『計量経済学の第一歩』

田中 隆一（著）

練習問題の解答

発行所　株式会社有斐閣
2015 年 12 月 20 日　初版第 1 刷発行

ISBN 978-4-641-15028-7

第１章　なぜ計量経済学が必要なのか

確認問題

1–1　次の 2 つの事柄の関係は因果関係でしょうか，それとも相関関係でしょうか。あなたの考えを述べてみてください。なお，相関関係と因果関係のどちらかが正しいというわけではありませんので，自分はどう思うのかを説明してみてください。

(1) 2 つの事柄：両親の所得，子どもの学力
　関係：両親の所得が高いと，子どもの学力が高い

(2) 2 つの事柄：クラブ活動への参加，友だちの数
　関係：クラブ活動へ参加している人は，友だちの数が多い

(3) 2 つの事柄：一国内の所得格差，経済成長率
　関係：所得格差の小さな国は，経済成長率が高い

(4) 2 つの事柄：友人の喫煙率，自身の喫煙
　関係：友人の喫煙率が高い人は，喫煙しやすい

(5) 2 つの事柄：都市の貧困率，犯罪率
　関係：貧困率の高い都市の犯罪率は高い

(6) 2 つの事柄：都市の凶悪犯罪発生率，1 人当たり警察官数
　関係：1 人当たり警察官数の多い都市の犯罪発生率は高い

【解答例】

(1) **相関関係**：子どもの学力に影響を与える要因としては学習時間の長さ，親の学力，児童書籍の読書量や十分な睡眠時間などが挙げられる。親の学力と学歴，学歴と所得の間には相関関係があると考えられるので，両親の所得が高いことと子どもの学力との間にも正の相関関係は見られるかもしれないが，子どもの高い学力の原因が両親の高い所得であるとはいえない。

(2) **因果関係，相関関係**：出会いの機会が増えるという理由によってクラブ活動に参加することで友達が増えるというケースを考える場合，因果関係といえる。しかし，もともと友達の数が多い社交的な人がクラブ活動へ積極的に参加するという傾向（相関関係）もあるかもしれない。その場合，個々の社交性を制御しなければ因果関係は観察できない。

(3) **相関関係**：所得格差の小さな国は国民への教育投資が大きく，人的資本の高さが経済成長率の高さをもたらすかもしれない。しかしその場合，因果関係は人的資本の高さ

と経済成長率にあり，所得格差が高いからといって必ずしも人的資本が高まるわけではないので，所得格差の小ささと経済成長率の高さの関係は因果関係とはいえない。

(4) 因果関係，相関関係：友人の薦めが喫煙のきっかけになるならば，友人の喫煙率が高いほど薦められる頻度が高くなるために自身の喫煙につながるかもしれない。また，禁煙への取り組みが禁煙に成功した友達の薦めによって広まる場合も友人の喫煙率と自身の喫煙可能性の両方を下げる。その場合は因果関係と言えるが，喫煙所での交流などによって喫煙することが喫煙者と友達になる機会を高める傾向を生む（相関関係）ということも考えられる。

(5) 因果関係，相関関係：貧困ライン以下の生活を送る人はそうでない人に比べて逮捕されたときの損失は比較的少ないかもしれない。また，耐え難い貧困によって最低限の衣食住が保障される刑務所の暮らしを選好するかもしれない。これらの理由から貧困ライン以下の人が罪を犯す可能性はそうでない人に比べて高く，都市の貧困率が高さと犯罪率の高さは因果関係であるといえる。しかし，犯罪率の高い町からは高所得者が去る，さらに犯罪率の高い町の地価は下がるために低所得者にとって流入しやすい場所になるという関係もあるかもしれない。その場合，犯罪率の高い都市の貧困率が高まるという相関関係も考えられる。

(6) 因果関係・相関関係：犯罪発生率の高い都市では，治安を守るために警察官数を多くする必要があるかもしれない。その場合は，犯罪発生率から警察官数への因果関係が考えられる。政治や経済活動の活発な都市では，警備のための警官が多いかもしれない。また，そのような都市では警官数と関係なく犯罪の発生率も高いということであれば，1 人当たり警察官数と犯罪発生率の間には見せかけの相関が生じるかもしれない。

1-2 **1-1** で見た関係が因果関係とすると，次の目標のためにどのような対策（政策）をとることができるでしょうか。また，これらの関係が相関関係であるときに，これらの対策は効果を持つでしょうか。6 つそれぞれの関係について，あなたの考えを説明してみてください。

(1) 目標：子どもの学力を高める
(2) 目標：友だちの数を増やす
(3) 目標：一国の経済成長率を高める
(4) 目標：喫煙率を下げる
(5) 目標：都市の犯罪率を下げる

(6) 目標：凶悪犯罪発生率を下げる

【解答例】

(1) 因果関係の場合，親の所得を高める政策をとることで子どもの学力が上がる。したがって，子育て世帯の親に給付金を配ることが学力向上につながる。しかし，相関関係の場合，子育て世代の親に給付金を配っても，子どもの学力向上につながるような変化には結びつかず，子どもの学習時間を少なくするような，TV ゲームや携帯の最新機種の購入，学習とは関係の薄い食費に支出されるかもしれない。子どもの学習環境を整えたいが費用面の困難がそれを阻害していた場合，給付金が学力の向上につながるかもしれないが，親の所得を高めることが必ず子どもの学力の向上という効果を持つとはいえない。

(2) 因果関係の場合，クラブ活動への参加を義務化する政策やクラブ活動への参加を奨励する政策をとることで友達の数が増える。相関関係の場合，施策による友達の増加は観測できない（0 でないとは言い切れない）。

(3) 因果関係の場合，国内の所得再配分機能の強化や所得移転などの政策によって経済格差を縮小することで経済成長率が高くなる。しかしこれらの関係が相関関係だった場合には所得格差の是正は経済成長を高めない。

(4) 因果関係の場合，友達と禁煙外来に行くと割引される，友達と一緒に禁煙パイプを購入すると特典がもらえるなどの政策が個人の禁煙運動を個別に奨励，優遇するよりも禁煙率が高まる。しかし，相関関係の場合，友人と一緒のキャンペーンと個別キャンペーンは同じ費用対効果しか持たないと考えられる。

(5) 因果関係の場合，貧困率を下げれば都市の犯罪率は下がる。したがって生活保障のような貧困ライン以下の人々への給付金を出すことで都市の犯罪率は下がる。相関関係の場合，給付金制度を設けても犯罪率は下がらない（ある都市だけ貧困者への給付金を設け，他の都市では設けなかった場合，周辺都市の貧困者が流入し（給付金なしの所得で判定した）貧困率が高まるとともに人口の増加が犯罪率の低下をもたらすかもしれない。その場合因果関係があったように見えるので政策評価の際には注意が必要である）。

(6) 因果関係の場合でも，因果の方向は犯罪発生率から警官数という方向なので，住民 1 人当たり警察官数を減らしても，犯罪発生率を減らすことはできない。相関関係の場合も，住民 1 人当たり警察官数と凶悪犯罪発生率の関係は見せかけの相関に過ぎない

ので，犯罪発生率に影響を与えている第 3 の要因（その都市における政治や経済活動の活発度）が変わらない限りは犯罪発生率を下げることはできない。

第 2 章　データの扱い方

確認問題

2-1　偶数の列

$$x_1 = 2, \quad x_2 = 4, \quad x_3 = 6, \quad x_4 = 8, \quad x_5 = 10$$

について，$\sum_{i=1}^{5} x_i$，$\sum_{i=1}^{3} 3x_i$，$\sum_{i=2}^{4}(2x_i + 3x_i)$ をそれぞれ求めましょう。

【解答例】

$$\sum_{i=1}^{5} x_i = 2 + 4 + 6 + 8 + 10 = 30$$

$$\sum_{i=1}^{3} 3x_i = 3 \times (2 + 4 + 6) = 36$$

$$\sum_{i=2}^{4}(2x_i + 3x_i) = 2\sum_{i=2}^{4} x_i + 3\sum_{i=2}^{4} x_i = 2 \times (4 + 6 + 8) + 3 \times (4 + 6 + 8) = 90$$

2-2　表 2. 3 で見た修学年数と年収のデータを使って以下の問題に答えましょう。

(1) 修学年数の平均と年収の平均をそれぞれ求めましょう。

(2) 修学年数の分散と，年収の分散をそれぞれ求めます。本文で見た 2 種類の分散の計算方法を使って，求めてみましょう。

【解答例】

(1) 修学年収の平均：13.8，年収の平均：607

(2) 修学年数の分散（σ^2）：5.96，修学年数の分散（s^2）：6.62，
年収の分散（σ^2）：24441，年収の分散（s^2）：27156.67

2-3　修学年数と年収の数値例で，xとyの共分散を$s_{xy} = (\frac{1}{n-1})\sum(x - \bar{x})(y - \bar{y})$とし，相関係数を$s_{xy}/(s_x s_y)$として相関係数を計算してみましょう。この方法で求めた相関係数は，本文中の相関係数ρと同じになります。

【解答例】

相関係数：0.69

$\frac{s_{xy}}{s_x s_y}$として定義した相関係数が，本文中で出てきた相関係数$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$と同じになることを確認しましょう。

$$\frac{s_{xy}}{s_x s_y} = \frac{(\frac{1}{n-1})\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

$$= \frac{(\frac{1}{n-1})\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\frac{1}{n-1}}\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n-1}}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

$$= \frac{(\frac{1}{n-1})\sum(x-\bar{x})(y-\bar{y})}{(\frac{1}{n-1})\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

$$= \frac{(\frac{1}{n})\sum(x-\bar{x})(y-\bar{y})}{(\frac{1}{n})\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{(\frac{1}{n})\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

$$= \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \rho$$

このように，どちらの共分散と分散を使っても，必ず相関係数は全く同じ値になります。

実証分析問題

本書のウェブサポートページにあるデータセット「2_income.csv」は，2007 年に東京大学社会科学研究所が実施した「東大社研パネル調査」を元にして作ったデータです。このデータと統計ソフトウェア（Excel や Stata など）を使って，次の問題の解答を考えてみましょう（東大社研パネル調査および統計ソフトウェアについては，本書末に収録してある参考文献およびサポート・ウェブサイトを参照してください）。

2-A 2007 年における，所得の平均と分散（σ^2とs^2の 2 種類）を計算しましょう。

【解答例】

所得の平均：258.13

分散（σ^2）：30140.32

分散（s^2）：30147.18

2-B 2007 年調査に含まれた人々の学歴（修学年数）の平均と分散（σ^2とs^2の 2 種類）を計算しましょう。

【解答例】

学歴の平均：13.85

分散（σ^2）：3.52

分散（s^2）：3.52

2-C このデータにおける所得と学歴の共分散（σ_{xy}とs_{xy}）を計算しましょう。

【解答例】

共分散（σ_{xy}）：80.85

共分散（s_{xy}）：80.87

2-D このデータにおける所得と学歴の相関係数を計算しましょう。所得と学歴の間にはどのような関係がありますか。

【解答例】

例えば，Excel の相関係数を求める関数（CORREL）を使って相関係数を計算すると，0.24815 となります。Stata の場合は，「`correlate`」というコマンドを使って相関係数を計算すると次の結果になります。

```
. correlate income yeduc
(obs=4392)

             |   income    yeduc
-------------+------------------
      income |   1.0000
       yeduc |   0.2482   1.0000
```

これらの結果から，修学年数と所得との間には正の相関があると言えます。

第 3 章　計量経済学のための確率論

確認問題

3-1　次の事象について考えます。

(1)　「修学年数が 12 年以上」という事象と，「修学年数が 16 年未満」という事象は排反事象でしょうか。

(2)　「修学年数が 12 年以上 16 年未満」という事象と，「修学年数が 12 年未満」という事情は排反事象でしょうか。

(3)　「修学年数が 12 年以上 16 年未満」という事象と，「修学年数が 12 年未満」という事象の和事象はどうなるでしょうか。

(4)　「修学年数が 12 年以上」という事象と，「修学年数が 16 年未満」という事象の積事象はどうなるでしょうか。

【解答例】

(1)　「12 年以上，16 年未満」の人が両方の事象に含まれるので，排反事象ではありません。

(2)　排反事象です。

(3)　「修学年数が 16 年未満」となります。

(4)　「修学年数が 12 年以上 16 年未満」となります。

3-2　日本人の血液型は A 型が 39%と最も多く，次は O 型で 29%，さらに B 型の 22%，AB 型の 10%となっていることが知られています。日本人全体から 1 人を無作為に選び出し，その人の血液型が何型なのかを考えます。

(1)　選ばれた人の血液型について，起こりうる結果である事象をすべて書き，標本空間も書きましょう。これらの事象はお互いに排反でしょうか。

(2)　それぞれの事象の起こりやすさとしての確率を求めましょう。これらの確率は，確率の「公理」を満たしていますか。確認してみましょう。

【解答例】

(1)　事象：「A 型」「B 型」「AB 型」「O 型」

標本空間：(「A 型」，「B 型」，「AB 型」，「O 型」)

1 人の人が異なる血液型を持っていることはなく，必ず 4 つのうちのいずれかになるので，排反事象になっています。

(2) 無作為に選ばれた人の血液型が 4 つのいずれかになる確率はそれぞれ,

$$P(\text{「A 型」}) = 0.39, \quad P(\text{「O 型」}) = 0.29,$$
$$P(\text{「B 型」}) = 0.22, \quad P(\text{「AB 型」}) = 0.10$$

となります。確率の公理それぞれについて，それらが満たされているのかを見てみましょう。

公理 1：すべての確率は 0 以上 1 未満なので満たされています。
公理 2：4 つの確率を足すと 1 になるので，満たされています。
公理 3：例えば,「A 型または B 型」という事象は「A 型」と「B 型」という排反事象の和事象になっていて，その確率は$P(\text{「A 型または B 型」}) = P(\text{「A 型」}) + P(\text{「B 型」}) = 0.61$となっているので, 公理が満たされていることが確認できます。その他のすべての場合についても，同様に満たされていることが確認できるので，公理は満たされています。

3-3 本文中で見た例 3.1（本書 34 ページ）において，すべての和事象について確率の公理が満たされていることを確認しましょう。

【解答例】

3-2 の解答例と同様に，それぞれの公理について確認をすると，満たされていることがわかります。

3-4 平成 25 年に国立教育政策研究所が発行した「平成 25 年度全国学力・学習状況調査　報告書　クロス集計」によると，平成 25 年度全国学力・学習状況調査において「朝食を毎日食べていますか」という質問に「している」「どちらかといえば，している」「あまりしていない」「まったくしていない」と答えた児童の割合は，それぞれ 88.6%，7.6%，3.0%，0.7%でした。また，この質問の答えごとに集計した算数 A の平均正答率は，それぞれ 78.4%，70.5%，65.1%，61.2%でした。このように,「朝食を毎日食べている」という事象と,「算数 A の正答率」の間には何らかの関係がありそうです。

そこで,「朝食を毎日食べている」という事象と,「算数 A の正答率が 75%以上」となる事象についての同時確率が次のようになっているとします。

朝食を食べている	算数 A の正答率が 75% 以上	
	はい	いいえ
はい	0.65	0.25
いいえ	0.05	0.05

(1) 全生徒から 1 人を無作為に抽出して，「朝食を食べている」という質問に「はい」と答える確率はいくらですか（ヒント：周辺確率を計算しましょう）。

(2) 「朝食を食べている」という質問に「はい」と答えた条件のもとで，算数 A の正答率が 75%以上になる条件付き確率を求めましょう。

(3) 「朝食を食べている」という事象と「算数 A の正答率が 75%以上」という事象は独立といえるでしょうか。

【解答例】

(1) 朝食を食べているに「はい」と答える確率は，「朝食を食べていて，算数の正答率が 70%以上」の時の確率と，「朝食を食べていて，算数の正答率か 75%未満」の時の確率の和なので，

$$0.65 + 0.25 = 0.90$$

となります。

(2) 「朝食を食べている」という質問に「はい」と答えた条件のもとで，算数 A の正答率が 75%以上になる条件付き確率は，「朝食を食べていて，かつ正答率が 75%以上」の確率を，(1)で求めた確率で割ったものになるので，

$$\frac{0.65}{0.9} = 0.722$$

となります。

(3) 「朝食を食べている」という事象と「算数 A の平均正答率が 75%以上」という事象が独立かどうかを調べるには，「朝食を食べている」という条件のもとで「算数 A の平均正答率が 75%以上」となる条件付き確率が，「算数 A の平均正答率が 75%以上」となる（条件なし）確率と同じになるかをしらべればわかります。

$$\mathrm{P}(\text{正答率が 75\%以上}) = 0.65 + 0.05 = 0.7$$

となりますが，これは

$$\mathrm{P}(\text{「正答率が 75\%以上」} \mid \text{「朝食を食べている」}) = 0.722$$

とは異なりますので，独立でないことがわかります。

3-5 確率変数Xが次の確率関数を持つときの累積分布関数$F(x)$のグラフを書きましょう。

(1) 確率変数Xは$x=(1,2,3)$のいずれかの値しかとらず，$\mathrm{P}(X=1)=\mathrm{P}(X=2)=\mathrm{P}(X=3)=1/3$

