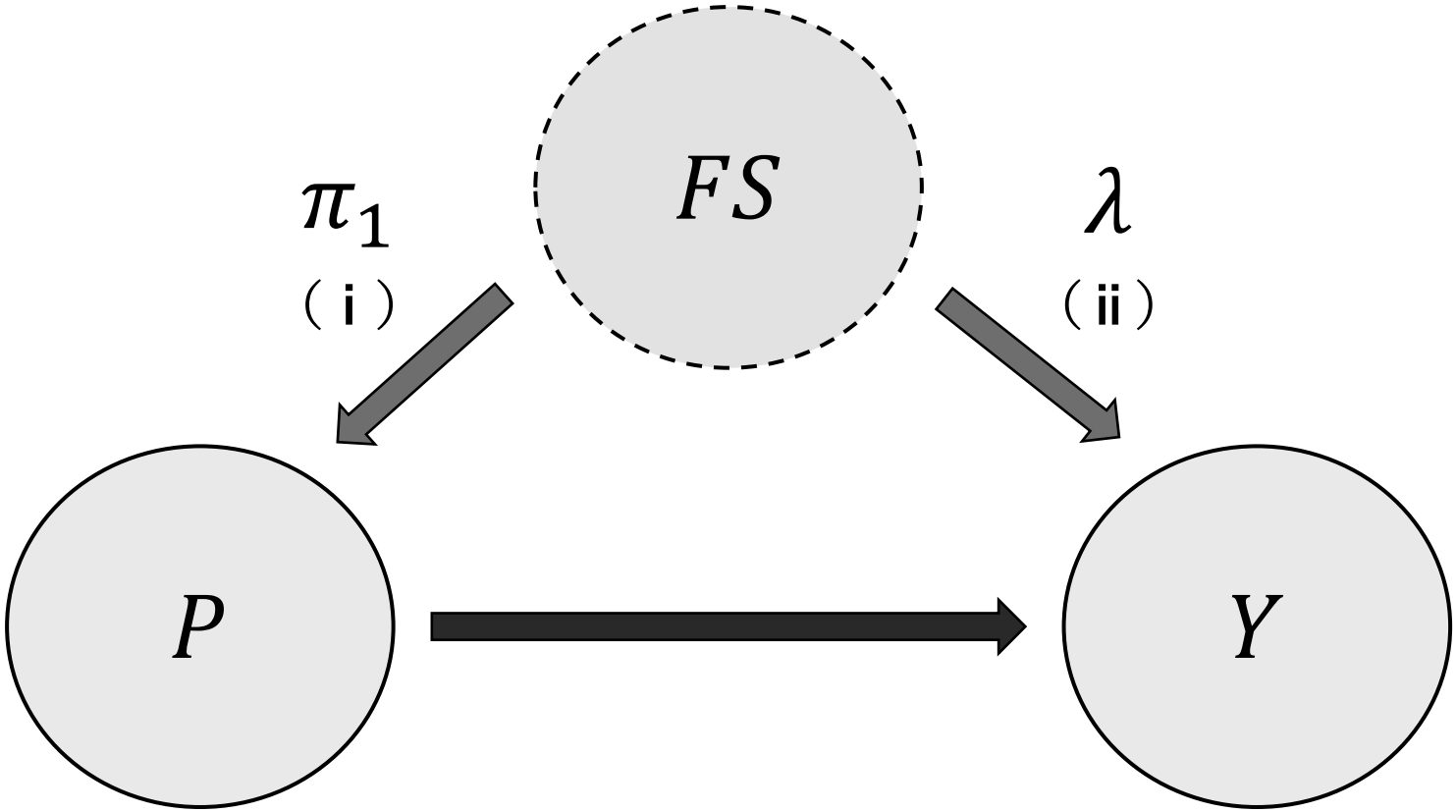
$$OVB = Short - Long = \beta - \beta^{l} = \pi_1 \times \lambda,$$

