第5章
重回帰分析

『入門　実践する計量経済学』
PPT

・重回帰モデル

・推定

・欠落変数バイアス

・コントロール

・自由度調整済み決定係数

・多重共線性

・ダミー変数を用いた実証分析

重回帰モデル

- **重回帰モデル**

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_K X_{Ki} + u_i$$

---説明変数はK個ある（非確率変数）

---誤差項 $u_i \sim i.i.d.N(0, \sigma^2)$

標準的仮定

- パラメータの解釈

β_1: (他の変数を一定とした上で) X_1だけを1単位増やしたとき、
Yは何単位変化するかを表す

[証明]

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_K X_{Ki}$$

X_{1i}は1単位増えるなら

$$Y_i' = \alpha + \beta_1 (X_{1i} + 1) + \beta_2 X_{2i} + \cdots + \beta_K X_{Ki}$$

Y_iの変化は

$$Y_i' - Y_i = \beta_1$$

推定

最小2乗法について

- 残差2乗和を最小にするようにパラメータを決める

$$\sum_{i=1}^{n} \tilde{u}_i^2 = \sum_{i=1}^{n} (Y_i - \tilde{\alpha} - \tilde{\beta}_1 X_{1i} - \cdots - \tilde{\beta}_K X_{Ki})^2$$

- $\tilde{\alpha}, \tilde{\beta}_1, \ldots, \tilde{\beta}_K$でそれぞれ偏微分して0と置くと、

$$\begin{aligned}
&1) \quad \sum (Y_i - \hat{\alpha} - \hat{\beta}_1 X_{1i} - \cdots - \hat{\beta}_K X_{Ki}) = 0 \\
&2) \quad \sum (Y_i - \hat{\alpha} - \hat{\beta}_1 X_{1i} - \cdots - \hat{\beta}_K X_{Ki}) X_{1i} = 0 \\
&\qquad \cdots \\
&K{+}1) \quad \sum (Y_i - \hat{\alpha} - \hat{\beta}_1 X_{1i} - \cdots - \hat{\beta}_K X_{Ki}) X_{Ki} = 0
\end{aligned}$$

- 残差を$\hat{u}_i = Y_i - \hat{\alpha} - \hat{\beta}_1 X_{1i} - \cdots - \hat{\beta}_K X_{Ki}$とすると、残差の性質は$K$+1個ある

$$\sum \hat{u}_i = 0, \ \sum \hat{u}_i X_{1i} = 0, \ \ldots, \sum \hat{u}_i X_{Ki} = 0$$

- 残差には、$K+1$個の制約がある

$$\sum \hat{u}_i = 0,\ \sum \hat{u}_i X_{1i} = 0,\ \ldots, \sum \hat{u}_i X_{Ki} = 0$$

$$\Rightarrow \quad \sum_{i=1}^{n} \left(\frac{\hat{u}_i}{\sigma}\right)^2 \sim \chi^2(n-K-1)$$

自由度は$n-(K+1) = n-K-1$

- 誤差項の分散σ^2の推定量

$$s^2 = \frac{1}{n-K-1} \sum_{i=1}^{n} \hat{u}_i^2$$

--- 期待値は $E[s^2] = E\left[\frac{\sigma^2}{n-K-1} \sum_{i=1}^{n} \left(\frac{\hat{u}_i}{\sigma}\right)^2\right]$

$$= \frac{\sigma^2}{n-K-1} E\left[\sum_{i=1}^{n} \left(\frac{\hat{u}_i}{\sigma}\right)^2\right] = \frac{\sigma^2}{n-K-1}(n-K-1) = \sigma^2$$

- t統計量は、自由度$n-K-1$のt分布に従う

$$t_{\hat{\beta}_1} = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} \sim t(n-K-1)$$

欠落変数バイアス

- 重回帰モデルの意義

  --- Yの動きを説明する変数は1つではない

  --- 一部の説明変数を欠落させると、関心ある変数の係数の推定に欠落変数バイアスが生じる

$$Y_i = \alpha + \beta_1 X_i + \beta_2 W_i + u_i$$

$$= \alpha + \beta_1 X_i + u_i^*$$

$$u_i^* = \beta_2 W_i + u_i$$

$$E[\hat{\beta}_1] = \beta_1 + \underbrace{\beta_2 \frac{s_{XW}}{s_X^2}}_{\text{欠落変数バイアス}}$$

---欠落変数バイアスが生じない条件

条件①: W_iはY_iの決定要因ではない($\beta_2 = 0$)

条件②: W_iはX_iと相関がない($s_{XW} = 0$)