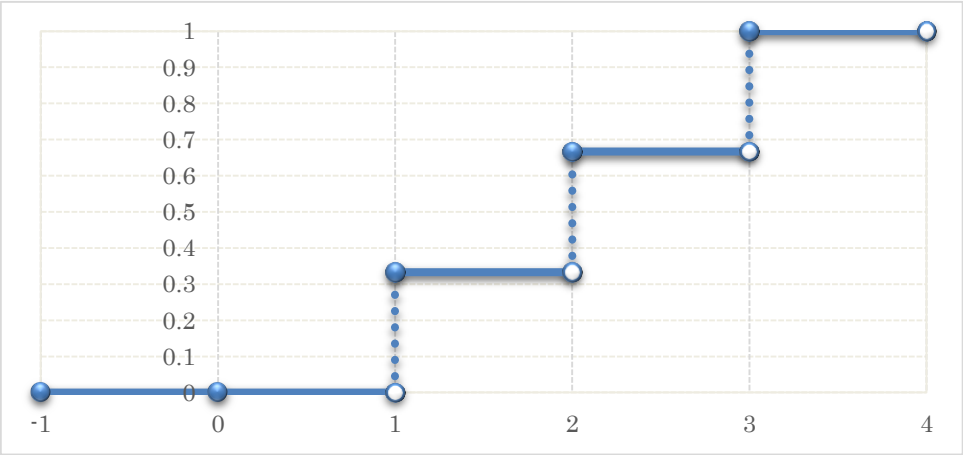
(2) 確率変数Xは$x=(-1,0,1)$のいずれかの値しかとらず，$\mathrm{P}(X=-1)=1/4$, $\mathrm{P}(X=0)=1/2$, $\mathrm{P}(X=1)=1/4$

【解答例】

(1) 累積分布関数$F(x)=P(X\leqq x)$は次のようになります。

$$F(x)=\begin{cases} 0 & x<1 \text{ のとき} \\ 1/3 & 1\leqq x<2 \text{ のとき} \\ 2/3 & 2\leqq x<3 \text{ のとき} \\ 1 & 3\leqq x \text{のとき} \end{cases}$$

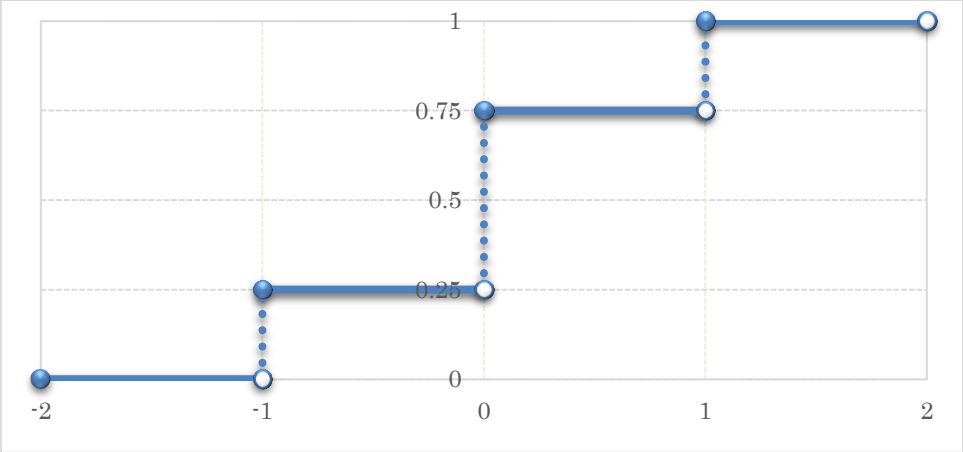
このグラフを描くと次のようになります。



(2) 累積分布関数$F(x)=P(X\leqq x)$は次のようになります。

$$F(x)=\begin{cases} 0 & x<-1 \text{ のとき} \\ 1/4 & -1 \leqq x<0 \text{ のとき} \\ 3/4 & 0 \leqq x<1 \text{ のとき} \\ 1 & 1 \leqq x \text{のとき} \end{cases}$$

このグラフを描くと次のようになります。



3-6 問題 **3-5** (2) の確率変数Xの期待値と分散を計算しましょう。

【解答例】

期待値と分散を計算するとそれぞれ次のようになります。

期待値：$-1\times\frac{1}{4}+0\times\frac{1}{2}+1\times\frac{1}{4}=0$

分散：$(-1-0)^2\times\frac{1}{4}+(0-0)^2\times\frac{1}{2}+(1-0)^2\times\frac{1}{4}=\frac{1}{2}$

3-7 表 3.1（本書 52 ページ）の学歴と年収の同時確率の例で，学歴と年収の期待値，共分散，および相関係数を自分で計算して確認しましょう。学歴と年収は独立ですか。

【解答例】

表 3-1　学歴と年収の同時確率

	年収 Y（万円）			
学歴 X	0	500	1000	周辺確率 P(x)
1	0.3	0.1	0.1	0.5
2	0.2	0.2	0.1	0.5
周辺確率 P(y)	0.5	0.3	0.2	

学歴Xと年収Yの期待値はそれぞれ

$$\mathrm{E}[X] = 1 \times 0.5 + 2 \times 0.5 = 1.5$$
$$\mathrm{E}[Y] = 0 \times 0.5 + 500 \times 0.3 + 1000 \times 0.2 = 350$$

となります。また，共分散は，

$$\begin{aligned}\mathrm{Cov}[X,Y] = {} & (1-1.5) \times (0-350) \times 0.3 + (1-1.5) \times (500-350) \times 0.1 + (1-1.5) \\ & \times (1000-350) \times 0.1 + (2-1.5) \times (0-350) \times 0.2 + (2-1.5) \\ & \times (500-350) \times 0.2 + (2-1.5) \times (1000-350) \times 0.1 = 25\end{aligned}$$

となります。さらに，学歴Xと年収Yの分散はそれぞれ，

$$\mathrm{V}[X] = (1-1.5)^2 \times 0.5 + (2-1.5)^2 \times 0.5 = 0.25$$
$$\mathrm{V}[Y] = (0-350)^2 \times 0.5 + (500-350)^2 \times 0.3 + (1000-350)^2 \times 0.2 = \ 152500$$

となるので，XとYの標準偏差はそれぞれ 0.5 と 391 になります。相関係数は共分散をこれらの標準偏差で割ったものなので，

$$\text{相関係数} = \frac{25}{0.5 \times 391} = 0.13$$

となります。

最後に，学歴と年収が独立かを調べましょう。学歴 X での条件付き期待値は

$$\mathrm{P}(Y = 0 \mid X = 1) = 0.6, \quad \mathrm{P}(Y = 500 \mid X = 1) = 0.2, \quad \mathrm{P}(Y = 1000 \mid X = 1) = 0.2$$
$$\mathrm{P}(Y = 0 \mid X = 2) = 0.4, \quad \mathrm{P}(Y = 500 \mid X = 2) = 0.4, \quad \mathrm{P}(Y = 1000 \mid X = 2) = 0.2$$

です。XとYが独立ならば, $\mathrm{P}(Y \mid X=1)=\mathrm{P}(Y \mid X=2)$が成り立っている必要がありますが, $\mathrm{P}(Y=0 \mid X=1) \neq \mathrm{P}(Y=0 \mid X=2)$なので，$X$と$Y$は独立ではありません。

3-8 Xは平均がμの確率変数のときに，次の式が正しいことを確認しましょう。

$$\mathrm{Var}[X]=\mathrm{E}[X^2]-\mu^2$$

【解答例】

分散の定義を期待値記号を使って書くと，

$$\mathrm{Var}[X]=\mathrm{E}[(X-\mu)^2]=\mathrm{E}[X^2-X\mu-\mu X+\mu^2]=\mathrm{E}[X^2]-E[X]\mu-\mu\mathrm{E}[X]+\mu^2=\mathrm{E}[X^2]-\mu^2$$

となって，この式が正しいことが確認できます。

3-9 確率変数 Xと Yの共分散$\mathrm{Cov}[X,Y]$が

$$\mathrm{Cov}[X,Y]=\mathrm{E}[(X-\mathrm{E}[X])(Y-\mathrm{E}[Y])]=\mathrm{E}[XY]-\mathrm{E}[X]\mathrm{E}[Y]$$

となることを確認しましょう。

【解答例】

共分散の定義を期待値記号を使って書くと，

$$\begin{aligned}\mathrm{Cov}[X,Y] &\equiv \mathrm{E}[(X-\mathrm{E}[X])(Y-\mathrm{E}[Y])]=\mathrm{E}\big[XY-X\mathrm{E}[Y]-\mathrm{E}[X]Y+\mathrm{E}[X]\mathrm{E}[Y]\big] \\ &=\mathrm{E}[XY]-\mathrm{E}[X\mathrm{E}[Y]]-\mathrm{E}[\mathrm{E}[X]Y]+\mathrm{E}[\mathrm{E}[X]\mathrm{E}[Y]]\end{aligned}$$

となります。ここで，$\mathrm{E}[X],\mathrm{E}[Y]$は確率変数ではないので，期待値記号の外に出すことができ，

$$\begin{aligned}&\mathrm{E}[XY]-\mathrm{E}[X\mathrm{E}[Y]]-\mathrm{E}[\mathrm{E}[X]Y]+\mathrm{E}[\mathrm{E}[X]\mathrm{E}[Y]] \\ &=\mathrm{E}[XY]-\mathrm{E}[X]\mathrm{E}[Y]-\mathrm{E}[X]\mathrm{E}[Y]+\mathrm{E}[X]\mathrm{E}[Y] \\ &=\mathrm{E}[XY]-\mathrm{E}[X]\mathrm{E}[Y]\end{aligned}$$

となって，共分散の式が成り立っていることが確認できます。

3-10 ある県の小学校で 6 年生全員に算数と国語のテストを行いました。テストは両教科とも 2 問の問題からなり，各 5 点の 10 点満点です。そのため，テストの結果は 0 点，5 点，10 点のいずれかになります。国語と算数のテストの点数をそれぞれXとYの確率変数とすると，同時確率は次のようになっているとします。

	算数		
国語	0 点	5 点	10 点
0 点	0.1	0.1	0.1
5 点	0.1	0.2	0.1
10 点	0.1	0.1	0.1

(1) 国語と算数のテストの点数の周辺確率をそれぞれ求めましょう。

(2) 国語のテストの点数が 0 点いう条件のもとでの算数のテストの点数の条件付き期待値を求めましょう。

(3) 国語のテストの点数が 5 点という条件のもとでの算数のテストの点数の条件付き期待値を求めましょう。

【解答例】

(1) 国語：$\mathrm{P}(0)=0.3,\mathrm{P}(5)=0.4,\mathrm{P}(10)=0.3$，算数：$\mathrm{P}(0)=0.3,\mathrm{P}(5)=0.4,\mathrm{P}(10)=0.3$

(2)

$$\mathrm{E}[Y|X=0]=0\times\mathrm{P}(Y=0|X=0)+5\times\mathrm{P}(Y=5|X=0)+10\times\mathrm{P}(Y=10|X=0)$$
$$=0\times\frac{1}{3}+5\times\frac{1}{3}+10\times 1/3=5$$

(3)

$$\mathrm{E}[Y|X=5]=0\times\mathrm{P}(Y=0|X=5)+5\times\mathrm{P}(Y=5|X=5)+10\times\mathrm{P}(Y=10|X=5)$$
$$=0\times\frac{1}{4}+5\times\frac{1}{2}+10\times\frac{1}{4}=5$$

実証分析問題

3-A t分布表およびF分布表はインターネットを使って検索するとたくさん見つけることができます。これらの分布表を自分でウェブから見つけ出してみましょう。

【解答例】 省略。

3-B 大卒であることと年収の関係について調べましょう。修学年数が 16 年以上の人を大卒とし，確率変数Xを使って$X = 1$，それ以外は$X = 0$とします。また，年収については，年収がゼロより大きい人々のみを対象とし，年収が 300 万円未満であれば 150，300 万円以上 600 万円未満であれば 450，600 万円以上であれば 700 の値をとる確率変数Yを考えます。

XとYの同時確率を推測するために，2007 年の東大社研パネル調査のデータを使って，それぞれの相対頻度を計算したところ，次のようになりました。

	年収（Y）		
大卒（X）	150	450	700
0	0.38	0.28	0.03
1	0.08	0.18	0.05

この相対度数表を同時確率分布として，次の問題について考えましょう。

(1) 年収（Y）の期待値と分散を求めましょう。

(2) 大卒（X）と年収（Y）の相関係数を求めましょう。

(3) 大卒である（$X = 1$）という条件のもとでの年収の期待値を求めましょう。

(4) 大卒ではない（$X = 0$）という条件のもとでの年収の期待値を求めて，(3) の答えと比較してみましょう。

【解答例】

(1) 年収の期待値を計算すると，332 万円になります。また，分散は$\mathrm{Var}[Y] = 32476$になります。

(2) 大卒ダミー変数と，年収の共分散を計算すると，25.08 になります。大卒ダミーの分散は，$\mathrm{Var}[X] = 0.2139$であり，(1) で求めた年収の平均と分散の情報を使うと，相関係数は，

$$\frac{25.08}{0.2139^{0.5} \times 32476^{0.5}} = 0.301$$

となります。

(3) 大卒である時の年収の条件付き確率は，

$$\mathrm{P}(Y = 150|X = 1) = \frac{0.08}{0.08 + 0.18 + 0.05}$$

$$\mathrm{P}(Y = 450|X = 1) = \frac{0.18}{0.08 + 0.18 + 0.05}$$

$$\mathrm{P}(Y = 700|X = 1) = \frac{0.05}{0.08 + 0.18 + 0.05}$$

この条件付き確率を使って年収の条件付き期待値を計算すると，412.9 万円になります。

(4) (3) と同様に条件付き期待値を計算すると，295.7 万円となります。この例では，大卒の方が年収の期待値が高くなっていることが確認できます。

第４章　統計学による推論

確認問題

4-1　同一の母集団分布（母平均はμ，母分散はσ^2）に独立に従っているn個の確率変数$(X_1, \ldots, X_n)$の標本平均$\bar{X}$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

が不偏性と一致性を持つことを確認しましょう。

【解答例】

母平均μの推定量としての標本平均$\bar{X}$が不偏性を持つとは，標本平均の期待値をとると，母平均となることを意味します。標本平均の期待値をとると，

$$\mathrm{E}[\bar{X}] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}[X_i] = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n}(n\mu) = \mu$$

となり，不偏性をもつことがわかります。

一致性を持つことは，標本サイズが無限大になると分散が 0 になることから確認することができます（一致性の証明には確率収束の確認が必要ですが，標本平均の一致性の直感的な確認にはこれで十分でしょう）。

参考までに，標本平均の分散が母分散を標本サイズで割ったものになることは，次のように確かめることができます。まず，$\mathrm{E}[\bar{X}] = \mu$なので，

$$\mathrm{V}[\bar{X}] = \mathrm{E}[(\bar{X} - \mathrm{E}[\bar{X}])^2] = \mathrm{E}[(\bar{X} - \mu)^2]$$

となります。ここに，$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$を代入して書き直すと，

$$\mathrm{E}[(\bar{X} - \mu)^2] = \mathrm{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right)^2\right] = \mathrm{E}\left[\frac{1}{n^2}\left(\sum_{i=1}^{n} X_i - n\mu\right)^2\right] = \frac{1}{n^2}\mathrm{E}\left[\left(\sum_{i=1}^{n}(X_i - \mu)\right)^2\right]$$

$$= \frac{1}{n^2}\mathrm{E}\left[\left(\sum_{i=1}^{n}(X_i - \mu)\right)^2\right] = \frac{1}{n^2}\mathrm{E}[((X_1 - \mu) + \cdots + (X_n - \mu))^2]$$