欠落変数バイアスの影響

- しかしながら、実際のデータセットには、**兄弟姉妹の数**変数は含まれない。
- よって、欠落変数バイアスの大きさを直接算出することはできない。

- 一方で、π_1 と λ の方向性に妥当な仮定を置くことで、**欠落変数バイアスの方向性を推論**することはできる。

欠落変数バイアスの影響



欠落変数バイアスの影響

i. 兄弟姉妹の数変数（FS）と私立大学進学変数（P）の関係：π_1

ii. 兄弟姉妹の数変数（FS）と結果変数（Y）の関係：λ

欠落変数バイアスの影響

$$OVB = Short - Long = \beta - \beta^{l} = \pi_1 \times \lambda,$$

- 仮に $\pi_1 < 0, \lambda < 0$ という仮定が妥当なら欠落変数バイアス（OVB）は正になる。

- 2.2節の結果より、Matchmakerを含んだ後の私立大学進学変数が結果変数に与える効果は 0 に近く、極めて小さかった。

- その小さい効果でさえ、正の欠落変数バイアスを含む結果なら、**私立大学の進学が年収を高める因果効果は「ない」**と結論づけることができる。

感度分析

- 回帰分析の推定値の確からしさを検証する方法として、感度分析がある。

- 感度分析では、制御変数の出し入れによって、介入変数の推定値が大きく動くかどうかを確認。

- 必要十分な制御変数が含まれているなら、それ以上に制御変数を追加しても、推定値はあまり変化しないはず（insensitive）。

計算確認①

- Matchmaker を使用しない推定モデルで、欠落変数はSAT の成績とする。
- １列が Shortモデルの推定結果、２列が Longモデルの推定結果。

	No selection controls			Selection controls		
	(1)	(2)	(3)	(4)	(5)	(6)
Private school	.212	.152	.139	.034	.031	.037
	(.060)	(.057)	(.043)	(.062)	(.062)	(.039)
Own SAT score ÷ 100		.051	.024		.036	.009
		(.008)	(.006)		(.006)	(.006)

計算確認①

Matchmakerを使用しないモデルにおける、私立大学進学変数と欠落変数の間の関係。

	Dependent		
	Own SAT score ÷ 100		
	(1)	(2)	(3)
Private school	1.165	1.130	.066
	(.196)	(.188)	(.112)
Female		−.367	
		(.076)	
Black		−1.947	
		(.079)	
Hispanic		−1.185	
		(.168)	
Asian		−.014	
		(.116)	
Other/missing race		−.521	
		(.293)	
High school top 10%		.948	
		(.107)	
High school rank missing		.556	
		(.102)	
Athlete		−.318	
		(.147)	
Average SAT score of schools applied to ÷ 100			.777
			(.058)
Sent two applications			.252
			(.077)
Sent three applications			.375
			(.106)
Sent four or more applications			.330
			(.093)

計算確認①

$$OVB = Short - Long = .212 - .152 = .06$$

$$OVB = \{Regression\ of\ omitted\ on\ included\}$$

$$\times \{Effect\ of\ omitted\ in\ long\}$$

$$= 1.165 \times .051 = .06.$$

- 欠落変数バイアスの大きさは、0.06。
- Longモデルの推定値は、Shortモデルの推定値から30％も低下している。

⇒ Sensitive

計算確認②

- Matchmaker を使用する推定モデルで、欠落変数は同じように SAT の成績とする。
- １列が Shortモデルの推定結果、２列が Longモデルの推定結果。

	No selection controls			Selection controls		
	(1)	(2)	(3)	(4)	(5)	(6)
Private school	.212	.152	.139	.034	.031	.037
	(.060)	(.057)	(.043)	(.062)	(.062)	(.039)
Own SAT score ÷ 100		.051	.024		.036	.009
		(.008)	(.006)		(.006)	(.006)

計算確認②

Matchmakerを使用するモデルにおける、私立大学進学変数と欠落変数の間の関係。

	Dependent		
	Own SAT score ÷ 100		
	(1)	(2)	(3)
Private school	1.165	1.130	.066
	(.196)	(.188)	(.112)
Female		−.367	
		(.076)	
Black		−1.947	
		(.079)	
Hispanic		−1.185	
		(.168)	
Asian		−.014	
		(.116)	
Other/missing race		−.521	
		(.293)	
High school top 10%		.948	
		(.107)	
High school rank missing		.556	
		(.102)	
Athlete		−.318	
		(.147)	
Average SAT score of schools applied to ÷ 100			.777
			(.058)
Sent two applications			.252
			(.077)
Sent three applications			.375
			(.106)
Sent four or more applications			.330
			(.093)

計算確認②

$$OVB = \{Regression\ of\ omitted\ on\ included\} \times \{Effect\ of\ omitted\ in\ long\} = .066 \times .036 = .0024.$$

- 欠落変数バイアスの大きさは、0.0024。
- Longモデルの推定値は、Shortモデルの推定値から7%しか変化しない。

⇒ Insensitive。**つまり Matchmaker を含めることで確からしい推定値を得られていることを示唆。**