2024/04/12

藪友良

第 5 章の答え

練習問題 1

5.3.1 節で学習したとおり、被説明変数をY_i、説明変数をX_iとした単回帰分析では、X_iの係数$\hat{\beta}_1$の期待値は、

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

となる。ここで、$\beta_2 > 0$、$s_{XW} < 0$であるから、欠落変数バイアスは負となる。

$$\beta_2 \frac{s_{XW}}{s_X^2} < 0$$

これは、$E[\hat{\beta}_1]$は真の値β_1より小さいことを意味する。

$$E[\hat{\beta}_1] < \beta_1$$

まとめると、生まれつきの能力W_iを除くことで、職業訓練ダミーX_iの係数(職業訓練の効果)は低めに推定されてしまう。

練習問題 2

ここで、$\beta_2 < 0$、$s_{XW} > 0$であるから、欠落変数バイアスは負となる($\beta_2 \frac{s_{XW}}{s_X^2} < 0$)。これは、$E[\hat{\beta}_1]$は真の値$\beta_1$より小さいことを意味する($E[\hat{\beta}_1] < \beta_1$)。つまり、移民の割合を除くことで、クラスの人数$X_i$の係数は低めに推定される。これは単回帰分析のほうが、重回帰分析よりも係数が小さくなる(クラスの人数を減らすことの効果が大きくなる)ことを意味する。

練習問題 3

双子の差をとると、

$$Y_i^{兄} - Y_i^{弟} = \left(\alpha + \beta_1 X_i^{兄} + \beta_2 W_i^{兄} + u_i^{兄}\right) - (\alpha + \beta_1 X_i^{弟} + \beta_2 W_i^{弟} + u_i^{弟})$$

$$= \beta_1\left(X_i^{兄} - X_i^{弟}\right) + \beta_2\left(W_i^{兄} - W_i^{弟}\right) + \left(u_i^{兄} - u_i^{弟}\right)$$

$$= \beta_1(X_i^{兄} - X_i^{弟}) + (u_i^{兄} - u_i^{弟})$$

となる。式展開では、$W_i^{兄} - W_i^{弟} = 0$とした。ここで、$Y_i = Y_i^{兄} - Y_i^{弟}$、$X_i = X_i^{兄} - X_i^{弟}$、$u_i = u_i^{兄} - u_i^{弟}$と定義すれば、上式は通常の単回帰分析によって推定できる。つまり、双子のデータを用いる利点は、双子の差をとることで、生まれつきの能力の要因を取り除くことが可能となる点にある。

現実問題として、双子であれば教育年数にあまり違いはない可能性がある。つまり、教育年数の差$X_i = X_i^{兄} - X_i^{弟}$はほぼ 0 の値をとり、説明変数の変動が非常に小さくなる。これでは、推定結果は不安定となる。3.3.3 節では、説明変数の変動が大きいほど、OLS 推定量の分散が小さくなることを説明している(図 3-4(a)参照)。

練習問題 4

5.6 節をもとに答えをまとめる。決定係数R^2は、

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

と定義され、説明変数X_iの数Kが増えるほど、その値が 1 に近づくという性質がある。これに対して、自由度調整済み決定係数は

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

と定義される。自由度調整済み決定係数では、調整項$\frac{n-1}{n-K-1}$を含めることで、説明変数を含めることに罰則を課している。つまり、自由度調整済み決定係数では、悪い説明変数を含めると、逆に、その値が下がることになる。

練習問題 5

都道府県の転入超過数とは、転入者数から転出者数を引いた値となる。つまり、

転入超過数　＝　転入者数　－　転出者数

という関係がある。これは恒等式であり、そもそも推定する意味はない。

練習問題 6

残差 2 乗和は、$X_2 = 10X_1$に注意すると、次のようになる。

$$\sum_{i=1}^{n}\tilde{u}_i^2=\sum_{i=1}^{n}(Y_i-\tilde{\alpha}-\tilde{\beta}_1X_{1i}-\tilde{\beta}_2X_{2i})^2=\sum_{i=1}^{n}(Y_i-\tilde{\alpha}-(\tilde{\beta}_1+10\tilde{\beta}_2)X_{1i})^2$$

これを各パラメータ$(\tilde{\alpha},\tilde{\beta}_1,\ \tilde{\beta}_2)$で偏微分して 0 と置くと 3 本の式が得られる。

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

②式と③式は、本質的に同じ式となる。これは②式の両辺を－2 で割る、③式の両辺を－20 で割ると、次式となることから明らかである。

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

以上から、正規方程式は 2 本の独立な式だけとなる。

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

独立な式は 2 本、パラメータは 3 つあるため、OLS 推定量を求めることはできない。OLS 推定量を導出するためには、独立な式がパラメータの数と同じだけ必要である。