となります。期待値記号の中身を展開すると，

$$\frac{1}{n^2}\mathrm{E}[((X_1 - \mu) + \cdots + (X_n - \mu))^2]$$

$$= \frac{1}{n^2}\mathrm{E}[(X_1-\mu)^2 + \cdots + (X_n-\mu)^2 + 2(X_1-\mu)(X_2-\mu) + \cdots + 2(X_{n-1}-\mu)(X_n-\mu)]$$

$$= \frac{1}{n^2}(\mathrm{E}[(X_1-\mu)^2] + \cdots + \mathrm{E}[(X_n-\mu)^2] + \mathrm{E}[2(X_1-\mu)(X_2-\mu)] + \cdots + \mathrm{E}[2(X_{n-1}-\mu)(X_n-\mu)])$$

となります。最初のn個の期待値は母分散σ^2になります。また，それ以外の項は，それぞれのX_iが独立なので，共分散は 0 となります。つまり，

$$\frac{1}{n^2}(\mathrm{E}[(X_1-\mu)^2] + \cdots + \mathrm{E}[(X_n-\mu)^2] + \mathrm{E}[2(X_1-\mu)(X_2-\mu)] + \cdots + \mathrm{E}[2(X_{n-1}-\mu)(X_n-\mu)])$$

$$= \frac{1}{n^2}(\sigma^2 + \cdots + \sigma^2) = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}$$

となるので，$\mathrm{V}[\bar{X}] = \frac{\sigma^2}{n}$となることが確認できます。

4-2 同一の母集団分布（母平均はμ，母分散はσ^2）に独立に従っているn個の確率変数$(X_1, \ldots, X_n)$の標本分散S^2

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

が不偏性を持つことを確認しましょう。

【解答例】

標本分散が不偏性を持つためには，$E[S^2] = \sigma^2$となっていなければなりませんので，標本分散の両辺の期待値を取ります。

$$\mathrm{E}[S^2] = \mathrm{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2\right] = \frac{1}{n-1}\sum_{i=1}^{n}\mathrm{E}[(X_i-\bar{X})^2]$$

この式の$\mathrm{E}[(X_i-\bar{X})^2]$を次のように書き換えます。

$$\mathrm{E}[(X_i-\bar{X})^2] = \mathrm{E}\left[\left(X_i-\mu-(\bar{X}-\mu)\right)^2\right]$$

$$= \mathrm{E}[(X_i-\mu)^2 - 2(X_i-\mu)(\bar{X}-\mu) + (\bar{X}-\mu)^2]$$
$$= \mathrm{E}[(X_i-\mu)^2] - 2\mathrm{E}[(X_i-\mu)(\bar{X}-\mu)] + \mathrm{E}[(\bar{X}-\mu)^2]$$
$$= \sigma^2 - 2\mathrm{E}[(X_i-\mu)(\bar{X}-\mu)] + \mathrm{E}[(\bar{X}-\mu)^2]$$

ぞれぞれのX_iは独立なので，共分散は0となっていることに注意すると，

$$\mathrm{E}[(X_i-\mu)(\bar{X}-\mu)] = \mathrm{E}\left[(X_i-\mu)(\frac{1}{n}\sum_{i=1}^{n}X_i-\mu)\right] = \frac{1}{n}\mathrm{E}\left[(X_i-\mu)\sum_{i=1}^{n}(X_i-\mu)\right] = \frac{1}{n}\sigma^2$$

となります。また，同様に

$$\mathrm{E}[(\bar{X}-\mu)^2] = \mathrm{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}X_i-\mu\right)^2\right] = \frac{1}{n^2}\mathrm{E}\left[\left(\sum_{i=1}^{n}X_i-n\mu\right)^2\right]$$

$$= \frac{1}{n^2}\mathrm{E}\left[\sum_{i=1}^{n}\mathrm{E}[(X_i-\mu)^2]\right] = \frac{1}{n^2}(n\sigma^2) = \frac{1}{n}\sigma^2$$

となります。これらを使うと，

$$\mathrm{E}[(X_i-\bar{X})^2] = \sigma^2 - \frac{2}{n}\sigma^2 + \frac{1}{n}\sigma^2 = \frac{n-1}{n}\sigma^2$$

となりますので，

$$\mathrm{E}[S^2] = \frac{1}{n-1}\sum_{i=1}^{n}\mathrm{E}[(X_i-\bar{X})^2] = \frac{1}{n-1}\sum_{i=1}^{n}\frac{n-1}{n}\sigma^2 = \frac{1}{n-1}\frac{n-1}{n}n\sigma^2 = \sigma^2$$

となり，母分散σ^2の推定量としての標本分散S^2は不偏性を持つことが確認できます。

標本分散が一致性を持つことも同様に標本分散の分散が標本サイズを無限大に大きくした時に 0 となることから確認できますが，詳細は本書の範囲を越えるため，省略します。

4-3 本文中に出てきた夏期講習の例において，6 校分のテストスコアの変化のデータから計算された標本分散は 8.267 となること，さらにt値が 0.284 となることを自分で確かめてみましょう。

【解答例】 省略。

実証分析問題

4-A 夏期講習の効果を調べるために別のモデル校 11 校で試験的に夏期講習を実施しました。その結果，テストスコアの変化は次のようになりました。

学校	A	B	C	D	E	F	G	H	I	J	K
前	55	52	61	55	63	47	45	54	59	50	43
後	55	58	63	59	62	55	51	52	62	49	50
変化	0	6	2	4	−1	8	6	−2	3	−1	7

このとき，夏期講習の効果はあったと言えるでしょうか。有意水準 5%の両側検定と片側検定で検定してみましょう。

【解答例】

まずテストの点数の変化の平均$\bar{X}$を計算すると，およそ 2.91 点となります。次に，標本分散を計算すると，12.69 となります。

本文で見た例と同様に，それぞれのテストの点数の変化は平均がμ，分散がσ^2の正規分布に独立に従っているとすると，片側検定，両側検定のいずれにおいても帰無仮説は「$\mu = 0$」となります。また，標準化された標本平均

$$\frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}}$$

に$\mu = 0$を代入し，分散を標本分散で置き換えたものは自由度 10 の t 分布に従う，つまり，

$$\frac{\bar{X}}{\sqrt{\frac{S^2}{n}}} \sim t(10)$$

となります。この t 検定統計量にテストの点数の変化の平均である 2.91 と標本分散 12.69，および標本サイズ$n = 11$を代入すると，t 値はおよそ 2.708 となります。

片側検定の場合から考えましょう。片側検定の対立仮説を「夏期講習には効果がある」つまり「$H_1: \mu > 0$」とします。自由度 10 の上側 5%の閾値はおよそ 1.81（この値は，例えばエクセルで「`=T.INV(0.95,10)`」とすると得られます）なので，t 値は 2.91 とこの閾値より大きく，有意水準 5%で帰無仮説を棄却できます。つまり，（有意水準 5%の片側検定において）夏期講習には学力向上の効果があったということになります。

両側検定の場合の対立仮説は「$H_1: \mu \neq 0$」となり，5%の有意水準の両側検定の時の閾値は自由度 10 の t 分布における上側 2.5%の値を見れば良いことになります。この閾値はおよそ 2.23 なので，t 値は 2.91 とこの閾値より大きく，有意水準 5%で帰無仮説を棄却できます。つまり，（有意水準 5%の両側検定においても）夏期講習には効果があったということになります。

4-B Excel や Stata を使って，正規分布とt分布の上側確率 2.5%を与える点を求めましょう。t分布に関しては自由度が 5, 10, 100, 1000 のときの点をそれぞれ求めて，比較してみましょう。

【解答例】

Excel では，t分布の上側 2.5%の値は「`=T.INV(0.95,`自由度`)`」で求めることができます。また，標準正規分布の上側 2.5%の値は「`=NORM.S.INV(0.975)`」で求めることができます。様々な自由度の時のt分布の上側 2.5%の閾値は次のようになります。

上側 2.5%の閾値	
自由度	t分布
5	2.570581836
10	2.228138852
100	1.983971519
1000	1.962339081
10000	1.96020124
正規分布	1.959963985

このように，自由度が十分に大きい時は，t分布の閾値と標準正規分布の閾値はあまり変わらないものになります。

Stata の場合，「invttail(自由度,0.025)」というコマンドを使えば，閾値を計算してくれます。計算結果を表示する「display」の後に，これらのt分布の上側 2.5%の閾値を計算するコマンドを実行した結果は次のようになります。

```
. display invttail(5,0.025)
2.5705818

. display invttail(10,0.025)
2.2281389

. display invttail(100,0.025)
1.9839715

. display invttail(1000,0.025)
1.9623391
```

なお，標準正規分布の上側 2.5%の閾値は，「invnorm(0.975)」とすること求めることができます。出力結果は次の通りです。

```
. display invnorm(0.975)
1.959964
```

第 5 章　単回帰分析

確認問題

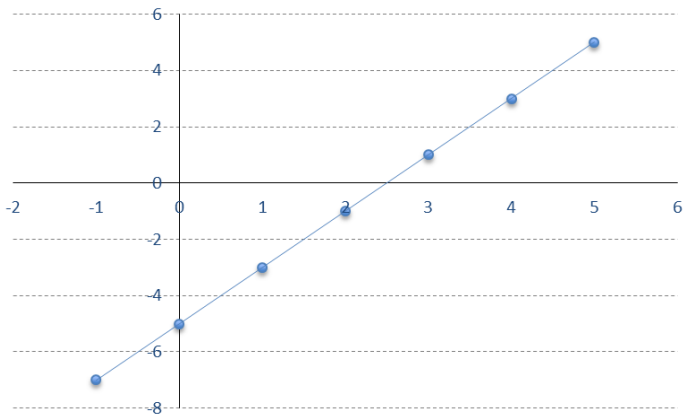
5-1　1 次関数$y = \beta_0 + \beta_1 x$について，

(1)　$\beta_0 = -5, \beta_1 = 2$のときのグラフを描きましょう。

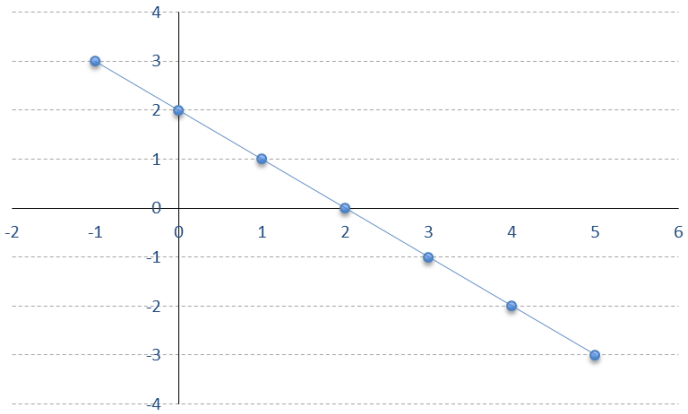
(2)　$\beta_0 = 2, \beta_1 = -1$のときのグラフを描きましょう。

【解答例】

(1)



(2)



5-2 単回帰モデル$Y_i = \beta_0 + \beta_1 X_i + U_i$を，次の標本サイズ 5 のデータを使って推定します。

$$x：-2, -1, 0, 1, 2$$
$$y：-4, -2, 3, 1, 2$$

(1) xとyの平均を求めましょう。

(2) xの標本分散と，xとyの標本共分散を求めましょう。

(3) (1) と (2) の結果を使って，傾きパラメターと切片パラメターの最小 2 乗法による推定値を求めましょう。

(4) (3) で求めた回帰パラメターの推定値を使って，5 つの観測値それぞれについて残差を求めましょう。

(5) (4) で求めた残差の 2 乗和を計算し，それをyの総変動（つまり標本分散に標本サイズを掛けたもの）で割ったものを 1 から引くことで，決定係数を求めましょう。

(6) (5) で計算した残差 2 乗和を（標本サイズ -2）で割ることによって，誤差項の分散の推定値を求めましょう。

(7) (6) で求めた誤差項の分散の推定値を使って，傾きパラメターの最小 2 乗推定量の分散の推定値を求めましょう。なお，この分散の推定値の平方根が「標準誤差」とよばれるものになり，第 6 章で詳しく説明します。

【解答例】

(1) $\bar{x} = 0,\ \bar{y} = 0$となります。

(2) 標本分散は：$[(-2)^2 + (-1)^2 + 0^2 + 1^2 + 2^2]/4 = 10/4$

標本共分散は：$[(-2) * (-4) + (-1) * (-2) + (0) * (3) + 1 * 1 + 2 * 2] = 15/4$となります。

(3) 傾きパラメターの推定値は，xとyの標本共分散をxの標本分散で割ったものになるので，

$$\hat{\beta}_1 = \frac{15/4}{10/4} = 1.5$$

となります。また，切片パラメターの推定値は，

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 0$$

となります。

(4) 5 つのそれぞれの観測値を代入することで，それぞれの残差を次のように求めることが

できます。

$-4-1.5\times(-2)=-1$

$-2-1.5\times(-1)=-0.5$

$3-0=3$

$1-1.5\times 1=-0.5$

$2-1.5\times 2=-1$

(5) 残差 2 乗和を計算すると，

$$1+0.25+9+0.25+1=11.5$$

となります。また，yの総変動は，

$$16+4+9+1+4=34$$

となります。決定係数は残差 2 乗和の総変動に対する比を 1 から引いたものなので，

$$\text{決定係数}：1-\frac{11.5}{34}=0.662$$

となります。

(6) 誤差項の分散の推定値は，残差 2 乗和 11.5 を自由度 $(5-2=3)$ で割ったものなので，

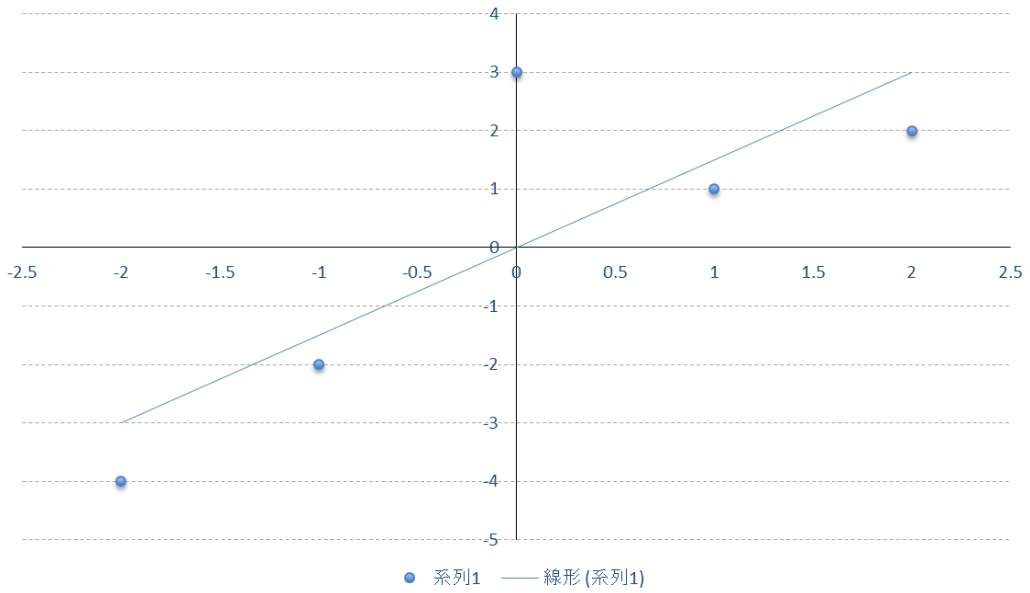
$$\hat{s}^2=\frac{11.5}{3}=3.83$$

となります。

(7) 傾きパラメターの分散の推定値は，(6) で求めた誤差項の分散の推定値を説明変数の変動$\sum(x_i-\bar{x})^2$で割ったものになるので，

$$\hat{\mathbb{V}}(\hat{\beta}_1|x_1,\ldots,x_n)=\frac{\hat{s}^2}{\sum(x_i-\bar{x})^2}=\frac{3.83}{10}=0.383$$

となります。ちなみに，標準誤差はこの平方根なので，0.619 になります。



5-3 次の表は，ある大学の 4 人の大学生の GPA（「grade point average」の略称で，学生の成績評価の指標です）と，大学入学試験での偏差値をまとめたものです。

学生	GPA	入試偏差値
1	2.7	50
2	2.4	55
3	3.3	65
4	3.5	70

(1) GPA と入試偏差値の関係を調べます。GPA を入試偏差値に回帰する単回帰モデル$\text{GPA} = \beta_0 + \beta_1(\text{偏差値}) + U$，の回帰パラメター（$\beta_0, \beta_1$）を最小 2 乗法で推定しましょう。入試偏差値と GPA の関係はどのようになっていますか。入試偏差値が 5 ポイント高いと，GPA がどれだけ変化するでしょうか。

(2) 入試偏差値から予測される GPA と残差をそれぞれの学生について求め，残差の和がゼロになることを確認しましょう。

(3) 決定係数を求めましょう。入試偏差値は GPA をどのくらい説明しているでしょうか。

【解答例】

(1) 傾きパラメターの推定値は$\hat{\beta}_1 = 0.05$。切片パラメターの推定値は $\hat{\beta}_0 = -0.025$になります。入試の偏差値が高いほど，GPA（大学での成績）も高くなる傾向があり，入試偏差値が 5 ポイント高いと，GPA は 0.25 高くなる傾向があることがわかります。

(2) GPA の当てはめ値と残差を 4 人それぞれについて計算すると，以下の表のようになります。

学生	GPA	入試偏差値	当てはめ値	残差
1	2.7	50	2.475	0.225
2	2.4	55	2.725	−0.325
3	3.3	65	3.225	0.075
4	3.5	70	3.475	0.025

残差を全て足すと，0 になっていることが確認できます。

(3) 決定係数は 0.7937 になります。これは，GPA の総変動のうち，79%は入試偏差値の変動で説明がつくことを意味しています。

5-4 次の条件付き期待値に関する式が成り立つことを確認しましょう。

(1) aは定数，Xは確率変数とすると，$\mathrm{E}[aX|X] = aX$.