- 欠落変数バイアスの方向

$$E\left[\hat{\beta}_1\right] = \beta_1 + \beta_2 \frac{s_{XW}}{s_X^2}$$

表 5 - 1　欠落変数バイアスの方向

	$s_{XW}>0$	$s_{XW}<0$
$\beta_2>0$	正のバイアス（$E[\hat{\beta}_1]>\beta_1$）	負のバイアス（$E[\hat{\beta}_1]<\beta_1$）
$\beta_2<0$	負のバイアス（$E[\hat{\beta}_1]<\beta_1$）	正のバイアス（$E[\hat{\beta}_1]>\beta_1$）

例：実験データー：　肥料量X_i、他要因W_i、収穫量Y_i

肥料量がランダムなら、X_iとW_iは無相関でバイアスなし

例：教育の所得への影響：教育年数X_i、能力W_i、所得Y_i

能力が上がると所得は増え$(\beta_2 > 0)$、教育年数と能力は正の相関がある$(s_{XW} > 0)$。よって、正のバイアスとなる

$$Y_i = \alpha + \beta_1 X_i + \beta_2 W_i + u_i$$
$$= \alpha + \beta_1 X_i + u_i^*$$

ここで、

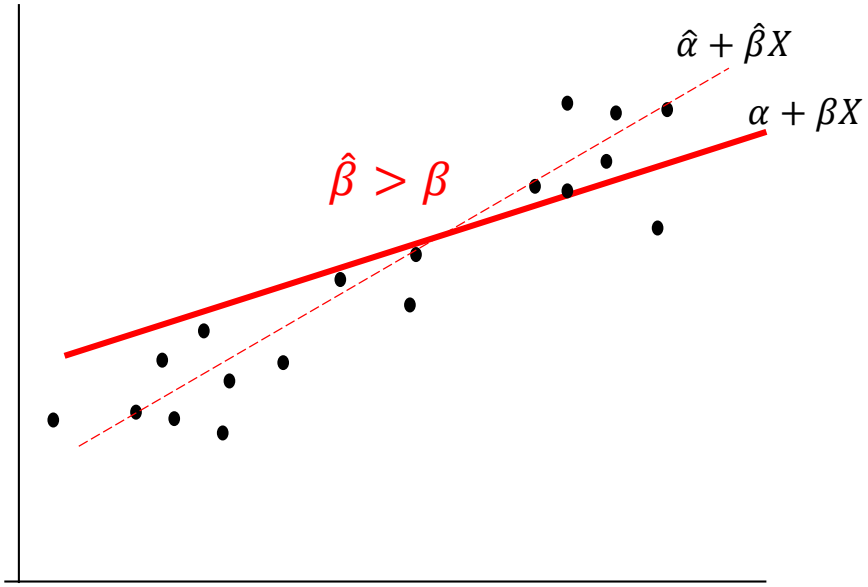
$$s_{XW} > 0 \qquad \beta_2 > 0$$

このため、

$$s_{Xu^*} > 0$$

このとき、$\hat{\beta}_1$はβ_1より大きくなる（正のバイアス）

$$u_i^* = \beta_2 W_i + u_i$$



例：賃料の決定要因

H駅周辺の賃料をY、駅からの所要時間をX、面積をWとすると、

$$\hat{Y} = 3.148 - 0.067X + 0.170W$$
$$(0.115)\ (0.009)\ \ (0.0035)$$

面積Wを除くと、Xの係数は正になる

$$\hat{Y} = 6.032 + 0.093X$$
$$(0.203)\ (0.017)$$

--- 駅からの所要時間Xと面積Wには正の相関がある($s_{XW} > 0$)

--- 面積が大きくなると賃料は高くなる($\beta_2 > 0$)

--- 欠落変数バイアスは正となる

$$E\left[\hat{\beta}_1\right] = \beta_1 + \beta_2 \frac{s_{XW}}{s_X^2}$$

例：親子の身長の関係

子供の身長をY_i、父親の身長をX_i、母親の身長をW_i

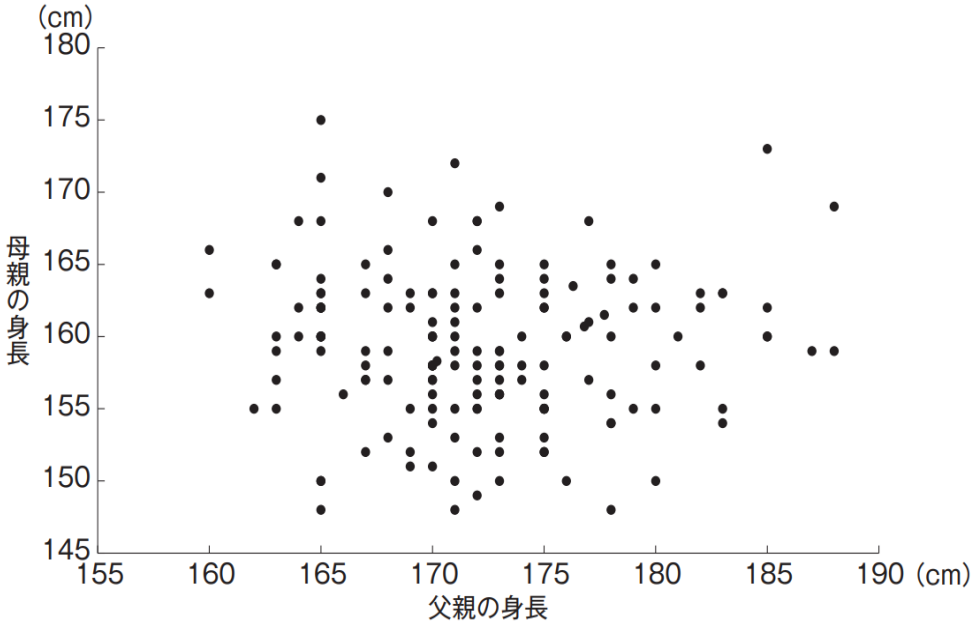
$$\hat{Y} = 58.418 + 0.331X + 0.330W$$
$$(13.88)\ \ (0.055)\ \ \ (0.057)$$