練習問題 7

残差 2 乗和は次のようになる。

$$\sum_{i=1}^{n}\tilde{u}_i^2=\sum_{i=1}^{n}(Y_i-\tilde{\alpha}-\tilde{\beta}_1X_{1i}-\tilde{\beta}_2X_{2i})^2$$

これを各パラメータ$(\tilde{\alpha},\tilde{\beta}_1,\ \tilde{\beta}_2)$で偏微分して 0 と置くと 3 本の式が得られる。

①	$\frac{\partial\sum\tilde{u}_i^2}{\partial\tilde{\alpha}}=\sum\frac{\partial\tilde{u}_i^2}{\partial\tilde{u}_i}\frac{\partial\tilde{u}_i}{\partial\tilde{\alpha}}=\sum 2\tilde{u}_i(-1)=-2\sum(Y_i-\tilde{\alpha}-\tilde{\beta}_1X_{1i}-\tilde{\beta}_2X_{2i})=0$
②	$\frac{\partial}{\partial\tilde{\beta}_1}\sum\tilde{u}_i^2=\sum\frac{\partial\tilde{u}_i^2}{\partial\tilde{u}_i}\frac{\partial\tilde{u}_i}{\partial\tilde{\beta}_1}=\sum 2\tilde{u}_i(-X_{1i})=-2\sum(Y_i-\tilde{\alpha}-\tilde{\beta}_1X_{1i}-\tilde{\beta}_2X_{2i})X_{1i}=0$

③	$\frac{\partial}{\partial\tilde{\beta}_2}\sum\tilde{u}_i^2=\sum\frac{\partial\tilde{u}_i^2}{\partial\tilde{u}_i}\frac{\partial\tilde{u}_i}{\partial\tilde{\beta}_2}=\sum 2\tilde{u}_i(-X_{2i})=-2\sum(Y_i-\tilde{\alpha}-\tilde{\beta}_1X_{1i}-\tilde{\beta}_2X_{2i})X_{2i}=0$

3本の式は独立に見えるが、これは誤りである。②式と③式を足すと、

$$-2\sum(Y_i-\tilde{\alpha}-\tilde{\beta}_1X_{1i}-\tilde{\beta}_2X_{2i})X_{1i}-2\sum(Y_i-\tilde{\alpha}-\tilde{\beta}_1X_{1i}-\tilde{\beta}_2X_{2i})X_{2i}=0$$

となる。ここで左辺をまとめると、

$$-2\sum(Y_i-\tilde{\alpha}-\tilde{\beta}_1X_{1i}-\tilde{\beta}_2X_{2i})(X_{1i}+X_{2i})=-2\sum(Y_i-\tilde{\alpha}-\tilde{\beta}_1X_{1i}-\tilde{\beta}_2X_{2i})$$

となる(式展開では、$X_1+X_2=1$を用いた)。つまり、②式+③式は、①式であり、独立な式は2本だけとわかる。独立な式は2本(②式と③式)、パラメータは3つ$(\tilde{\alpha},\ \tilde{\beta}_1,\ \tilde{\beta}_2)$あるため、OLS推定量を求めることはできない。

練習問題 8

残差2乗和$\sum_{i=1}^n\tilde{u}_i^2=\sum_{i=1}^n\left(Y_i-\tilde{\alpha}-\tilde{\beta}_1X_{1i}-\tilde{\beta}_2X_{2i}\right)^2$を$\tilde{\alpha}$で偏微分して0と置くと、

$$-2\sum_{i=1}^n(Y_i-\hat{\alpha}-\hat{\beta}_1X_{1i}-\hat{\beta}_2X_{2i})=0$$

となり、両辺を-2で割ると

$$\sum_{i=1}^n(Y_i-\hat{\alpha}-\hat{\beta}_1X_{1i}-\hat{\beta}_2X_{2i})=0$$

となる。上式を満たす$\tilde{\alpha}$、$\tilde{\beta}_1$、$\tilde{\beta}_2$は最小2乗推定量なので、「^(ハット)」を付けた。この式を展開すると、

$$\sum_{i=1}^nY_i-n\hat{\alpha}-\hat{\beta}_1\sum_{i=1}^nX_{1i}-\hat{\beta}_2\sum_{i=1}^nX_{2i}=0$$

となり、さらに両辺をnで割って、$\hat{\alpha}$について解くと次式が得られる。

$$\hat{\alpha}=\bar{Y}-\hat{\beta}_1\bar{X}_1-\hat{\beta}_2\bar{X}_2$$

次に、残差2乗和$\sum_{i=1}^n\tilde{u}_i^2=\sum_{i=1}^n\left(Y_i-\tilde{\alpha}-\tilde{\beta}_1X_{1i}-\tilde{\beta}_2X_{2i}\right)^2$を$\tilde{\beta}_1$で偏微分して0と置くと、

$$-2\sum_{i=1}^n(Y_i-\hat{\alpha}-\hat{\beta}_1X_{1i}-\hat{\beta}_2X_{2i})\,X_{1i}=0$$