(2) X，Uはともに確率変数とすると，$\mathrm{E}[UX^2|X] = X^2\mathrm{E}[U|X]$.

(3) X，Y，Uはともに確率変数で，$\mathrm{E}[YU|X] = 0$とすると，$\mathrm{E}[UXY] = 0$.

【解答例】

(1) aは定数なので，期待値記号の外に出すことができ，$\mathrm{E}[aX|X] = a\mathrm{E}[X|X]$。また，確率変数$X$で条件付けた$X$の期待値は$X$なので，$\mathrm{E}[X|X] = X$となります。これらを合わせると，$\mathrm{E}[aX|X] = aX$となることがわかります。

(2) 確率変数Xで条件づけすると，Xの値によって決まるものは全て定数と同じように扱うことができるので，X^2も定数として扱うことができ，期待値記号の外に出すことができ，

$$\mathrm{E}[UX^2|X] = X^2\mathrm{E}[U|X]$$

となります。

(3) 繰り返し期待値の法則を使うと，

$$\mathrm{E}[UXY] = \mathrm{E}\big[\mathrm{E}[UXY|X]\big]$$

となります。中の条件付き期待値では，確率変数Xで条件づけしているので，Xを条件付き期待値記号の外に出すことができ，

$$\mathrm{E}\big[\mathrm{E}[UXY|X]\big] = \mathrm{E}\big[X\mathrm{E}[UY|X]\big]$$

となります。最後は，$\mathrm{E}[YU|X] = 0$を代入することで，

$$\mathrm{E}\big[X\mathrm{E}[UY|X]\big] = E[X0] = 0$$

となることがわかります。

5-5 繰り返し期待値の法則を使って，確率変数XとUが平均独立$\mathrm{E}[U|X] = \mathrm{E}[U]$であれば，$X$と$U$の共分散$\mathrm{Cov}[X,U]$は必ず 0 になることを確認しましょう。

【解答例】

XとUの共分散は，

$$\mathrm{Cov}[X,U] = \mathrm{E}[X - \mathrm{E}(X)]\big[U - \mathrm{E}[U]\big] = \mathrm{E}[XU] - 2\mathrm{E}[\mathrm{E}[X]U] + \mathrm{E}\big[\mathrm{E}[X]\mathrm{E}[U]\big]$$

です。右辺の第 2 項目に着目すると，内側のXの期待値$\mathrm{E}[X]$はすでに定数となっていますので，外側の期待値記号の外に出してやることができ，$-2\mathrm{E}[\mathrm{E}[X]U] = -2\mathrm{E}[X]\mathrm{E}[U]$となります。また同じように，第 3 項目の中の期待値$\mathrm{E}[X]$および$\mathrm{E}[U]$は 2 つともすでに定数になっていますので，外側の期待値記号の外に出して$\mathrm{E}\big[\mathrm{E}[X]\mathrm{E}[U]\big] = \mathrm{E}[X]\mathrm{E}[U]$となります。その結果，

$$\mathrm{Cov}[X,U] = \mathrm{E}\big[X - \mathrm{E}[X]\big]\big[U - \mathrm{E}[U]\big] = \mathrm{E}[XU] - \mathrm{E}[X]\mathrm{E}[U]$$

になります。

さて，ここで繰り返し期待値の法則の登場です。右辺の第 1 項は繰り返し期待値の法則を使うと次のように書くことができます。

$$\mathrm{E}[XU] = \mathrm{E}\big[\mathrm{E}[XU|X]\big]$$

右辺の内側の期待値$\mathrm{E}[XU|X]$はXで条件付けすることによって$\mathrm{E}[XU|X] = X\mathrm{E}[U|X]$と書くことができますので，$\mathrm{E}[XU] = \mathrm{E}\big[\mathrm{E}[XU|X]\big] = \mathrm{E}\big[X\mathrm{E}[U|X]\big]$となります。さらに$X$と$U$は平均独立なので，$\mathrm{E}\big[X\mathrm{E}[U|X]\big] = \mathrm{E}\big[X\mathrm{E}[U]\big] = \mathrm{E}[X]\mathrm{E}[U]$となります。このことから，$\mathrm{E}[XU] = \mathrm{E}[X]\mathrm{E}[U]$,

つまり$\mathrm{Cov}[X,U]=0$となることがわかります。

5-6 単回帰モデルの傾きパラメターの推定量が不偏性を持つ，つまり$\mathrm{E}[\hat{\beta}_1]=\beta_1$となることを確認しましょう。

【解答例】 ウェブ補論 5 を参照してください。

5-7 単回帰モデルを標本サイズnのデータを使って最小 2 乗法で推定し，残差 2 乗和を自由度$n-2$で割った誤差項の分散の推定量s^2が不偏性を持つ，つまり$\mathrm{E}[s^2]=\mathrm{s}^2$となることを確認しましょう。

【解答例】

$\mathrm{V}[U|X]=\mathrm{V}[U]=s^2$のときに，$\hat{s}^2=\frac{1}{n-2}\sum\hat{u}_i^2$の期待値が$s^2$になる，つまり，$\mathrm{E}[\hat{s}^2]=s^2$が成り立てば，不偏性があると言えます。

$$\hat{s}^2=\frac{1}{n-2}\sum_{i=1}^{n}(\hat{u}_i)^2=\frac{1}{n-2}\sum_{i=1}^{n}\left(y_i-\hat{\beta}_0-\hat{\beta}_1x_i\right)^2$$

$y_i=\beta_0+\beta_1x_i+u_i$を代入し，さらに$\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}$に注意して整理すると，

$$\begin{aligned}\hat{s}^2&=\frac{1}{n-2}\sum_{i=1}^{n}\left(\beta_0+\beta_1x_i+u_i-\hat{\beta}_0-\hat{\beta}_1x_i\right)^2\\&=\frac{1}{n-2}\sum_{i=1}^{n}\left(\beta_0+\beta_1x_i+u_i-\left(\bar{y}-\hat{\beta}_1\bar{x}\right)-\hat{\beta}_1x_i\right)^2\\&=\frac{1}{n-2}\sum_{i=1}^{n}\left(\beta_0+\beta_1x_i+u_i-\bar{y}-\hat{\beta}_1(x_i-\bar{x})\right)^2\end{aligned}$$

さらに，$\bar{y}=\beta_0+\beta_1\bar{x}+\bar{u}$を代入すると，

$$\begin{aligned}\hat{s}^2&=\frac{1}{n-2}\sum_{i=1}^{n}\left(\beta_0+\beta_1x_i+u_i-(\beta_0+\beta_1\bar{x}+\bar{u})-\hat{\beta}_1(x_i-\bar{x})\right)^2\\&=\frac{1}{n-2}\sum_{i=1}^{n}\left(\beta_0+\beta_1x_i+u_i-\beta_0-\beta_1\bar{x}-\bar{u}-\hat{\beta}_1(x_i-\bar{x})\right)^2\\&=\frac{1}{n-2}\sum_{i=1}^{n}\left(\beta_1(x_i-\bar{x})+u_i-\bar{u}-\hat{\beta}_1(x_i-\bar{x})\right)^2\\&=\frac{1}{n-2}\sum_{i=1}^{n}\left((\beta_1-\hat{\beta}_1)(x_i-\bar{x})+u_i-\bar{u}\right)^2\\&=\frac{1}{n-2}\sum_{i=1}^{n}\left[(\beta_1-\hat{\beta}_1)^2(x_i-\bar{x})^2+(u_i-\bar{u})^2+2(\beta_1-\hat{\beta}_1)(x_i-\bar{x})(u_i-\bar{u})\right]\end{aligned}$$

となります。

また，

$$\hat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})u_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

であることに注意して，上の式に代入すると，

$$\hat{s}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left[\left[\frac{\sum_{j=1}^{n}(x_j - \bar{x})u_j}{\sum_{j=1}^{n}(x_j - \bar{x})^2}\right]^2 (x_i - \bar{x})^2 + (u_i - \bar{u})^2 - 2\left[\frac{\sum_{j=1}^{n}(x_j - \bar{x})u_j}{\sum_{j=1}^{n}(x_j - \bar{x})^2}\right](x_i - \bar{x})(u_i - \bar{u})\right]$$

まずは，すべての説明変数の観測地で条件付けした$\hat{s}^2$の条件付き期待値$E[\hat{s}^2|x_1, \dots, x_n]$を求めます。そのために，それぞれの項ごとの条件付き期待値を見てみましょう。

$$\mathrm{E}\left[\left[\frac{\sum_{j=1}^{n}(x_j - \bar{x})u_j}{\sum_{j=1}^{n}(x_j - \bar{x})^2}\right]^2 (x_i - \bar{x})^2 | x_1, \dots, x_n\right] = \frac{s^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}(x_i - \bar{x})^2$$

$$\mathrm{E}[(u_i - \bar{u})^2 | x_1, \dots, x_n] = \mathrm{E}[(u_i - \frac{1}{n}\sum_{j=1}^{n} u_j)^2 | x_1, \dots, x_n] = \frac{1}{n^2}\mathrm{E}[(nu_i - \sum_{j=1}^{n} u_j)^2 | x_1, \dots, x_n]$$

$$= \frac{1}{n^2}\mathrm{E}[(\,(n-1)u_i - \sum_{j\neq i} u_j)^2 | x_1, \dots, x_n] = \frac{1}{n^2}[(n-1)^2 s^2 + (n-1)s^2]$$

$$= \frac{(n-1)s^2}{n^2}[n - 1 + 1] = \frac{(n-1)s^2}{n}$$

$$\mathrm{E}\left[\frac{\sum_{j=1}^{n}(x_j-\bar{x})u_j}{\sum_{j=1}^{n}(x_j-\bar{x})^2}(x_i-\bar{x})(u_i-\bar{u})|x_1,\dots,x_n\right]$$

$$=\frac{(x_i-\bar{x})}{\sum_{j=1}^{n}(x_j-\bar{x})^2}\mathrm{E}\left[\sum_{j=1}^{n}(x_j-\bar{x})u_j(u_i-\bar{u})\,|x_1,\dots,x_n\right]$$

$$=\frac{(x_i-\bar{x})}{\sum_{j=1}^{n}(x_j-\bar{x})^2}\frac{1}{n}\mathrm{E}\left[\sum_{j=1}^{n}(x_j-\bar{x})u_j(nu_i-\sum_{j=1}^{n}u_j)\,|x_1,\dots,x_n\right]$$

$$=\frac{(x_i-\bar{x})}{\sum_{j=1}^{n}(x_j-\bar{x})^2}\frac{1}{n}\mathrm{E}[\sum_{j}(x_j-\bar{x})u_j\,((n-1)u_i-\sum_{j\neq i}u_j)|x_1,\dots,x_n]$$

$$=\frac{(x_i-\bar{x})}{\sum_{j=1}^{n}(x_j-\bar{x})^2}\frac{1}{n}\mathrm{E}\left[(n-1)u_i^2(x_i-\bar{x})-\sum_{j\neq i}(x_j-\bar{x})u_j^2\,|x_1,\dots,x_n\right]$$

$$=\frac{(x_i-\bar{x})}{\sum_{j=1}^{n}(x_j-\bar{x})^2}\frac{1}{n}\left[(n-1)s^2(x_i-\bar{x})-\sum_{j\neq i}(x_j-\bar{x})s^2\right]$$

$$=\frac{(x_i-\bar{x})s^2}{\sum_{j=1}^{n}(x_j-\bar{x})^2}\left[(x_i-\bar{x})-\frac{1}{n}\sum_{j}(x_j-\bar{x})\right]=\frac{(x_i-\bar{x})s^2}{\sum_{j=1}^{n}(x_j-\bar{x})^2}(x_i-\bar{x})=\frac{(x_i-\bar{x})^2s^2}{\sum_{j=1}^{n}(x_j-\bar{x})^2}$$

これらの 3 つの項を$E[\hat{s}^2|x_1,\dots,x_n]$の式に代入すると，

$$\mathrm{E}[\hat{s}^2|x_1,\dots,x_n]=\frac{1}{n-2}\sum[\frac{s^2}{D}(x_i-\bar{x})^2)+\frac{(n-1)s^2}{n}-2\frac{(x_i-\bar{x})^2s^2}{D}]$$

$$=\frac{s^2}{n-2}\sum[\frac{1}{D}(x_i-\bar{x})^2+\frac{n-1}{n}-2\frac{(x_i-\bar{x})^2}{D}]=\frac{s^2}{n-2}\left[1+\sum\frac{n-1}{n}-2\right]$$

$$=\frac{s^2}{n-2}[n-2]=s^2$$

となります。

最後に，$\hat{s}^2$の期待値$\mathrm{E}[\hat{s}^2|x_1,\dots,x_n]$が$s^2$となることは，繰り返し期待値の法則を使うと簡単に確認できます。

$$\mathrm{E}[\hat{s}^2]=\mathrm{E}[\mathrm{E}[\hat{s}^2|x_1,\dots,x_n]]=\mathrm{E}[s^2]=s^2$$

これで$\hat{s}^2$は誤差分散の不偏推定量であることが確認できました。

実証分析問題

5-A 本文中で見た年収と修学年数の関係についての単回帰モデルを使った分析結果を確認しましょう。本書のウェブサポートページから「5_1_income.csv」をダウンロードして，Excel や Stata を使って年収と修学年数の関係を表す 4 つの単回帰モデルをそれぞれ推定し，結果が同じになることを確認しましょう。

【解答例】 省略。

5-B 通勤時間が長いと，睡眠時間が短くなるか調べてみましょう。本書のウェブサポートページにある「5_2_sleep.csv」には，3726 人分の通勤時間（commute，単位は分）と睡眠時間（sleep，単位は分）のデータが収録されています。このデータを使って，睡眠時間を通勤時間に回帰する単回帰モデル

$$sleep_i = \beta_0 + \beta_1 commute_i + U_i$$

の回帰パラメターを推定しましょう。通勤時間が 1 分長くなると，睡眠時間がどれだけ短くなるでしょうか。

【解答例】

単回帰モデルを Stata で推定した結果は次のようになります。

```
. regress sleep commute;

      Source |       SS       df       MS              Number of obs =    3726
-------------+------------------------------           F(  1,  3724) =  308.71
       Model |   621252.33     1   621252.33           Prob > F      =  0.0000
    Residual |  7494252.54  3724  2012.42012           R-squared     =  0.0766
-------------+------------------------------           Adj R-squared =  0.0763
       Total |  8115504.87  3725  2178.65903           Root MSE      =   44.86

------------------------------------------------------------------------------
       sleep |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     commute |  -.5530016    .031474   -17.57   0.000    -.6147095   -.4912937
```

```
       _cons |   431.7653   1.292579   334.03   0.000     429.231    434.2995
------------------------------------------------------------------------------
```

この表から変数commuteの傾きパラメターの推定値は−0.5530016となっていることがわかります。つまり，通勤時間が 1 分長くなると，睡眠時間が 0.55 分短くなる」ことがわかります。