母親の身長Wを除いても、父親の身長の係数はほぼ同じ

$$\hat{Y} = 116.73 + 0.322X$$
$$(10.35)\ \ (0.06)$$

父親と母親の身長に相関はなく($s_{XW} = 0$)、欠落変数バイアスなし

図 5 - 1　父親と母親の身長の関係



コントロール

- **コントロール**：関心はX_{1i}の係数β_1にあるが、

バイアスを避けるため、他要因を説明変数

X_{2i}、…、X_{Ki}として追加する

- **コントロール変数**：コントロールのための説明変数である

$$X_{2i}、\ \ldots、X_{Ki}$$

例(169カ国における高齢化とGDP成長率の関係)

被説明変数Y_iは、1990年から2015年にかけてのGDP成長率

$$\frac{2015年のGDP_i - 1990年のGDP_i}{1990年のGDP_i}$$

説明変数X_1は、1990年から2015年にかけての高齢者割合

(50歳以上人口÷20歳～49歳人口)の変化

$$\hat{Y} = 0.420 + 0.335X_1$$
$$(0.042)\ \ (0.226)$$

コントロール変数X_2を追加

X_2: 1990年のGDPの対数

$$\hat{Y} = 1.693 + 1.036X_1 - 0.153X_2$$
$$(0.270)\ (0.258)\quad (0.032)$$

コントロール変数X_3、X_4、地域別ダミー

X_3: 1990年の人口の対数

X_4: 1990年の高齢者割合

地域別ダミー (東アジア、南アジア、アフリカなど)

$$\hat{Y} = 1.594 + 0.773X_1 - 0.155X_2 - 0.013X_3 + 0.014X_4 + \text{地域別効果}$$
$$(0.331)\ (0.293)\quad (0.040)\quad (0.018)\quad (0.332)$$

--- 高齢者割合の変化X_1の係数は有意に正

高齢化が成長率にプラスの影響を与えるのは、高齢化した国ほど

ロボットやAIの活用に熱心であるため

不必要なコントロール

$$Y_i = \alpha + \beta_1 X_i + \beta_2 W_{1i} + \beta_3 W_{2i} + u_i$$

---教育年数X_iが所得Y_iに与える影響に関心がある

---能力W_{1i}は**コントロールすべき説明変数**

- **冗長なコントロール**：　$\beta_3 = 0$としたケース
  --- W_{2i}は意味のない説明変数なので、コントロールしてもβ_3は0と推定されるだけ
  ---推定するパラメータが増えると標準誤差が大きくなる
  --- W_{2i}のコントロールは不要だが、コントロールしても問題は小さい