となり、さらに両辺を-2で割ると、

$$\sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i})\, X_{1i} = 0$$

となる。ここで、$\hat{\alpha} = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2$を上式に代入すると、下式となる。

$$\sum_{i=1}^{n}\left(Y_i - (\bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2) - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i}\right) X_{1i}$$

$$= \sum_{i=1}^{n}\left((Y_i - \bar{Y}) - \hat{\beta}_1 (X_{1i} - \bar{X}_1) - \hat{\beta}_2 (X_{2i} - \bar{X}_2)\right) X_{1i}$$

$$= \sum_{i=1}^{n}(Y_i - \bar{Y}) X_{1i} - \hat{\beta}_1 \sum_{i=1}^{n}(X_{1i} - \bar{X}_1) X_{1i} - \hat{\beta}_2 \sum_{i=1}^{n}(X_{2i} - \bar{X}_2) X_{1i} = 0$$

偏差の和は 0 から、上式は次のように書き換えられる[1]。

$$\sum_{i=1}^{n}(Y_i - \bar{Y})(X_{1i} - \bar{X}_1) - \hat{\beta}_1 \sum_{i=1}^{n}(X_{1i} - \bar{X}_1)(X_{1i} - \bar{X}_1) - \hat{\beta}_2 \sum_{i=1}^{n}(X_{2i} - \bar{X}_2)(X_{1i} - \bar{X}_1) = 0$$

ここで、X_{1i}とX_{2i}の標本共分散は 0 ならば、上式の左辺第 3 項は 0 となる$(\sum_{i=1}^{n}(X_{2i} - \bar{X}_2)(X_{1i} - \bar{X}_1) = 0)$。よって、上式は、

$$\sum_{i=1}^{n}(Y_i - \bar{Y})(X_{1i} - \bar{X}_1) - \hat{\beta}_1 \sum_{i=1}^{n}(X_{1i} - \bar{X}_1)^2 = 0$$

となり、これを$\hat{\beta}_1$について解けば次式が得られる。

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)^2}$$

以上より、重回帰分析における OLS 推定量$\hat{\beta}_1$は、Y_iをX_{1i}だけで単回帰したときの OLS 推定量の式と同じである。つまり、説明変数間の相関が 0 であるなら、単回帰分析であっても欠落変数バイアスが生じないことがわかる。

練習問題 9

被説明変数を所得Y_iとし、説明変数を教育年数X_iと職種W_iとした、次の重回帰

---

[1]偏差の和は 0（$\sum_{i=1}^{n}(X_{1i} - \bar{X}_1) = 0$、$\sum_{i=1}^{n}(X_{2i} - \bar{X}_2) = 0$）であるから、以下の式が成立する。

$$\sum_{i=1}^{n}(X_{1i} - \bar{X}_1) X_{1i} = \sum_{i=1}^{n}(X_{1i} - \bar{X}_1) X_{1i} - \sum_{i=1}^{n}(X_{1i} - \bar{X}_1)\bar{X}_1 = \sum_{i=1}^{n}(X_{1i} - \bar{X}_1)(X_{1i} - \bar{X}_1)$$

$$\sum_{i=1}^{n}(X_{2i} - \bar{X}_2) X_{1i} = \sum_{i=1}^{n}(X_{2i} - \bar{X}_2) X_{1i} - \sum_{i=1}^{n}(X_{2i} - \bar{X}_2)\bar{X}_1 = \sum_{i=1}^{n}(X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)$$

式展開では、$\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)\bar{X}_1 = \bar{X}_1 \sum_{i=1}^{n}(X_{1i} - \bar{X}_1) = \bar{X}_1 \times 0 = 0$、$\sum_{i=1}^{n}(X_{2i} - \bar{X}_2)\bar{X}_1 = \bar{X}_1 \sum_{i=1}^{n}(X_{2i} - \bar{X}_2) = 0$を用いた。

モデルを考える。

$$Y_i = \alpha + \beta_1 X_i + \beta_2 W_i + u_i \qquad ①$$

ここで、教育年数X_iの係数β_1は、「職種W_iを一定とし、教育年数X_iが 1 年増えたとき、所得Y_iがいくら変化するか」を示している。こうした教育の効果に関心があるなら、この推定をしてもよいだろう。