5-C 所得や生活水準といった経済的要因は，政治的な選好に影響を与えるのでしょうか。本書のウェブサポートページにある「5_3_abe.csv」には，4276 人分の年収（income，単位は万円）と当時の首相であった安倍晋三氏への感情を 0 から 100 までの数値で表したもの（abe，大きいほど支持が強い）が収録されています。このデータを使って，単回帰モデル

$$abe_i = \beta_0 + \beta_1 income_i + U_i$$

の回帰パラメターを推定し，賃金所得と支持感情の関係について議論してみましょう（ヒント：傾きパラメターの符号はどうなるでしょうか)。

【解答例】

この回帰モデルの回帰パラメターを Stata で推定した結果は次のようになります。

```
. regress abe income;

      Source |       SS       df       MS              Number of obs =    4276
-------------+------------------------------           F(  1,  4274) =   10.85
       Model |  2245.59031     1  2245.59031            Prob > F      =  0.0010
    Residual |  884349.376  4274  206.9137              R-squared     =  0.0025
-------------+------------------------------           Adj R-squared =  0.0023
       Total |  886594.966  4275  207.390635           Root MSE      =  14.384

------------------------------------------------------------------------------
         abe |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |  -.0030593   .0009287    -3.29   0.001      -.00488   -.0012387
       _cons |   43.43716   .3277382   132.54   0.000     42.79462     44.0797
------------------------------------------------------------------------------
```

この表から，変数 income の係数は−0.0030593となっていることがわかります。つまり，年収が高いほど，支持感情が弱くなることがわかります。

第 6 章　重回帰分析の基本

確認問題

6-1　k個の説明変数と切片からなる重回帰モデルにおいて，残差 2 乗和の最小化問題の 1 階条件式を書いてみましょう。求めた 1 階条件の式は，モーメント条件の式と同じでしょうか。

【解答例】

$k+1$個の回帰パラメターそれぞれについて偏微分して 0 とする 1 回条件の式を書くと，モーメント条件の標本版と全く同じになります。

残差は$\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \cdots - \hat{\beta}_k x_{ki}$なので，この 2 乗和は，

$$\sum_{\square=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \cdots - \hat{\beta}_k x_{ki}\right)^2$$

になります。この式を$(\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k)$でそれぞれ偏微分すると，

$$\frac{\partial}{\partial \hat{\beta}_0} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \cdots - \hat{\beta}_k x_{ki}\right)^2 = \sum_{i=1}^{n} 2\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \cdots - \hat{\beta}_k x_{ki}\right)(-1)$$

$$\frac{\partial}{\partial \hat{\beta}_1} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \cdots - \hat{\beta}_k x_{ki}\right)^2 = \sum_{i=1}^{n} 2\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \cdots - \hat{\beta}_k x_{ki}\right)(-x_{1i})$$

$$\vdots$$

$$\frac{\partial}{\partial \hat{\beta}_k} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \cdots - \hat{\beta}_k x_{ki}\right)^2 = \sum_{i=1}^{n} 2\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \cdots - \hat{\beta}_k x_{ki}\right)(-x_{ki})$$

という$k+1$本の式になります。これらの式をそれぞれ$=0$として整理すると，

$$\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \cdots - \hat{\beta}_k x_{ki}\right) = 0$$

$$\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \cdots - \hat{\beta}_k x_{ki}\right)(x_{1i}) = 0$$

$$\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \cdots - \hat{\beta}_k x_{ki}\right)(x_{ki}) = 0$$

となります。これらの式の両辺を標本サイズで割れば，モーメント法における$k+1$本のモーメント条件の標本版と等しいものになっていることがわかります。

実証分析問題

6-A 本書のウェブサポートページにある「6_1_income.csv」を使って，実際にミンサー方程式を推定してみましょう。

【解答例】

Stata による回帰分析の結果は次のようになります。

```
. regress lincome yeduc exper exper2

      Source |       SS       df       MS              Number of obs =    4299
-------------+------------------------------           F(  3,  4295) =  372.81
       Model |  712.697229     3  237.565743            Prob > F      =  0.0000
    Residual |   2736.9054  4295  .637230593           R-squared     =  0.2066
-------------+------------------------------           Adj R-squared =  0.2060
       Total |  3449.60263  4298  .802606474           Root MSE      =  .79827

------------------------------------------------------------------------------
     lincome |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       yeduc |   .1175467   .0070603    16.65   0.000      .103705    .1313885
       exper |   .1961736   .0074935    26.18   0.000     .1814824    .2108649
      exper2 |  -.0063811   .0003162   -20.18   0.000     -.007001   -.0057613
       _cons |   2.485502   .1107823    22.44   0.000     2.268312    2.702693
------------------------------------------------------------------------------
```

この推定結果から，教育の収益率は 11.75467%であることがわかります。

6-B 本書のウェブサポートページにある「6_2_yeduc.csv」には修学年数（yeduc）と父親（母親）が大学卒業以上であれば 1 となるダミー変数 pacograd（mocograd）が収録されています。

(1) 母親の学歴（mocograd）を父親の学歴（pacograd）と定数項に回帰する単回帰モデルを推定し，残差を計算しましょう。

(2) 子どもの修学年数（yeduc）を (1) で作った残差と定数項に回帰する単回帰モデルを推定しましょう。残差の係数パラメターの推定値は，例 6.4（141 ページ）の重回帰モデルにおける母親の学歴変数の係数パラメターの推定値と同じになり

ましたか（このように，重回帰分析を多段階に分けて行う方法は「回帰解剖」とよばれ，ウェブ補論 4 で詳しく解説しています）。

【解答例】

(1) Stata で推定した結果と残差の作り方は以下の通りです。

```
. regress mocograd pacograd

      Source |       SS       df       MS              Number of obs =    3954
-------------+------------------------------           F(  1,  3952) =  757.73
       Model |  50.9279652     1  50.9279652           Prob > F      =  0.0000
    Residual |  265.619582  3952  .067211433           R-squared     =  0.1609
-------------+------------------------------           Adj R-squared =  0.1607
       Total |  316.547547  3953  .080077801           Root MSE      =  .25925

------------------------------------------------------------------------------
    mocograd |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    pacograd |   .2497677   .0090736    27.53   0.000     .2319783     .267557
       _cons |   .0149893   .0048976     3.06   0.002     .0053871    .0245914
------------------------------------------------------------------------------

. predict res, residual
```

回帰モデルを推定した後に，「predict 新しい変数名, residual」というコマンドで残差（residual）を作って新しい変数名をつけてデータセットに追加します。ここでは，新しい変数名として「res」という名前をつけています。

(2) Stata による推定結果は次の通りです。

```
. regress yeduc res

      Source |       SS       df       MS              Number of obs =    3954
-------------+------------------------------           F(  1,  3952) =   35.28
       Model |  65.6143573     1  65.6143573           Prob > F      =  0.0000
    Residual |  7350.46531  3952  1.85993555           R-squared     =  0.0088
-------------+------------------------------           Adj R-squared =  0.0086
       Total |  7416.07967  3953  1.87606366           Root MSE      =  1.3638
```

```
------------------------------------------------------------------------------
      yeduc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------+----------------------------------------------------------------
        res |   .4970149   .0836795     5.94   0.000     .3329559    .6610739
      _cons |   13.96131   .0216886   643.72   0.000     13.91878    14.00383
------------------------------------------------------------------------------
```

なお，重回帰モデルの推定結果は次の通りで，母親の学歴の変数が上の結果と全く同じになっていることが確認できます。

```
. regress yeduc mocograd pacograd

      Source |       SS       df       MS              Number of obs =    3954
-------------+------------------------------           F(  2,  3951) =  422.54
       Model |  1306.72283     2  653.361415            Prob > F      =  0.0000
    Residual |  6109.35684  3951  1.54628115            R-squared     =  0.1762
-------------+------------------------------           Adj R-squared =  0.1758
       Total |  7416.07967  3953  1.87606366            Root MSE      =  1.2435

------------------------------------------------------------------------------
      yeduc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------+----------------------------------------------------------------
    mocograd |   .4970149   .0762982     6.51   0.000     .3474274    .6466024
    pacograd |   1.108861   .0475107    23.34   0.000     1.015713    1.202009
      _cons |   13.59462   .0235193   578.02   0.000     13.54851    13.64073
------------------------------------------------------------------------------
```

このように，重回帰分析を残差を唯一の説明変数とする単回帰モデルを推定することで得ようとする方法は，回帰分析をバラバラに解剖するという意味で，「回帰解剖」と呼ばれることがあります（ウェブ補論 4 参照）。この回帰解剖を行うと，重回帰モデルの推定結果と全く同じ係数パラメターの推定値を得ることができます。しかし，標準誤差は異なっている点に注意してください。これは，回帰解剖における残差は推定された変数なので，推定誤差があるのですが，回帰解剖ではそれを考慮していないため，標準誤差が誤ったものとなります。正しい標準誤差は重回帰モデルを推定した結果得られたものです。

6-C 通勤時間が長いと，仕事の満足度が低くなるか調べてみましょう。本書のウェブ

サポートページにある「6_3_happy_work.csv」には，3604 人分の通勤時間（commute，単位は分）と仕事に対する満足度（happy_work，不満から満足までの 5 段階）のデータが収録されています。

(1) 仕事に対する満足度を通勤時間に回帰する単回帰モデル

$$happy\_work_i = \beta_0 + \beta_1 commute_i + U_i$$

の回帰パラメターを推定しましょう。通勤時間の係数の符号はどうなるでしょうか。また，係数パラメターは統計的に有意ですか。

(2) 年収（income）と修学年数（yeduc）を共変量として追加した重回帰モデル

$$happy\_work_i = \beta_0 + \beta_1 commute_i + \beta_2\, income_i + \beta_3 yeduc_i + U_i$$

の回帰パラメターを推定しましょう。通勤時間の係数の値はどのように変化しましたか。

(3) で推定した重回帰モデルに年収（income）と修学年数（yeduc）を追加することに統計的に意味はあるのでしょうか。複合仮説

$$H_0: \beta_2 = \beta_3 = 0$$

をF検定することで調べてみましょう。

【解答例】

(1) Stata による回帰分析の結果は次のようになります。

```
. regress happy_work commute

      Source |       SS       df       MS              Number of obs =    3604
-------------+------------------------------           F(  1,  3602) =    2.96
       Model |  3.98644047     1  3.98644047            Prob > F      =  0.0852
    Residual |  4843.00662  3602  1.34453265           R-squared     =  0.0008
-------------+------------------------------           Adj R-squared =  0.0005
       Total |  4846.99306  3603  1.34526591           Root MSE      =  1.1595

------------------------------------------------------------------------------
  happy_work |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
```

```
    commute |  -.0014337   .0008326    -1.72   0.085    -.0030661    .0001988
      _cons |    2.22551    .033748    65.94   0.000     2.159343    2.291677
------------------------------------------------------------------------------
```

符号は負なので，通勤時間が長いほど仕事の満足度は低いことになります。この係数パラメターのp値は 8.5%なので, 5%有意水準の両側検定では帰無仮説を棄却できませんが，10%有意水準の両側検定では帰無仮説を棄却します。

(2) Stata による推定結果は以下の通りです。

```
. regress happy_work commute income yeduc

      Source |       SS       df       MS              Number of obs =    3604
-------------+------------------------------           F(  3,  3600) =   13.82
       Model |  55.1758678     3  18.3919559           Prob > F      =  0.0000
    Residual |   4791.8172  3600  1.33106033           R-squared     =  0.0114
-------------+------------------------------           Adj R-squared =  0.0106
       Total |  4846.99306  3603  1.34526591           Root MSE      =  1.1537

------------------------------------------------------------------------------
  happy_work |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     commute |  -.0024914   .0008488    -2.94   0.003    -.0041557   -.0008272
      income |   .0004735   .0000876     5.41   0.000     .0003018    .0006452
       yeduc |   .0202123   .0104646     1.93   0.054    -.0003049    .0407294
       _cons |    1.83565   .1417575    12.95   0.000     1.557717    2.113584
------------------------------------------------------------------------------
```

通勤時間の係数の値は負で，大きさは 2 倍弱になりました。これは高い所得を得られる仕事であれば通勤時間が長くても就くことを選ぶので，年収を制御することによって通勤時間の負の効果がより大きく検出されたことになります。逆にいうと，2．1 で推定した通勤時間の係数パラメターの推定値は，年収を考慮しなかったことによって上方バイアスがかかっていたと言えます。

(3) Stata による推定結果は以下の通りです。

```
. test income yeduc

 ( 1)  income = 0
 ( 2)  yeduc = 0
```

```
      F(  2,  3600) =   19.23
           Prob > F =    0.0000
```

F値は 19.23 で，p 値はほぼ 0%なので，これらの変数の係数が同時に 0 であるという帰無仮説は棄却され，これらの変数を回帰式に含めることには統計的に意味があると言えます。

6-D 所得や教育水準といった要因は，政治的な選好に影響を与えるのでしょうか。本書のウェブサポートページにある「6_4_minshu.csv」には 4218 人分の年収（income, 単位は万円）と民主党に対する支持感情を 0 から 100 までの数値で表したもの（minshu，大きいほど支持が強い）が収録されています。

(1) このデータを使って，単回帰モデル

$$minshu_i = \beta_0 + \beta_1\, income_i + U_i$$

の回帰パラメターを推定し，年収と民主党への支持感情の関係について議論してみましょう。年収の傾きパラメターは統計的に有意でしょうか。

(2) (1) で推定した単回帰モデルに修学年数（yeduc）を追加した重回帰モデル

$$minshu_i = \beta_0 + \beta_1\, income_i + \beta_2\, yeduc_i + U_i$$

の回帰パラメターを推定しましょう。賃金の傾きパラメターの係数は統計的に有意ですか。

【解答例】

(1) Stata による推定結果は以下の通りです。

```
. regress minshu income

      Source |       SS       df       MS              Number of obs =    4218
-------------+------------------------------           F(  1,  4216) =    4.32
       Model |  1478.47709     1  1478.47709            Prob > F      =  0.0378
    Residual |   1443813.6  4216  342.460532            R-squared     =  0.0010
-------------+------------------------------           Adj R-squared =  0.0008
       Total |  1445292.08  4217  342.729922            Root MSE      =  18.506
```

```
------------------------------------------------------------------------------
     minshu |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------+----------------------------------------------------------------
     income |   .0024904   .0011986     2.08   0.038     .0001406    .0048403
      _cons |   43.82368   .4239929   103.36   0.000     42.99243    44.65493
------------------------------------------------------------------------------
```

年収の係数は正で，5%水準で統計的にも有意ですので，年収が高いほど民主党への支持感情が強いと言えます。

(2) Stata による推定結果は以下の通りです。

```
. regress minshu income yeduc

      Source |       SS       df       MS              Number of obs =    4218
-------------+------------------------------           F(  2,  4215) =    4.52
       Model |  3091.44429     2  1545.72215            Prob > F      =  0.0110
    Residual |  1442200.64  4215  342.159107           R-squared     =  0.0021
-------------+------------------------------           Adj R-squared =  0.0017
       Total |  1445292.08  4217  342.729922           Root MSE      =  18.498

------------------------------------------------------------------------------
     minshu |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------+----------------------------------------------------------------
     income |   .0020023    .001219     1.64   0.101    -.0003876    .0043921
      yeduc |   .3349071   .1542502     2.17   0.030     .0324955    .6373186
      _cons |   39.30279   2.124906    18.50   0.000     35.13686    43.46873
------------------------------------------------------------------------------
```

修学年数を追加すると，年収の係数は 10%水準でも有意ではなくなりました。この結果から，(1)で検出した年収の効果は主に学歴の効果を拾っていたことがわかります。

第 7 章　重回帰分析の応用

確認問題

7-1　次の重回帰モデルでXの限界効果を求めましょう。

(1) $Y = 3 + 2X + U$

(2) $Y = -5 - 3X + 2X^2 + U$

(3) $\ln Y = X - X^2 - X^3 + U$

【解答例】

(1) $\dfrac{dY}{dX} = 2$

(2) $\dfrac{dY}{dX} = -3 + 4X$

(3) $\dfrac{d\ln Y}{dX} = 1 - 2X - 3X^2$

ちなみに，$\frac{d\ln Y}{dX} = \frac{dY/Y}{dX}$となっているので，「$X$の増分が$Y$の変化分（%）に与える効果」という解釈になります。

7-2　全国学力調査をはじめとするさまざまな学力テストにおいて，所得の高い家庭の子どもの学力は高い傾向が見られます。この傾向は大都市部や中小都市部，町村部で異なるのでしょうか。