$$Y_i = \alpha + \beta_1 X_i + \beta_2 W_{1i} + \beta_3 W_{2i} + u_i$$

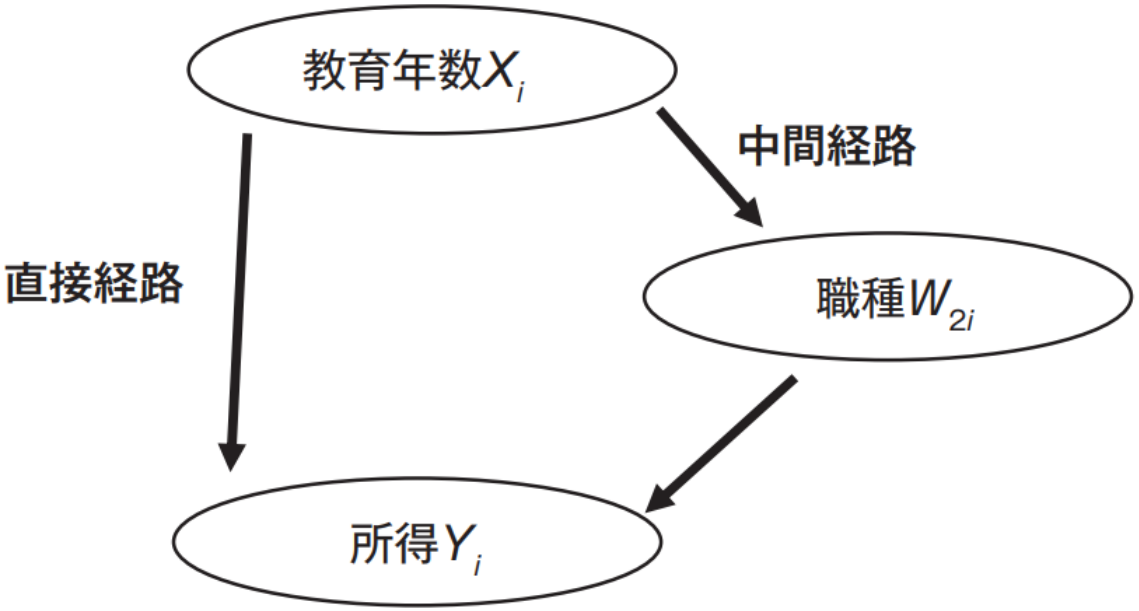
- **悪いコントロール**：　$\beta_3 \neq 0$、W_{2i}はX_iに依存して決まる

  --- W_{2i}はコントロールすべきではない

(例) W_{2i}は職種(ホワイトカラー、ブルーカラー)とすると、

職種W_{2i}は教育年数X_iに依存して決まる

図 5 - 2　教育年数が所得に与える効果



$$Y_i = \alpha + \beta_1 X_i + \beta_2 W_{1i} + \beta_3 W_{2i} + u_i$$

(例) 高校の質X_iが収入Y_iに与える影響が知りたい

W_{1i}はIQ、W_{2i}は大学に進学したか

⇒ W_{2i}は悪いコントロール

(例) 殺虫剤散布量X_iが農家の医療費Y_iに与える影響が知りたい

W_{1i}は年齢、W_{2i}は医療機関に通院した回数

⇒ W_{2i}は悪いコントロール

(例) 米国i州のビール税X_iが交通事故死亡者数Y_iに与える影響が知りたい

W_{1i}は自動車保有台数、W_{2i}はビールの消費額

⇒ W_{2i}は悪いコントロール

自由度調整済み決定係数

決定係数の問題

・決定係数は回帰モデルの当てはまりを測る指標

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \hat{u}_i^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}$$

・説明変数の数Kが増えると、残差2乗和$\sum_{i=1}^{n} \hat{u}_i^2$は低下し、$R^2$は大きくなる

---説明変数の数は多いほど、良いモデルとなってしまう

[証明] 説明変数が2個の方が、1個のときより残差2乗和が小さい

・$K = 1$のとき、最小2乗推定量は$\hat{\alpha}, \hat{\beta}_1$

・$K = 2$のとき、最小2乗推定量は$\hat{\alpha}^*, \hat{\beta}_1^*, \hat{\beta}_2^*$

・ どのような$\tilde{\alpha}, \tilde{\beta}_1, \tilde{\beta}_2$に対しても

$$\sum_{i=1}^{n} (Y_i - \hat{\alpha}^* - \hat{\beta}_1^* X_{1i} - \hat{\beta}_2^* X_{2i})^2 \leq \sum_{i=1}^{n} (Y_i - \tilde{\alpha} - \tilde{\beta}_1 X_{1i} - \tilde{\beta}_2 X_{2i})^2$$

・ これは$\tilde{\alpha} = \hat{\alpha}, \tilde{\beta}_1 = \hat{\beta}_1, \tilde{\beta}_2 = 0$でも成立する。

$$\sum_{i=1}^{n} (Y_i - \hat{\alpha}^* - \hat{\beta}_1^* X_{1i} - \hat{\beta}_2^* X_{2i})^2 \leq \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta}_1 X_{1i})^2$$

自由度調整済み決定係数

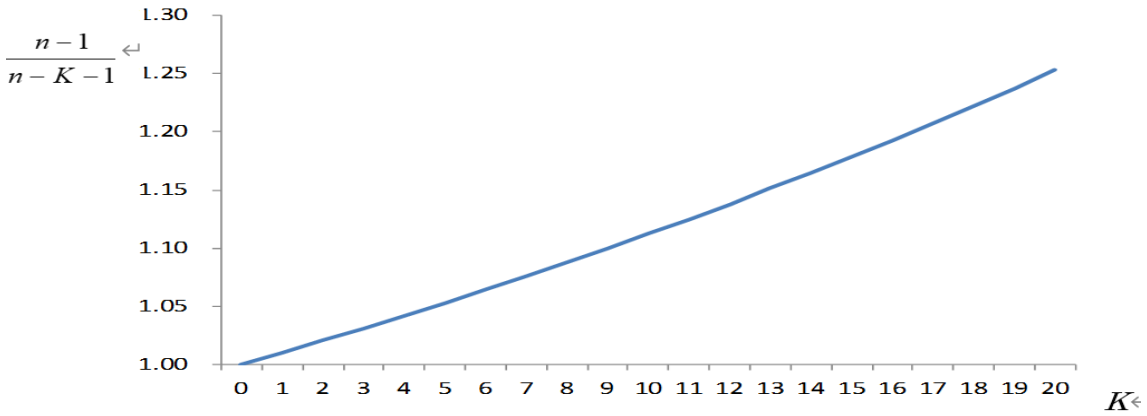
$$\bar{R}^2 = 1 - \frac{n-1}{n-K-1} \frac{\sum_{i=1}^{n} \hat{u}_i^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}$$