ここで、職種は教育年数に依存して、次のように決まるとする。

$$W_i = \theta_0 + \theta_1 X_i + e_i \qquad ②$$

この式を、所得Y_iの式に代入すると、

$$Y_i = \alpha + \beta_1 X_i + \beta_2(\theta_0 + \theta_1 X_i + e_i) + u_i$$

$$= \underbrace{(\alpha + \beta_2\theta_0)}_{=\delta_0} + \underbrace{(\beta_1 + \beta_2\theta_1)}_{=\delta_1} X_i + \underbrace{(u_i + \beta_2 e_i)}_{=\varepsilon_i}$$

となり、次の単回帰モデルが得られる。

$$Y_i = \delta_0 + \delta_1 X_i + \varepsilon_i \qquad ③$$

ただし、$\delta_0 = \alpha + \beta_2\theta_0$、$\delta_1 = \beta_1 + \beta_2\theta_1$、$\varepsilon_i = u_i + \beta_2 e_i$とした。③式は、①式と②式を統合したモデルとなっている。また、教育年数X_iは確率変数ではないため、X_iは$\varepsilon_i = u_i + \beta_2 e_i$と無相関になる[2]。つまり、単回帰モデルを推定しても、欠落変数バイアスは生じない。ここで、教育年数の係数δ_1は、次のようになる。

$$\delta_1 = \underbrace{\beta_1}_{\text{直接的効果}} + \underbrace{\beta_2\theta_1}_{\text{間接的効果}}$$

つまり、係数β_1は教育年数の「直接的効果」、$\beta_2\theta_1$は教育年数が職種に影響を与えることから生じる「間接的効果」を合わせたものとなる。

練習問題 10、11、12、13

ウェブサイトから、データと再現コードをダウンロードしてもらいたい。

---

[2] ここで、u_iとe_iは誤差項であり、期待値はそれぞれ 0 となる。このため、ε_iの期待値も次のとおり 0 となる。

$$E[\varepsilon_i] = E[u_i + \beta_2 e_i] = E[u_i] + \beta_2 E[e_i] = 0 + \beta_2 \times 0 = 0$$

X_iは確率変数ではないことに注意すると、X_iとε_iとの共分散は次のとおり 0 となる。

$$Cov(X_i, \varepsilon_i) = E[(X_i - \bar{X})\varepsilon_i] = (X_i - \bar{X})E[\varepsilon_i] = 0$$

＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊

こちらは初版(第 1 刷)には掲載されていない新しい問題になります。

練習問題 14

14. 以下の 4 ケースにおいて多重共線性が生じる理由を述べよ。

(a) 勤続年数が分からないため、勤続年数を年齢－教育年数－6 として計算した。説明変数は、年齢、教育年数、勤続年数とする。

(b) 説明変数は、ダミー変数D_i、その 2 乗D_i^2とする。

(c) 説明変数X_{1i}は、すべて 0 となる変数とする。

(d) 説明変数X_{1i}は、すべて 1 となる変数とする。

練習問題 14 の答え

(a)　勤続年数は、次のように定義される[3]。

勤続年数=年齢－教育年数－6

ここで、教育年数は、小卒なら 6 年、中卒なら 9 年、高卒なら 12 年、大卒なら 16 年となる。たとえば、40 歳の大卒なら、勤続年数は 18 年(=40－16－6)である。勤続年数=年齢－教育年数－6 という関係式から、－6－年齢+教育年数－勤続年数=0 となり、多重共線性が成立する。

(b) ダミー変数は 0 から 1 の値をとるので、2 乗しても値がかわらない(0 の 2 乗は 0 であり、1 の 2 乗は 1 である)。つまり、 $D_i = D_i^2$であり、多重共線性が生じることになる。

(c) $X_{1i} = 0$であるから、c_1以外をすべて 0 としても($c_1 \neq 0$、$c_0 = c_2 = \cdots = c_K = 0$)、

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

が成立する。ここで、$c_1 \neq 0$であるから、多重共線性が成立している。

(d)この場合、$c_0 = -1$、$c_1 = 1$、$c_2 = \cdots = c_K = 0$と設定すれば、

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

が成立する。

なお、(c)(d)といった状況は、データのサブサンプルを考えるときに生じやす

[3] 勤続年数がデータとして利用できない場合、こうした計算式を用いることになる。かりに勤続年数が正確にわかるなら、この計算式が成立しないこともあるだろう。たとえば、大学入学前に浪人していれば、この式は成立しなくなり、多重共線性の問題も生じない。

い。たとえば、職業別の男女賃金格差に関心があり、データを収集したとしよう。このデータには男女のデータが含まれているが、歯科衛生士だけをデータから取り出した場合、男性の人数はかなり少なくなってしまう(歯科衛生士の多くは女性である)。偶然、男性が全く含まれないなら、女性ダミーは 1 だけになってしまう。