(1) すべての生徒は大都市部，中小都市部，町村部のいずれかに住んでいるとします。これら 3 グループをダミー変数で分類してみましょう。

(2) テストの点数（score）が家計所得（income）に依存するかどうかを調べるために，単回帰モデル

$$score = \beta_0 + \beta_1\, income + U$$

を考えます。所得がテストの点数に与える影響が大都市部，中小都市部，町村部で異なる場合には，どのような回帰モデルを推定すればよいでしょうか。重回帰モデルを書いてみましょう。

(3) (2)で書いた重回帰モデルを使って，3 つの居住都市規模で所得がテストの点数に与える影響が異なるかどうかを検定するためには，どうすればよいでしょうか（ヒント：どのような帰無仮説を検定すればよいでしょうか）。

【解答例】

(1) 大都市に住んでいる場合には 1,そうでなければ 0 となるダミー変数（lcity）と，中小都市に住んでいる場合には 1,そうでなければ 0 となるダミー変数（mcity）を使うと，3 つのグループは次のように表すことができます。

	大都市部	中小都市部	町村部
lcity	1	0	0
mcity	0	1	0

(2) 2 つのダミー変数 lcity と mcity を使って，次の式を推定します。

$$score = \beta_0 + \beta_1 income + \beta_2 lcity + \beta_3 mcity + \beta_4 (income \times lcity) + \beta_5 (income \times mcity) + U$$

もし町村部と大都市部（中都市部）で所得がテストスコアに与える影響が異なるのであれば，β_4（β_5）が 0 とは異なります。

(3) 次の複合帰無仮説

$$H_0: \beta_4 = \beta_5 = 0$$

をF検定し，（例えば有意水準 5%で）帰無仮説を棄却すれば所得がテストスコアに与える影響は大都市部，中小都市部，町村部で異なると言えます。

実証分析問題

7-A 本書のウェブサポートページにある「7_1_income.csv」を使って，本文 170 ページで出てきたミンサー方程式の推定結果を自分で確認してみましょう。

【解答例】

Stata で分析した結果は次のようになります。

```
. regress lincome yeduc female female_yeduc

     Source |       SS       df       MS              Number of obs =    4286
------------+------------------------------           F(  3,  4282) =  421.49
      Model |  856.086525     3  285.362175            Prob > F      =  0.0000
   Residual |  2899.05329  4282  .677032529            R-squared     =  0.2280
------------+------------------------------           Adj R-squared =  0.2274
      Total |  3755.13981  4285  .876345347            Root MSE      =  .82282

------------------------------------------------------------------------------
     lincome |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      yeduc |   .0240947   .0085334     2.82   0.005      .0073647    .0408246
     female |  -2.079202    .192386   -10.81   0.000    -2.456378   -1.702025
female_yeduc |   .0902285   .0137996     6.54   0.000     .0631742    .1172828
      _cons |   5.346895   .1209202    44.22   0.000      5.109829    5.583962
------------------------------------------------------------------------------

. test female female_yeduc

 ( 1)  female = 0
 ( 2)  female_yeduc = 0

       F(  2,  4282) =  565.45
            Prob > F =    0.0000

. regress lincome yeduc if female==0

     Source |       SS       df       MS              Number of obs =    2150
------------+------------------------------           F(  1,  2148) =    9.61
      Model |  5.39763576     1  5.39763576            Prob > F      =  0.0020
   Residual |  1205.95084  2148  .561429628            R-squared     =  0.0045
------------+------------------------------           Adj R-squared =  0.0040
      Total |  1211.34848  2149  .563680073            Root MSE      =  .74929
```

```
------------------------------------------------------------------------------
    lincome |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------+----------------------------------------------------------------
      yeduc |   .0240947   .0077708     3.10   0.002     .0088556    .0393338
      _cons |   5.346895   .1101138    48.56   0.000     5.130955    5.562836
------------------------------------------------------------------------------


. regress lincome yeduc if female==1

     Source |       SS       df       MS              Number of obs =    2136
------------+------------------------------           F(  1,  2134) =   94.83
      Model |  75.2380817     1  75.2380817            Prob > F      =  0.0000
   Residual |  1693.10245  2134  .793393837           R-squared     =  0.0425
------------+------------------------------           Adj R-squared =  0.0421
      Total |  1768.34053  2135  .828262543           Root MSE      =  .89073


------------------------------------------------------------------------------
    lincome |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------+----------------------------------------------------------------
      yeduc |   .1143232   .0117398     9.74   0.000     .0913006    .1373458
      _cons |   3.267694   .1619845    20.17   0.000      2.95003    3.585358
------------------------------------------------------------------------------
```

7-B 第 6 章の実証分析問題 **6-C** で推定した通勤時間と仕事の満足度の関係が，男女間で異なるかどうかを調べてみましょう。本書のウェブサポートページにある「7_3_happy_work.csv」には 3097 人分の通勤時間（commute，単位は分）と仕事に対する満足度（happy_work，不満から満足までの 5 段階）のデータが収録されています。

(1) 仕事に対する満足度と通勤時間のモデルにおいて，切片パラメターと通勤時間の傾きパラメターが男女間で異なる重回帰モデルを書いてみましょう（ヒント：女性であれば 1 となる女性ダミー変数（female）およびそれと通勤時間との交差項を使いましょう）。

(2) (1)の重回帰モデルを，本書のウェブサポートページにあるデータセット「7_3_happy_work.csv」を使って推定しましょう。

(3) (2)で推定した重回帰モデルの結果を使って，男女間で回帰モデルが異なるかどうかを検定しましょう（ヒント：F検定しましょう）。

(4) 男性と女性それぞれにとって，通勤時間が長くなると仕事の満足度はどうなるでしょうか。男性の標本と女性の標本を別々に使って，それぞれの回帰モデルにおける通勤時間の係数パラメターを推定し，統計的有意性についても確認しましょう。

【解答例】

(1) $happy\_work = \beta_0 + \beta_1 \text{commute} + \beta_2 income + \beta_3 \text{yeduc} + \beta_4 female$
$+\beta_5(\text{commute} \times \text{female}) + \beta_6(income \times female) + \beta_7(\text{yeduc} \times \text{female}) + U$

(2) Stata で分析した結果は次のようになります。

```
. regress happy_work commute income yeduc female female_commute female_yeduc femal
 e_income

      Source |       SS       df       MS              Number of obs =    3097
-------------+------------------------------           F(  7,  3089) =    7.83
       Model |  70.9394375     7  10.1342054            Prob > F      =  0.0000
    Residual |  3999.80128  3089  1.29485312           R-squared     =  0.0174
-------------+------------------------------           Adj R-squared =  0.0152
       Total |  4070.74072  3096  1.31483873           Root MSE      =  1.1379

------------------------------------------------------------------------------
   happy_work |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
       commute |  -.0041476   .0013111    -3.16   0.002    -.0067183    -.001577
        income |   .0008851   .0001623     5.45   0.000     .0005668    .0012034
         yeduc |   .0241805   .0147209     1.64   0.101    -.0046832    .0530442
        female |   .3662832   .3389055     1.08   0.280    -.2982198    1.030786
female_commute |  -.0008275   .0021858    -0.38   0.705    -.0051132    .0034582
  female_yeduc |   .0007677   .0254203     0.03   0.976    -.0490746      .05061
 female_income |  -.0005877   .0003028    -1.94   0.052    -.0011813    5.93e-06
         _cons |   1.603713   .2009123     7.98   0.000     1.209777    1.997648
------------------------------------------------------------------------------
```

(3) Stata の「`test`」コマンドを使ってF検定を行うと次のようになります。

```
. test female female_commute female_yeduc female_income

 ( 1)  female = 0
 ( 2)  female_commute = 0
 ( 3)  female_yeduc = 0
 ( 4)  female_income = 0

       F(  4,  3089) =    5.75
            Prob > F =    0.0001
```

この結果, 男性と女性の回帰モデルが同じであるという複合帰無仮説は, 有意水準 5%で棄却されます。つまり，男性と女性の回帰モデルは異なると統計的にいえます。

(4) まずは男性の標本のみを用いて回帰モデルを推定します。推定結果は以下のようになります。

男性の標本を使った結果

```
. regress happy_work commute income yeduc if female==0

      Source |       SS       df       MS              Number of obs =    1659
-------------+------------------------------           F(  3,  1655) =   13.59
       Model |   53.422757     3  17.8075857            Prob > F      =  0.0000
    Residual |  2168.85331  1655  1.31048539           R-squared     =  0.0240
-------------+------------------------------           Adj R-squared =  0.0223
       Total |  2222.27607  1658  1.34033539           Root MSE      =  1.1448

------------------------------------------------------------------------------
  happy_work |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     commute |  -.0041476   .0013189    -3.14   0.002    -.0067346   -.0015607
      income |   .0008851   .0001633     5.42   0.000     .0005648    .0012054
       yeduc |   .0241805   .0148095     1.63   0.103    -.0048668    .0532278
       _cons |   1.603713   .2021214     7.93   0.000     1.207272    2.000153
------------------------------------------------------------------------------
```

この推定結果から，通勤時間が伸びると，仕事の満足度は統計的に有意に下がることがわかります。

次に，女性の標本のみを使った推定結果は次の通りです。

```
. regress happy_work commute income yeduc if female==1

      Source |       SS       df       MS              Number of obs =    1438
-------------+------------------------------           F(  3,  1434) =    3.05
       Model |  11.6925064     3  3.89750214           Prob > F      =  0.0276
    Residual |  1830.94797  1434  1.27681169           R-squared     =  0.0063
-------------+------------------------------           Adj R-squared =  0.0043
       Total |  1842.64047  1437  1.28228286           Root MSE      =    1.13

------------------------------------------------------------------------------
  happy_work |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     commute |  -.0049752   .0017367    -2.86   0.004    -.0083819   -.0015684
      income |   .0002974   .0002538     1.17   0.241    -.0002004    .0007952
       yeduc |   .0249482   .0205791     1.21   0.226    -.0154202    .0653167
       _cons |   1.969996   .2710227     7.27   0.000     1.438352    2.501639
------------------------------------------------------------------------------
```

この推定結果から，女性にとっても通勤時間が伸びると仕事の満足度は統計的に有意に下がることがわかります。

7-C 第 6 章の実証分析問題 **6-D** で，年収（income，単位は万円）や修学年数（yeduc）と民主党に対する支持感情（minshu）の関係を重回帰モデル $minshu_i = \beta_0 + \beta_1\, income_i + \beta_2 yeduc_i + U_i$ を使って調べました。この関係は都市部とそれ以外では異なるのでしょうか。

(1) 切片パラメターと年収および修学年数の傾きパラメターが都市部とそれ以外で異なる重回帰モデルを書いてみましょう（ヒント：都市部であれば 1 となる都市ダミー変数（city）およびそれと年収，修学年数との交差項を使いましょう）。

(2) (1) で書いた重回帰モデルを，本書のウェブサポートページにある「7_4_minshu.csv」を使って推定しましょう。

(3) (2)で推定した重回帰モデルの結果を使って，都市部とそれ以外で年収や修学年数が民主党支持感情に与える影響の回帰モデルが異なるかどうかを検定しましょう（ヒント：F検定しましょう）。

【解答例】

(1) $minshu_i = \beta_0 + \beta_1 income_i + \beta_2 yeduc_i + \beta_3 city$
$+\beta_4(income_i \times city) + \beta_5(yeduc_i \times city) + U_i$

(2) Stata で分析した結果は次のようになります。

```
. regress minshu income yeduc city city_income city_yeduc

      Source |       SS       df       MS              Number of obs =    4218
-------------+------------------------------           F(  5,  4212) =    5.03
       Model |  8308.94998     5     1661.79            Prob > F      =  0.0001
    Residual |  1390662.72  4212  330.166837            R-squared     =  0.0059
-------------+------------------------------           Adj R-squared =  0.0048
       Total |  1398971.67  4217  331.745712            Root MSE      =   18.17

------------------------------------------------------------------------------
      minshu |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |   .0038985   .0019409     2.01   0.045     .0000934    .0077036
       yeduc |   .2197197   .2414717     0.91   0.363    -.2536922    .6931315
        city |  -5.990619   4.266497    -1.40   0.160     -14.3552    2.373965
 city_income |  -.0006672   .0024663    -0.27   0.787    -.0055025    .0041681
  city_yeduc |   .3477803   .3113798     1.12   0.264    -.2626883    .9582489
       _cons |    40.9451   3.267236    12.53   0.000      34.5396    47.35061
------------------------------------------------------------------------------
```

(3) Stata の「test」コマンドを使ってF検定を行うと次のようになります。

```
. test city city_income city_yeduc

 ( 1)  city = 0
 ( 2)  city_income = 0
 ( 3)  city_yeduc = 0

       F(  3,  4212) =    2.31
            Prob > F =    0.0740
```

このF検定の結果から，p値は 7.4%であることがわかりました。このことは，5%の有意水準では「重回帰モデルは都市部とそれ以外で同じである」という帰無仮説を棄却で

きませんが，10%水準であれば棄却できることを意味しています。

7-D 175 ページで出てきた妻の労働供給の決定要因についての例では，妻が労働市場で働くかどうかは夫の所得と 6 歳以下の子ども数に強く依存することがわかりました。それ以外にも，母親が働いていた女性は，自分が結婚した後も働き続ける傾向があるかもしれません。このことを調べるために，15 歳のときに母親が労働市場で働いていたのであれば 1，そうでなければ 0 となるダミー変数（mowork15）を追加した重回帰モデル

$$work = \beta_0 + \beta_1\, income_s + \beta_2 childu_6 + \beta_3\, mowork_{15} + U$$

を推定することを考えます。

(1) 本書のウェブサポートページにある「7_2_work.csv」には，15 歳の時に母親が労働市場で働いていたのであれば 1，そうでなければ 0 となるダミー変数（mowork15）も収録されています。このデータセットを使って，重回帰モデルを推定しましょう。15 歳のときに母親が働いていた女性は，労働市場で働き続ける傾向があるのでしょうか。

(2) 誤差項の分散不均一性に対して頑健な標準誤差を計算し，分散が均一であるという仮定のもとで計算した（通常の）標準誤差と比較してみましょう。

【解答例】

(1) Stata で分析した結果は次のようになります。

```
. regress work income_s childu6 mowork15

      Source |       SS       df       MS              Number of obs =    1053
-------------+------------------------------           F(  3,  1049) =   22.85
       Model |   15.880939     3  5.29364633           Prob > F      =  0.0000
    Residual |  243.042138  1049  .231689359           R-squared     =  0.0613
-------------+------------------------------           Adj R-squared =  0.0587
       Total |  258.923077  1052  .246124598           Root MSE      =  .48134

------------------------------------------------------------------------------
        work |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    income_s |   -.000225    .000062    -3.63   0.000    -.0003467   -.0001032
```

```
     childu6 |  -.2048748   .0297838    -6.88   0.000    -.2633174   -.1464322
    mowork15 |    .124343   .0339083     3.67   0.000     .0578073    .1908788
       _cons |   .6805748   .0442131    15.39   0.000     .5938186    .7673311
------------------------------------------------------------------------------
```

この推定結果から，mowork15 の傾きパラメターの推定値が正で，統計的にも有意であることがわかります。このことから，「15 歳の時に母親が働いていた女性は，労働市場で働き続ける傾向がある」と言うことができます。

(2) Stata で頑健な標準誤差を推定するためには，回帰分析のコマンドの最後に，「robust」と付け加えます。

```
. regress work income_s childu6 mowork15, robust

Linear regression                               Number of obs =    1053
                                                F(  3,  1049) =   23.38
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0613
                                                Root MSE      =  .48134

------------------------------------------------------------------------------
             |               Robust
        work |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    income_s |   -.000225    .000062    -3.63   0.000    -.0003467   -.0001032
     childu6 |  -.2048748    .029725    -6.89   0.000    -.2632021   -.1465475
    mowork15 |    .124343   .0343364     3.62   0.000     .0569673    .1917188
       _cons |   .6805748   .0443839    15.33   0.000     .5934835    .7676662
------------------------------------------------------------------------------
```