調整項(ペナルティ)

- **Kが増えると、$\sum_{i=1}^{n} \hat{u}_i^2$は下がるが $\frac{n-1}{n-K-1}$ は上がる**
- **説明力の高い変数を加えると$\bar{R}^2$は上がる**

--- Kが増えたとき、モデルの説明力が大きく改善する($\sum_{i=1}^{n} \hat{u}_i^2$ が大きく低下する)なら、$\bar{R}^2$は増加する

--- Kが増えても、モデルの説明力はあまり改善しない($\sum_{i=1}^{n} \hat{u}_i^2$ があまり低下しない)なら、$\bar{R}^2$は低下する



$$\bar{R}^2 = 1 - \frac{n-1}{n-K-1}\frac{\sum_{i=1}^{n}\hat{u}_i^2}{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}$$

- **$\bar{R}^2$は負になることもある**

  --- 説明力のない変数ばかりだと$\bar{R}^2$は負になる

[証明]

$$\sum_{i=1}^{n}(Y_i-\bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2 + \sum_{i=1}^{n}\hat{u}_i^2$$

であるから、説明変数に意味がなければ、

$\sum_{i=1}^{n}(Y_i-\bar{Y})^2 \approx \sum_{i=1}^{n}\hat{u}_i^2$となる。よって

$$\frac{\sum_{i=1}^{n}\hat{u}_i^2}{\sum_{i=1}^{n}(Y_i-\bar{Y})^2} \approx 1$$

また、$\frac{n-1}{n-K-1} > 1$から $\bar{R}^2$は負となる。

- **自由度調整済み決定係数はいくつあればよいか**

--- 分析対象に依存して異なる

--- 水準のままの時系列データなら0.9を超えても高いといえない

--- 変化率などの時系列データ、また横断面データなら
0.2や0.3程度でも高いかもしれない

- **被説明変数が同じであれば、自由度調整済み決定係数でモデル選択をしてもよい**

--- 決定係数は、モデルで説明された変動÷全変動だが、
被説明変数が異なれば全変動の値が異なる

(例) 被説明変数として、YとLn(Y)を用いたら相互比較はできない

--- モデルによって、サンプルサイズが異なるなら、決定係数の比較はできない

(例) 説明変数の一部が欠損している

例：賃料Yの決定要因

X_1:専有面積、X_2: 駅までの所要時間、X_3:築年数、X_4:階数

$$\hat{Y} = 2.688 + 0.160X_1$$
$$(0.101)\ (0.003)$$
$$\hat{Y} = 3.148 + 0.170X_1 - 0.067X_2$$
$$(0.115)\ (0.004)\ \ (0.009)$$
$$\hat{Y} = 4.735 + 0.165X_1 - 0.078X_2 - 0.066X_3 + 0.216X_3$$
$$(0.122)\ \ (0.003)\ \ (0.006)\ \ \ (0.003)\ \ \ (0.029)$$

表 5 - 2　賃料の決定要因

	(1)式		(2)式		(3)式	
定数項	2.688	***	3.148	***	4.735	***
	(0.101)		(0.115)		(0.122)	
専有面積	0.160	***	0.170	***	0.165	***
	(0.003)		(0.004)		(0.003)	
駅までの所要時間			-0.067	***	-0.078	***
			(0.009)		(0.006)	
築年数					-0.066	***
					(0.003)	
階数					0.216	***
					(0.029)	
$\bar{R}^2$	0.756		0.773		0.883	
n	724		724		724	

***は1%有意、**は5%有意、*は10%有意を表す。カッコ内は標準誤差

多重共線性

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + u$$

完全な多重共線性

任意の$c_0, c_1, \ldots, c_K$に対し($c_0 = c_1 = \cdots = c_K = 0$でない)、

$$c_0 + c_1 X_{1i} + c_2 X_{2i} + \cdots + c_K X_{Ki} = 0$$

--- ある説明変数が、他の説明変数の線形関数になっている

$c_1 \neq 0$ なら、$X_{1i} = -\frac{c_0}{c_1} - \frac{c_2}{c_1} X_{2i} - \cdots - \frac{c_K}{c_1} X_{Ki}$