この例では，通常の標準誤差と頑健な標準誤差では若干の違いはあるものの，値自体はほとんど同じになっていました。

第 8 章　操作変数法

確認問題

8-1　次の単回帰モデルの政策変数は外生変数でしょうか，それとも内生変数でしょうか。

(1) $(\text{所得}) = \beta_0 + \beta_1(\text{修学年数}) + U$

(2) $(\text{修学年数}) = \beta_0 + \beta_1(\text{父親の修学年数}) + U$

(3) $(\text{仕事の満足度}) = \beta_0 + \beta_1(\text{通勤時間}) + U$

(4) $(\text{妻の就業ダミー変数}) = \beta_0 + \beta_1(\text{夫の所得}) + U$

【解答例】

すべての場合において，政策変数が回帰モデルの誤差項Uと無相関であれば政策変数は外生変数となります。政策変数が回帰モデルの誤差項Uとなんらかの相関をもっているのであれば，政策変数は内生変数となります。以下では，それぞれの場合において，（観察データを用いていると）内生変数と考えられる理由について説明します。

(1) 生まれ持っての能力や，幼少期に形成された非認知的能力などは所得と修学年数の両方に相関しているかもしれません。このような個人特性（観測されない異質性と呼ばれたりもします）がある場合には，修学年数は内生変数になってしまいます。

(2) 父親の修学年数は父親の観測されない能力の高さと相関していて，それが子どもに遺伝しているとすると，子どもの観測されない能力は父親の修学年数と相関してしまい，その結果内生変数となっているかもしれません。ちなみに，このような遺伝的な能力の世代間相関を制御するために，養子のデータを使った研究もあります。

(3) 仕事の満足度は，その仕事から得られる報酬額に依存しているかもしれません。また，報酬額が高ければ，通勤時間が少々長くても通勤することも考えられます。その場合には，誤差項に含まれている仕事から得られる報酬を通じて，通勤時間と誤差項が相関してしまう，つまり内生変数となっているかもしれません（つまり欠落変数バイアスの原因になります)。もし報酬の情報があるのであれば，共変量として制御することによってこの内生性を部分的に制御することができます。

(4) 妻の就業状態によって，夫の就業選択が影響を受けるのであれば（逆方向の因果関係），妻の就業に影響を与える誤差項Uは夫の所得と相関するかもしれません。

8-2 ミンサー方程式を使って教育の収益率を推定したいのですが，修学年数が観測できない要因を通じて誤差項と相関している可能性があるので，操作変数法を使うことを考えています。操作変数の候補として，以下の変数を考えていますが，これらは操作変数として適切でしょうか。それぞれの変数について，簡単に議論してみましょう。

(1) 母親の修学年数
(2) 名前の頭文字の 50 音での順番
(3) 現在住んでいる場所の都市規模
(4) 15 歳のときに住んでいた場所の都市規模

【解答例】

それぞれの場合について，操作変数としての適切さ，つまり

① 内生変数と（偏）相関しているか，
② 操作変数自体は外生変数（回帰モデルの誤差項と相関していない）か，

の 2 点について考えてみましょう。

(1) 母親の学歴が高いと，子どもの学歴も高くなる傾向があるので，①は満たされているように思われます。②については，問題 **8-1** の(2)で見たように，観測されない能力が母親から子どもに遺伝しているとすると，外生変数ではない可能性があります。

(2) 名前の頭文字の 50 音での順番と修学年数の間には，関係があるようには思えないので，操作変数の要件①は満たされていないのではないかと思われます。②については，名前の頭文字と所得の関係について何らかの直接的な関係があるようには思われないので，満たされていると考えられます。

(3) 操作変数の要件①については，進学の決定時に都市規模が大きいほど高校や大学が多くあり，通学費用が低くなるため，修学年数と都市規模には正の相関関係があるように思われます。操作変数の要件②については，現在の居住地は必ずしも進学決定時のものと同じとは限らず，高学歴者ほど大都市に移住する傾向があり，かつ大都市居住者の所得は高い傾向があるので，内生変数の疑いも残ると思われます。

(4) 操作変数の要件①については，進学の決定時に都市規模が大きいほど高校や大学が多くあり，通学費用が低くなるため，修学年数と都市規模には正の相関関係があるように思

われます。操作変数の要件②については，15 歳時の居住地は親が決定していると思われるため，子どもの所得の決定要因そのものに直接的な影響は与えないかもしれないので，満たされているかもしれません。ただし，大都市に住むことを選んだ親は観測されない能力が高く，それが子どもの能力に遺伝していたりすると，この限りではありません。

実証分析問題

8-A 例 8.1（197 ページ）に出てきた教育の収益率の推定結果を，本書のウェブサポートページにある「8_income.csv」を使って確認してみましょう。

【解答例】 省略。

8-B 例 8.4（207 ページ）で見た教育の収益率の操作変数法による推定では，修学年数の内生性を考慮するために父親の修学年数と兄弟姉妹数を操作変数として使いました。本書のウェブサポートページにあるデータセット「8_income.csv」には，母親の修学年数（moyeduc）も収録されています。

(1) 父親の修学年数と兄弟姉妹数に加えて，母親の修学年数も操作変数として使って 2 段階最小 2 乗法で教育の収益率を推定してみましょう。

(2) (1) の操作変数に加えて，生まれ月も操作変数として使ってみると，教育の収益率の推定値はどのように変化しますか。

【解答例】

(1) Stata を使って推定した結果は次のようになります。

```
. ivregress 2sls lincome (yeduc=payeduc moyeduc sibs) exper exper2

Instrumental variables (2SLS) regression              Number of obs =     734
                                                      Wald chi2(3)  =   69.74
                                                      Prob > chi2   =  0.0000
                                                      R-squared     =  0.2422
                                                      Root MSE      =  .30972

------------------------------------------------------------------------------
    lincome |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
```

```
-------------+----------------------------------------------------------------
       yeduc |   .0685564   .0211268     3.25   0.001     .0271487    .1099641
       exper |   .0612705   .0160012     3.83   0.000     .0299087    .0926322
      exper2 |  -.0010616   .0006044    -1.76   0.079    -.0022463     .000123
       _cons |   4.543446   .3197084    14.21   0.000     3.916829    5.170063
------------------------------------------------------------------------------
Instrumented:  yeduc
Instruments:   exper exper2 payeduc moyeduc sibs
```

推定結果から，母親の修学年数を操作変数として追加しても，父親の修学年数と兄弟姉妹数を操作変数として行って二段階最小 2 乗法で得られた結果と大きくは変わりませんでした。なお，母親の修学年数が子どもの修学年数と相関しているかを調べるには，第 1 段階目の推定結果を見ればわかります。推定に用いたコマンドの一番最後に，「**, first**」を追加すると，上で見た 2 段階目の推定結果の前に，一段階目の推定結果を表示してくれます。

```
First-stage regressions
-----------------------

                                            Number of obs   =       734
                                            F(   5,    728) =     54.15
                                            Prob > F        =    0.0000
                                            R-squared       =    0.2711
                                            Adj R-squared   =    0.2661
                                            Root MSE        =    1.7543

------------------------------------------------------------------------------
       yeduc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |   .1527395   .0869367     1.76   0.079    -.0179371    .3234162
      exper2 |  -.0107426   .0029764    -3.61   0.000     -.016586   -.0048992
     payeduc |   .1655239   .0334356     4.95   0.000     .0998821    .2311657
     moyeduc |   .0891105   .0450941     1.98   0.049     .0005806    .1776405
        sibs |  -.2345904   .0805249    -2.91   0.004    -.3926791   -.0765017
       _cons |   11.72137   .7638968    15.34   0.000     10.22166    13.22107
------------------------------------------------------------------------------
```

この結果から，母親の修学年数は子どもの修学年数と正の，5%の有意水準で統計的に

も有意な関係があることが確認できますので，操作変数の要件のうちの一つは少なくとも満たされていたことがわかります。

(2) Stata を使って第一段階と第二段階の回帰モデルを推定した結果は次のようになります。

```
. ivregress 2sls lincome (yeduc=payeduc moyeduc sibs mbirth) exper exper2, first

First-stage regressions
-----------------------

                                          Number of obs   =       734
                                          F(   6,    727) =     45.16
                                          Prob > F        =    0.0000
                                          R-squared       =    0.2715
                                          Adj R-squared   =    0.2655
                                          Root MSE        =    1.7550

------------------------------------------------------------------------------
       yeduc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |   .1493101   .0871183     1.71   0.087    -.0217234    .3203437
      exper2 |  -.0106342   .0029819    -3.57   0.000    -.0164883     -.00478
     payeduc |   .1667067   .0334943     4.98   0.000     .1009497    .2324637
     moyeduc |   .0863505   .0452968     1.91   0.057    -.0025776    .1752786
        sibs |  -.2324421   .0806182    -2.88   0.004    -.3907145   -.0741698
      mbirth |   .0124357   .0184644     0.67   0.501    -.0238142    .0486856
       _cons |   11.68161   .7664605    15.24   0.000     10.17687    13.18635
------------------------------------------------------------------------------


Instrumental variables (2SLS) regression          Number of obs =     734
                                                  Wald chi2(3)  =   70.65
                                                  Prob > chi2   =  0.0000
                                                  R-squared     =  0.2438
                                                  Root MSE      =  .30937

------------------------------------------------------------------------------
```

```
    lincome |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+----------------------------------------------------------------
      yeduc |   .0707394    .021035     3.36   0.001     .0295115    .1119673
      exper |   .0607683   .0159786     3.80   0.000     .0294508    .0920859
     exper2 |  -.0010295   .0006032    -1.71   0.088    -.0022119    .0001528
      _cons |   4.512304   .3184353    14.17   0.000     3.888182    5.136425
------------------------------------------------------------------------------
Instrumented:  yeduc
Instruments:   exper exper2 payeduc moyeduc sibs mbirth
```

まず，1 段階目の推定結果から，生まれ月の変数と修学年数の間には統計的に有意な関係がないことがわかります（p 値は 50.1%）ので，操作変数としての要件は満たしていないことがわかります。しかしながら，父母の修学年数や兄弟姉妹数といったその他の操作変数を同時に使っているため，推定された結果は，生まれ月を操作変数として用いない時のものと大きくは異なっていません。

第 9 章　パネル・データ分析

確認問題

9-1　本文中で見たゴミ処理場建設による補償問題で，平均的な宅地の地価（千円/m²）がゴミ処理場の建設前と後で次のようになったとします。

	ゴミ処理場から近い	ゴミ処理場から遠い
建設前	480	700
建設後	520	800

差の差の推定量を使って，ゴミ処理場建設による地価の下落はどれだけかを求めましょう。

【解答例】

ゴミ処理場建設前と後の平均的な宅地の地価の変化を，ゴミ処理場の近くと遠くで比較します。その変化の差を求めると，

$$(520 - 480) - (800 - 700) = 40 - 100 = -60$$

となり，ゴミ処理場建設による地価の下落分は 1 ㎡あたり 6 万円だったことがわかります。

9-2　本文 218 ページで紹介したデュフロの論文では，インドネシアで 1973 年から始まったインプレス（INPRES）小学校建設プログラムが子どもたちの修学年数にどれだけの影響を与えたのかが議論されています。この学校建設プログラムは，もともとの就学率の低い地域に重点的に学校を建設したので，学校建設が盛んに行われた地域とそうでない地域の就学率を比較すると，学校建設が盛んに行われた地域のほうが就学率は低い傾向があります。その結果，修学年数も短い傾向があります。

学校建設プログラムが始まったのは 1973 年ですので，1974 年に 2 歳から 6 歳だった子どもたちはプログラムの影響を受けますが，12 歳から 17 歳だった子どもはすでに小学校を卒業しているのでプログラムの影響は受けません。また，1973 年以前の就学率をもとにして，就学率の高い地域 A と低い地域 B の 2 つに分けると，もともと就学率の低い地域は小学校が盛んに建設されたので，地域 B はプログラムの影響を強く受けた地域となり，地域 A はプログラムの影響をあまり受けなかった地域となります。

1974 年時点の年齢と居住地域で分類した 4 つのグループの平均的な修学年数を計算

すると，次のようになっていました。

	地域 A	地域 B
1974 年に 2 歳から 6 歳	9.76	8.49
1974 年に 12 歳から 17 歳	9.40	8.02

小学校建設プログラムは，修学年数を何年引き上げたでしょうか。

【解答例】

1974 年に 2 歳から 6 歳だった人たちは，学校建設プログラムの影響を受けた世代で，1974 年に 12 歳から 17 歳だった人たちはその影響を受けていない世代です。また，地域 A は学校建設プログラムの影響をあまり受けていない地域であり，地域 B は学校建設プログラムの影響を強く受けた地域になります。地域 A に住んでいた人々を対照群，地域 B に住んでいた人々を処置群とし，1974 年に 2 歳から 6 歳だった世代は政策実施後，12 歳から 17 歳だった世代は政策実施前と考えると，差の差の推定量による政策効果の推定値は次のようになります。

$$(8.49 - 8.02) - (9.76 - 9.40) = 0.47 - 0.36 = 0.11$$

つまり，学校建設プログラムの影響で修学年数は 0.11 年（1.3 ヶ月）伸びたことがわかります。

9-3 $Y_{it} = \beta_0 + \beta_1 T_i + \beta_2\, AFTER_t + \beta_3 (T_i \times AFTER_t) + U_{it}$を用いても，$\beta_3$の問題 **9-2** で求めた最小 2 乗推定量が差の差の推定量とまったく同じになることを確認しましょう。

【解答例】

$T = 0, AFTER = 0$のとき，回帰モデルは$Y_{it} = \beta_0 + U_{it}$となるので，$\beta_0$の最小 2 乗推定値は$T = 0, AFTER = 0$のグループの$Y$の平均（$\bar{Y}_{00}$ とします）となります。

同様に，$T = 1, AFTER = 0$のとき，回帰モデルは$Y_{it} = \beta_0 + \beta_1 + U_{it}$となるので，$\beta_0 + \beta_1$の最小 2 乗推定値は$T = 1, AFTER = 0$のグループの$Y$（$\bar{Y}_{10}$ とします）の平均となります。

同様に，$T = 0, AFTER = 1$のとき，回帰モデルは$Y_{it} = \beta_0 + \beta_2 + U_{it}$となるので，$\beta_0 + \beta_2$の最小 2 乗推定値は$T = 0, AFTER = 1$のグループの Y の平均（$\bar{Y}_{01}$ とします）となります。

最後に，$T = 1, AFTER = 1$のとき，回帰モデルは$Y_{it} = \beta_0 + \beta_1 + \beta_2 + \beta_3 + U_{it}$となるので，$\beta_0 + \beta_1 + \beta_2 + \beta_3$の最小 2 乗推定値は$T = 0, AFTER = 1$のグループの Y の平均（$\bar{Y}_{11}$ とします）となります。

以上をまとめると，$\hat{\beta}_0 = \bar{Y}_{00}$，$\hat{\beta}_0 + \hat{\beta}_1 = \bar{Y}_{10}$，$\hat{\beta}_0 + \hat{\beta}_2 = \bar{Y}_{01}$，$\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3 = \bar{Y}_{11}$とな

ります。差の差の推定量は$(\bar{Y}_{11} - \bar{Y}_{10}) - (\bar{Y}_{01} - \bar{Y}_{00})$なので，これに代入すると，

$$(\bar{Y}_{11} - \bar{Y}_{10}) - (\bar{Y}_{01} - \bar{Y}_{00}) = [\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3 - (\hat{\beta}_0 + \hat{\beta}_1)] - (\hat{\beta}_0 + \hat{\beta}_2 - \hat{\beta}_0) = \hat{\beta}_3$$

になります。

実証分析問題

9-A 例 9.1（224 ページ）で見た生活満足度と喫煙本数の関係について，もう少し調べてみましょう。データセット「9_1_cig_xt.csv」には所得（income，単位は万円）も収録されています。所得を共変量として追加して，生活の満足度が喫煙本数に与える影響を推定してみましょう。

【解答例】

Stata による推定結果は次のようになります。

```
. regress d_ncig d_life d_income, noconstant

      Source |       SS       df       MS              Number of obs =    3022
-------------+------------------------------           F(  2,  3020) =   15.39
       Model |  865.884277     2  432.942139            Prob > F      =  0.0000
    Residual |  84935.4907  3020  28.1243347            R-squared     =  0.0101
-------------+------------------------------           Adj R-squared =  0.0094
       Total |   85801.375  3022  28.3922485            Root MSE      =  5.3032

------------------------------------------------------------------------------
      d_ncig |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      d_life |  -.3006544   .1094783    -2.75   0.006    -.5153139   -.0859948
    d_income |   .0031436   .0006431     4.89   0.000     .0018828    .0044045
------------------------------------------------------------------------------
```

この推定結果の表より，生活の満足度の変化の係数パラメターの推定値は負で統計的にも有意になりました。このことから，共変量として所得を追加したとしても，生活の満足度が上がると喫煙本数が減ることがわかります。