--- パラメータを識別できない

--- 統計ソフトで分析するとErrorメッセージが出るか、問題のある変数を除いて(omitted)して推定される

例) $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + u$

- Y:成績、X_1:勉強時間、X_2:バイト時間、X_3：睡眠時間、X_4: 余暇時間
- $24 = X_1 + X_2 + X_3 + X_4$であり、

$$24 - X_1 - X_2 - X_3 - X_4 = 0$$

- 他の変数（バイト時間、睡眠時間、余暇時間）を一定とおいたうえで、勉強時間を1時間増やすことはできない

例) $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + u$

- $X_2 = 10X_1$
- 残差2乗和 $\sum_{i=1}^{n} \tilde{u}_i^2 = \sum_{i=1}^{n} (Y_i - \tilde{\alpha} - \tilde{\beta}_1 X_{1i} - \tilde{\beta}_2 X_{2i})^2$

$$= \sum_{i=1}^{n} (Y_i - \tilde{\alpha} - (\tilde{\beta}_1 + 10\tilde{\beta}_2) X_{1i})^2$$

- $\tilde{\alpha}, \tilde{\beta}_1, \tilde{\beta}_2$でそれぞれ偏微分して0と置く

① $$\frac{\partial \sum \tilde{u}_i^2}{\partial \tilde{\alpha}} = \sum \frac{\partial \tilde{u}_i^2}{\partial \tilde{u}_i} \frac{\partial \tilde{u}_i}{\partial \tilde{\alpha}}$$

$$= \sum 2\tilde{u}_i(-1) = -2 \sum (Y_i - \tilde{\alpha} - (\tilde{\beta}_1 + 10\tilde{\beta}_2) X_{1i}) = 0$$

② $$\frac{\partial}{\partial \tilde{\beta}_1} \sum \tilde{u}_i^2 = \sum \frac{\partial \tilde{u}_i^2}{\partial \tilde{u}_i} \frac{\partial \tilde{u}_i}{\partial \tilde{\beta}_1}$$

$$= \sum 2\tilde{u}_i(-X_{1i}) = -2 \sum (Y_i - \tilde{\alpha} - (\tilde{\beta}_1 + 10\tilde{\beta}_2) X_{1i}) X_{1i} = 0$$

③ $$\frac{\partial}{\partial \tilde{\beta}_2} \sum \tilde{u}_i^2 = \sum \frac{\partial \tilde{u}_i^2}{\partial \tilde{u}_i} \frac{\partial \tilde{u}_i}{\partial \tilde{\beta}_2}$$

$$= \sum 2\tilde{u}_i(-10X_{1i}) = -20 \sum (Y_i - \tilde{\alpha} - (\tilde{\beta}_1 + 10\tilde{\beta}_2) X_{1i}) X_{1i} = 0$$

---②③は同じ式であるから、2本の独立な式しか存在しない

多重共線性があるとき、独立な正規方程式の数は、パラメータの数より少ないため、パラメータを識別できない

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + u$$

弱い多重共線性

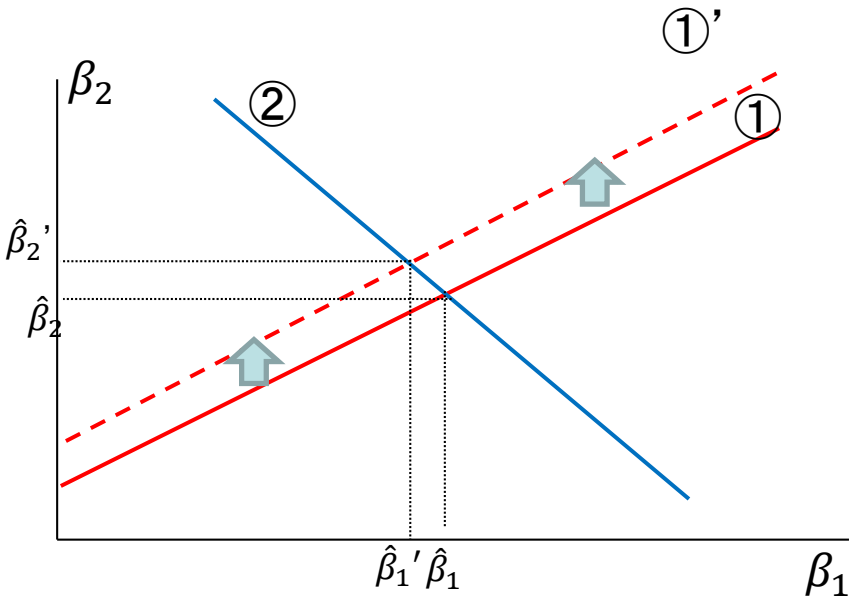
任意の定数$c_1, c_2, \ldots, c_K$に対して、

$$c_0 + c_1 X_{1i} + c_2 X_{2i} + \cdots + c_K X_{Ki} \approx 0$$

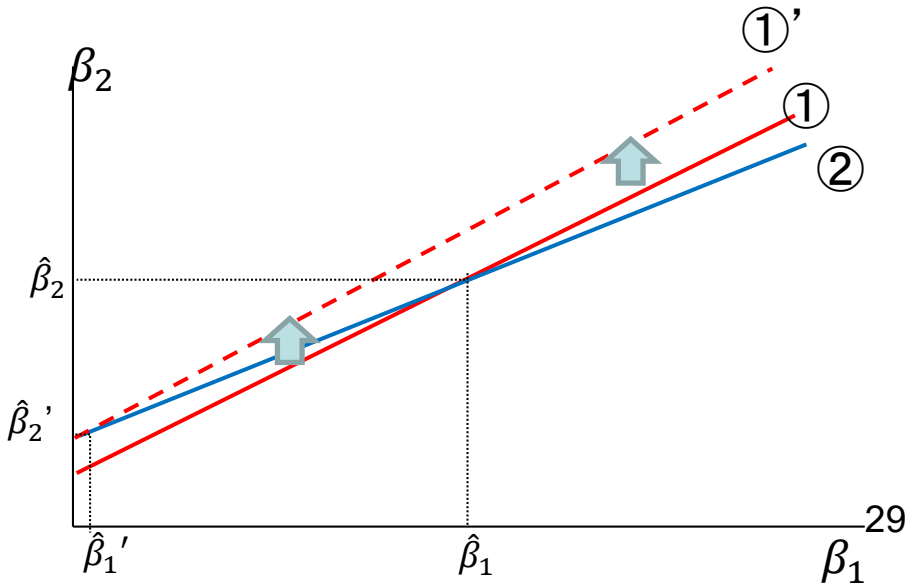
--- 全てのパラメータを推定できる

--- 推定は不安定になる（推定量の分散が大きい）

(a) 通常のケース



(b) 弱い多重共線性



弱い多重共線性は問題か？

- **理論モデルをきちんと考えたら弱い多重共線性は生じない**
  --- 生産関数なら、生産量は資本ストックと労働量に依存する
  --- 貨幣需要関数なら、貨幣量はGDPと金利に依存する

- **理論なきモデルでは、弱い多重共線性は生じやすい**
  --- 適当に説明変数をどんどん追加していくと多重共線性は生じやすい

- **理論モデルの推定で、弱い多重共線性が生じても問題ではない**
  ---もし変数が重要と考えるなら、多重共線性があっても除く必要はない。線形関数になっている変数の係数は不安定になるので、有意な結果が得られ難いだけ
  --- 多重共線性があってもサンプルサイズが大きければ安定的な推定ができる