9-B 例 9.2（227 ページ）で見た病気と生活の満足度の関係について，本書のウェブサポートページにあるデータセット「`9_2_life_xt.csv`」を使って結果を確認してみましょう。

【解答例】 省略。

第 10 章　マッチング法

確認問題

10-1　6 人の修学年数と，15 歳のときに母親が就業していたかどうかを調べたところ，以下の表のようになりました。

子どもの修学年数	母親の就業	母親就業の傾向スコア
10	1	0.7
9	0	0.7
12	1	0.5
10	0	0.5
16	0	0.2
18	0	0.2

母親の就業は子どもの修学年数にどのような影響を与えているのでしょうか。

(1) 母親が就業していた人々の平均修学年数と母親が就業していなかった人々のそれとを比較してみましょう。どちらのグループの平均修学年数が高くなっていますか。

(2) 15 歳時点での暮らし向きや家計所得といった家庭環境に関する変数を使って母親が就業する傾向スコアを推定すると，表の 3 列目のようになりました。傾向スコアが同じ人々の修学年数を比較してみましょう。母親の就業は子どもの修学年数にどのような影響を与えていると言えるでしょうか。

【解答例】

(1) 母親の就業していた人の平均修学年数は，

$$\frac{10+12}{2}=11$$

なので 11 年になっています。母親の就業していなかった平均修学年数は，

$$\frac{9+10+16+18}{4}=13.25$$

となり，13.25 年になっていますので，母親の就業していなかった人々の方が，母親の就業していた人々よりも平均修学年数は長いことがわかります。

(2) 傾向スコアが 0.7 の 2 人を比べると，母親が就業していた人の方がそうでない人よりも 1 年修学年数が長くなっていたことがわかります。同様に，傾向スコアが 0.5 の 2 人を比べてみても，母親が就業していた人の方がそうでない人よりも 2 年修学年数が長くなっていたことがわかります。これらのことから，母親が就業するという確率を表す傾向スコアの同じ人々の比較から，この例では母親の就業は子どもの修学年数に正の影響を与えていた可能性を示唆しています。

実証分析問題

10-A　母親の就業が既婚の娘の就業選択に与える影響を調べましょう。本書のウェブサポートページにあるデータセット「10_2_work.csv」には既婚女性 1132 人のデータが収録されています。このデータセットには現在就業していれば 1，そうでなければ 0 となる就業ダミー変数（work）をはじめ，15 歳のときに母親が就業していたら 1，そうでなければ 0 となる母親就業ダミー変数（mowork15），両親の学歴（mocograd および pacograd），および 15 歳時点の暮らし向き（life15）や学業成績（academic15），家庭の蔵書数（books15）が含まれています。

(1) 母親就業ダミー変数（mowork15）を両親の学歴（mocograd および pacograd），15 歳時点の暮らし向き（life15），学業成績（academic15），家庭の蔵書数（books15）に回帰する線形確率モデルを推定し，その予測値（母親が就業していた傾向スコア（score））を計算しましょう。

(2) 傾向スコアが（0, 0.65），（0.65, 0.7），（0.7, 0.74），（0.74, 0.78），（0.78, 0.82），（0.82, 1）のいずれかとなる標本からなる 6 つのグループを作り，それぞれのグループに対して，母親が就業していた人々とそうでない人々の現在の就業者割合を計算して，比較してみましょう。母親の就業は，既婚の娘の就業選択に影響を与えていますか。

(3) 現在の就業ダミー変数（work）を母親の就業ダミー変数（mowork15）と傾向スコア（score）に回帰して，母親就業ダミー変数の係数パラメターを推定してみましょう。傾向スコアを含めない場合の結果と比較して，どのような違いが見られますか。

【解答例】

(1) Stata を使って線形確率モデルを推定した結果は次のようになります。

```
regress mowork15 mocograd pacograd life15 academic15 books15
```

```
      Source |       SS       df       MS              Number of obs =    1132
-------------+------------------------------           F(  5,  1126) =    6.35
       Model |  5.92641325     5  1.18528265           Prob > F      =  0.0000
    Residual |   210.26705  1126  .186738055           R-squared     =  0.0274
-------------+------------------------------           Adj R-squared =  0.0231
       Total |  216.193463  1131  .191152487           Root MSE      =  .43213

------------------------------------------------------------------------------
    mowork15 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    mocograd |   .0737983   .0701761     1.05   0.293    -.0638923    .2114889
    pacograd |  -.1056511   .0357027    -2.96   0.003    -.1757025   -.0355998
      life15 |  -.0284479   .0170811    -1.67   0.096    -.0619622    .0050665
  academic15 |  -.0022165   .0120106    -0.18   0.854    -.0257822    .0213492
     books15 |  -.0214158   .0059494    -3.60   0.000     -.033089   -.0097427
       _cons |   .9057155   .0468994    19.31   0.000     .8136956    .9977355
------------------------------------------------------------------------------
```

傾向スコアを計算するためには，この推定結果を使って，被説明変数の当てはめ値を計算すれば良いです。Stata ではこの推定の直後に「predict（新しい変数名）」とすれば，当てはめ値を新しい名前の変数としてデータセットに追加してくれます。ここでは新しい変数名を「score」として傾向スコアを作ってみましょう。Stata に「predict score」と書くと，

```
predict score
(option xb assumed; fitted values)
```

と表示され，データセットに score という名前の傾向スコアが作成されます。

(2) (1) で計算した傾向スコアの記述統計を見ると，次のようになります。

```
. summarize score

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       score |      1132    .7429329    .0723877    .527496   .9057155
```

この表から，傾向スコアの最小値は 0.4848947，最大値は 0.9128741 であることがわかります。これらの 1132 人を，傾向スコアの似た人たちの 6 つのグループに分けてみましょう。ここでの 6 つのグループへの分け方は特に意味はないのですが，一つの目安として，それぞれのグループに含まれる人の数が 100 人を下回らないように，$(0, 0.65)$,

(0.65, 0.7), (0.7, 0.74), (0.74, 0.78), (0.78, 0.82), (0.82, 1)としてあります。

例えば，傾向スコアの値が 0.65 未満の人々について，母親の就業状態ごとの平均就業割合を計算すると次のようになります。

母親が就業していなかった人々

```
. summarize work if mowork15==0 & score<.65

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        work |        55    .5090909     .504525         0          1
```

母親が就業していた人々

```
. summarize work if mowork15==1 & score<.65

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        work |        76    .6052632    .4920419         0          1
```

この 2 つの就業割合の比較から，母親の就業の傾向スコアが 0.65 未満の人々は，母親が就業していた方がそうでなかった場合よりも就業割合は高く，その差は$0.60526 - 0.50909 = 0.09617$だけ高かったことがわかります。

なお，この差は次の単回帰モデルを推定して，母親就業ダミー変数の係数パラメターの推定値としても求めることができます。

```
. regress work mowork15 if score<.65

      Source |       SS       df       MS              Number of obs =     131
-------------+------------------------------           F(  1,   129) =    1.19
       Model |     .295124     1     .295124           Prob > F      =  0.2767
    Residual |  31.9033493   129  .247312785           R-squared     =  0.0092
-------------+------------------------------           Adj R-squared =  0.0015
       Total |  32.1984733   130  .247680564           Root MSE      =  .49731

------------------------------------------------------------------------------
        work |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    mowork15 |   .0961722   .0880381     1.09   0.277    -.0780133    .2703578
```

```
       _cons |   .5090909   .0670567     7.59   0.000     .3764177    .6417642
------------------------------
```

他のグループに関しても同様に就業割合の差を比較してまとめたものが次の表になります。

傾向スコア	(0, 0.65)	(0.65, 0.7)	(0.7, 0.74)	(0.74, 0.78)	(0.78, 0.82)	(0.82, 1)
就業割合の差	0.09617	0.1453808	0.0192513	0.1543903	−0.0648439	0.102657

このように, 5つのグループにおいて, 15歳時に母親が就業していた既婚女性の方が, 母親が就業していなかった人々よりも現在の就業割合が高くなっていることがわかります。

(3) Stata を使って，傾向スコアを含めない場合と含めた場合の推定結果はそれぞれ次のようになります。

傾向スコアを制御しない場合

```
. regress work mowork15

      Source |       SS       df       MS              Number of obs =    1132
-------------+------------------------------           F(  1,  1130) =    5.86
       Model |  1.44371394     1  1.44371394            Prob > F      =  0.0157
    Residual |  278.481198  1130  .246443538            R-squared     =  0.0052
-------------+------------------------------           Adj R-squared =  0.0043
       Total |  279.924912  1131  .247502132            Root MSE      =  .49643

------------------------------------------------------------------------------
        work |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    mowork15 |   .0817183   .0337627     2.42   0.016     .0154736     .147963
       _cons |   .4914089   .0291013    16.89   0.000     .4343103    .5485076
------------------------------------------------------------------------------
```

傾向スコアを制御する場合

```
regress work mowork15 score

      Source |       SS       df       MS              Number of obs =    1132
```

```
-------------+------------------------------           F(  2,  1129) =    3.50
       Model |   1.7268664     2  .863433199            Prob > F      =  0.0304
    Residual |  278.198045  1129  .246411023            R-squared     =  0.0062
-------------+------------------------------          Adj R-squared =  0.0044
       Total |  279.924912  1131  .247502132            Root MSE      =   .4964


------------------------------------------------------------------------------
        work |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    mowork15 |   .0756426    .034233     2.21   0.027     .0084752    .1428099
       score |   .2216409   .2067616     1.07   0.284    -.1840392     .627321
       _cons |   .3312585   .1522067     2.18   0.030     .0326188    .6298982
------------------------------------------------------------------------------
```

これらの推定結果から，傾向スコアを含めない場合は，15 歳時点で母親が就業していると，現在の就業確率が約 8.2%高くなる統計的にも有意な関係がわかります。また，傾向スコアを含めた場合では，15 歳時点で母親が就業していると，現在の就業確率が約 7.6%高くなる統計的にも有意な関係があることがわかります。これらの二つの結果から，傾向スコアを考慮しないと影響は過大に計測されていたと見ることもできますが，その差はあまり大きなものではありませんでした。このことは，二つ目の回帰モデルにおいて傾向スコアの傾きパラメターの推定値は統計的に有意ではないということとも整合的だといえます。

第 11 章　回帰不連続デザイン

実証分析問題

マサチューセッツ工科大学（MIT）の経済学部のアングリスト（Joshua Angrist）教授のウェブサイト（http://economics.mit.edu/faculty/angrist）で，「Data and Programs」→「Angrist Data Archive」と進んでいったページには，ここで紹介した Angrist and Levy（1999）で用いられたデータ（Stata 用）と，論文中にある表を作成するのに使われた Stata のプログラムファイル（do ファイル）が掲載されています（do ファイルの使い方については，ソフトウェアの参考書として紹介しているものや本書のウェブサポートページを参照してください)。

教科書では，アングリストとレヴィの作成したデータと Stata 用 do ファイルを使って，Joshua D. Angrist and Victor Lavy “Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement,” *Quarterly Journal of Economics*, **114** (2): 533-575, 1999，の結果を再現する手順を紹介しました。ホームページには，その手順で修正を施した Stata 用の do ファイル（表 2 と表 4 用）があります。

ここでは，アングリストとレヴィのデータを使って，ほぼ同じ推定結果を得るためのより簡単なプログラム紹介します。ホームページからダウンロードできる Stata 用 do ファイル「11_1_ex.do」は，小学 5 年生のデータを使って，表 2 および表 4 の結果を再現するものです（標準誤差が若干異なるのは，計算に用いるプログラムが若干違うためですが，大きな差はありません)。プログラムにはないコメントは（　）に入れてあります。

```
===プログラムここから===
* Estimation Code for Angrist and Lavy (1999)
* November 1st, 2015
* Ryuichi Tanaka (University of Tokyo)

clear　（すでに読み込んであるデータを削除します）
use final5.dta　（小学 5 年生のデータを読み込みます。あらかじめ final5.dta をこの do ファイルと一緒に同じフォルダに入れておき，そのフォルダを「working directory」にしておいてください。）

* clearning data（データを整理します）
replace avgverb= avgverb-100 if avgverb>100
replace avgmath= avgmath-100 if avgmath>100
```

```
replace avgverb=. if verbsize==0
replace avgmath=. if mathsize==0
keep if 1<classize & classize<45 & c_size>5
generate c_size2= (c_size^2)/100
```
（一学年生徒数の 2 乗の項を作ります。）

```
* predicted class size
generate func1= c_size/(int((c_size-1)/40)+1)
```
（一学年の生徒数から予想される学級生徒数を作ります。）

```
* generate discontinuity sample (+-5)
```
（不連続点の前後 5 人のデータのみを使った分析に使います。変数 disc はダミー変数で，一学年の生徒数が不連続点の前後 5 人ならば 1 となります。）

```
generate disc= (c_size>=36 & c_size<=45) | (c_size>=76 & c_size<=85) | (c_size>=116 & c_size<=125)
```

```
* generate trend
```
（トレンド項を作ります。）

```
generate trend= c_size if c_size>=0 & c_size<=40
replace trend= 20+(c_size/2) if c_size>=41 & c_size<=80
replace trend= (100/3)+(c_size/3) if c_size>=81 & c_size<=120
replace trend= (130/3)+(c_size/4) if c_size>=121 & c_size<=160
```

```
keep avgverb avgmath classize tipuach trend func1 disc schlcode c_size c_size2
```
（分析に使う変数のみを残して, 使わない変数をデータセットから削除します。この後の分析に影響はないので, 削除しなくても大丈夫です。）

（ここで label コマンドを使って変数の説明を追記します。この後の分析には影響はありません。）

```
label variable classize "class size"
label variable c_size "enrollment"
label variable c_size2 "c_size^2"
label variable tipuach "share of students from disadvantaged household"
label variable func1 "predicted class size"
label variable trend "trend"
label variable disc "=1 for discontinuity sample"
```

```
* Table 2
```
（表 2 の数値を計算します。）

```
* verbal
```
（読解力）

```
summarize avgverb avgmath
```
（読解力と算数のクラス平均点の記述統計を計算します）

regress avgverb classize, cluster(schlcode) （読解力のクラス平均点を学級生徒数に回帰する単回帰モデルを最小 2 乗法で推定します。最後の「cluster(schlcode)」は，同一学校内で誤差項が相関していても頑健な標準誤差を計算するためのオプションコマンドです。）

regress avgverb classize tipuach, cluster(schlcode) （共変量として貧困家庭比率を追加して分析を行います。）

regress avgverb classize tipuach c_size, cluster(schlcode) （共変量としてさらにトレンド項を追加して分析を行います。）

* mathematics（算数）

regress avgmath classize, cluster(schlcode)

regress avgmath classize tipuach, cluster(schlcode)

regress avgmath classize tipuach c_size, cluster(schlcode)

* Table 4（表 4）

* verbal（読解力）

（実際の学級生徒数に対して，予想される学級児童数を操作変数とする 2 段階最小 2 乗法で推定を行っています。共変量として貧困家庭比率を含めています。）

ivregress 2sls avgverb (classize=func1) tipuach, cluster(schlcode)

ivregress 2sls avgverb (classize=func1) tipuach c_size, cluster(schlcode)（一学年生徒数を共変量として追加しています。）

ivregress 2sls avgverb (classize=func1) tipuach c_size c_size2, cluster(schlcode) （一学年生徒数の 2 乗項を共変量として追加しています。）

ivregress 2sls avgverb (classize=func1) trend, cluster(schlcode) （共変量として，トレンド項のみを用いた推定を行います。）

ivregress 2sls avgverb (classize=func1) tipuach if disc==1, cluster(schlcode)（不連続点の前後 5 人のデータのみを使って推定しています。）

ivregress 2sls avgverb (classize=func1) tipuach c_size if disc==1, cluster(schlcode) 不連続点の前後 5 人のデータのみを使って推定しています。）

* mathematics （算数）

ivregress 2sls avgmath (classize=func1) tipuach, cluster(schlcode)

ivregress 2sls avgmath (classize=func1) tipuach c_size, cluster(schlcode)

ivregress 2sls avgmath (classize=func1) tipuach c_size c_size2, cluster(schlcode)

ivregress 2sls avgmath (classize=func1) trend, cluster(schlcode)

ivregress 2sls avgmath (classize=func1) tipuach if disc==1, cluster(schlcode)

```
ivregress 2sls avgmath (classize=func1) tipuach c_size if disc==1, cluster(schlcode)
```

===プログラムここまで===