ダミー変数を用いた実証分析

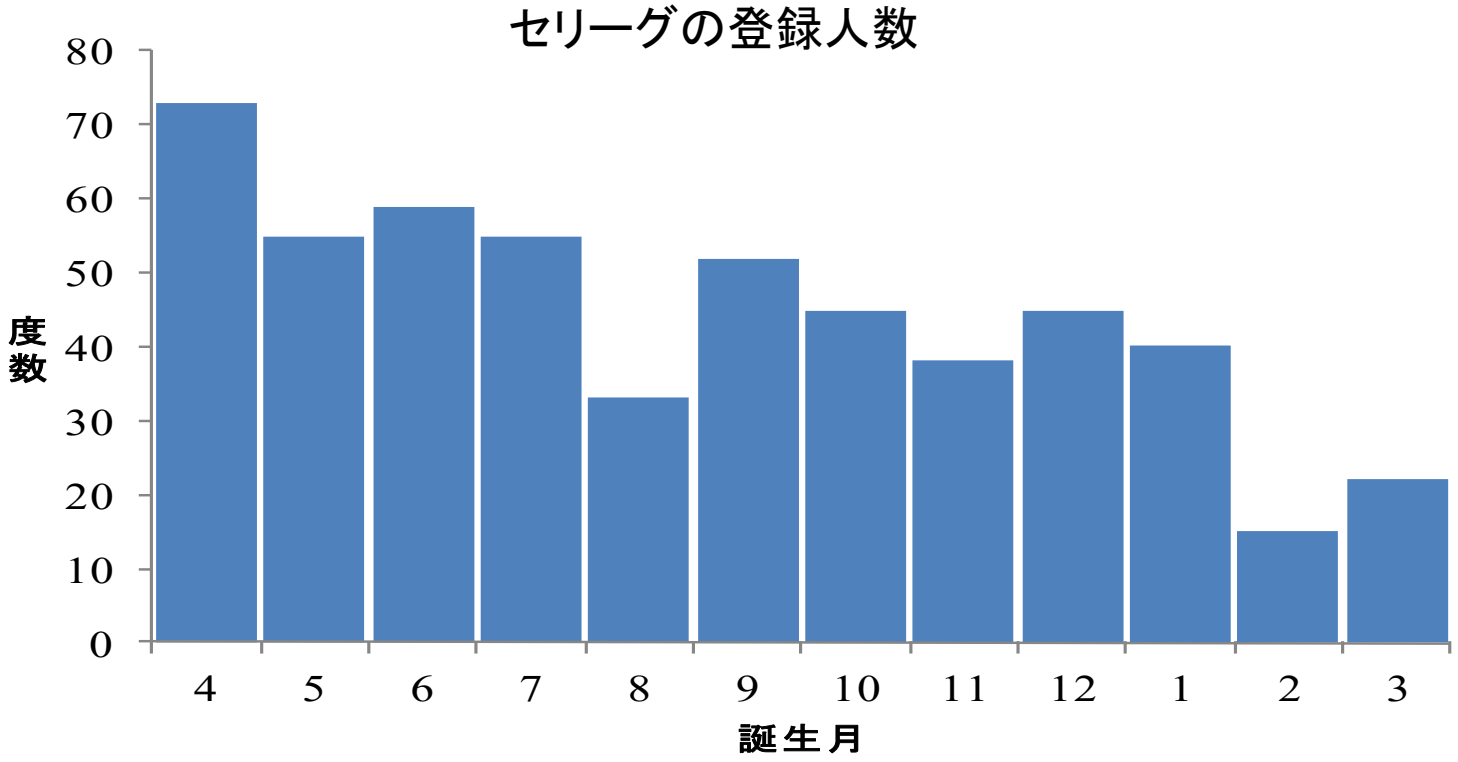
(例) 相対年齢効果：実年齢の違いが成績などに影響を与える

--- 同一学年には、4月2日から翌年の4月1日生まれの子供が入学

--- 年齢差は最大で1年ある

--- 月齢差は、子供のときに大きな影響を与える

--- 大人になっても影響は残る（3月生まれは偏差値や所得が低い）



- Y は中学生の偏差値換算した数学点数(平均50、標準偏差10)
- 生まれ月ダミー

  Q_{1i} : 4-6月生まれなら1となるダミー変数

  Q_{2i} : 7-9月生まれなら1となるダミー変数

  Q_{3i} : 10-12月生まれなら1となるダミー変数

  Q_{4i} : 1-3月生まれなら1となるダミー変数

ID	数学	Q1	Q2	Q3	Q4
1	32	0	0	0	1
2	46	0	0	1	0
3	40	0	0	1	0
4	45	0	1	0	0
5	64	0	1	0	0
6	40	0	0	1	0
7	55	0	0	0	1
8	60	0	0	0	1

- 重回帰モデル

$$Y_i = \alpha + \beta_1 Q_{2i} + \beta_2 Q_{3i} + \beta_3 Q_{4i} + u_i$$

--- 全てのダミー変数をいれると多重共線性が生じる

$$Q_{1i} + Q_{2i} + Q_{3i} + Q_{4i} = 1$$

--- 4-6月生まれなら、$Q_{2i} = Q_{3i} = Q_{4i} = 0$となる

αは4-6月生まれの平均点

--- 7-9月生まれなら、$Q_{2i} = 1$、$Q_{3i} = Q_{4i} = 0$であり

$\alpha + \beta_1$は7-9月生まれの平均点

β_1は、7-9月と4-6月生まれの平均点の差($\beta_1 = \alpha + \beta_1 - \alpha$)

--- 1-3月生まれなら、$Q_{4i} = 1$、$Q_{2i} = Q_{3i} = 0$となる

$\alpha + \beta_3$は1-3月生まれの平均点

β_3は、1-3月と4-6月生まれの平均点の差($\beta_3 = \alpha + \beta_3 - \alpha$)

	(1)式	(2)式
定数項	50.416***	28.661***
	(0.297)	(1.426)
7～9月生まれ（Q_2）	0.115	0.101
	(0.411)	(0.501)
10～12月生まれ（Q_3）	−0.322	−0.790
	(0.421)	(0.513)
1～3月生まれ（Q_4）	−1.603***	−1.861***
	(0.430)	(0.523)
母親の教育年数		0.632***
		(0.114)
父親の教育年数		1.023***
		(0.095)
$\bar{R}^2$	0.004	0.103
n	4508	2770

（注）***，**，*は，有意水準 1 %，5 %，10%で有意になることを示します．カッコ内は標準誤差，n はサンプルサイズを表します．

まとめ

- **重回帰モデル**

  ---説明変数が2個以上あるケース

- **推定**

  ---最小2乗法による推定

- **欠落変数バイアス**

- **コントロール**

  ---重回帰モデルが本当なら、全ての説明変数を含める

- **自由度調整済み決定係数**

  ---重回帰モデルにおける当てはまりの尺度

- **多重共線性**

- **ダミー変数を用いた実証分析**