2024/04/22

藪友良

第 9 章の答え

練習問題 1

被説明変数がダミー変数であるとき、$Y_i = 1$となる確率P_iは、次のような線形モデルで表せる。

$$P_i = \alpha + \beta X_i$$

これが線形確率モデルと言われる理由である(詳しくは例 9-2 参照)。

練習問題 2

OLS 推定量は不偏性と一致性を持つ。しかし、ガウス=マルコフの条件が満たされないため、有効性は満たされない。なお、OLS 推定量$\hat{\beta}$の分散は、

$$\sigma_{\hat{\beta}}^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2\sigma_i^2}{\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^2}$$

であるため、通常の標準誤差ではなく、ロバスト標準誤差を用いることが必要である。

練習問題 3

被説明変数Y_iは、企業から連絡があれば 1、連絡がなければ 0 となるダミー変数となり、説明変数X_iは、黒人固有の名前なら 1、白人固有の名前なら 0 となるダミー変数となる。誤差項としては、個人iの属性が考えられる。名前はランダムに割り当てられたため、説明変数X_iは個人属性などを表す誤差項と無相関となっており、OLS 推定にバイアスは生じない(8.3 節参照)。

練習問題 4

ここで、$u_i^* = \sqrt{N_i}u_i$である。よって、u_i^*の分散は、次のようになる。

$$E[u_i^{*2}] = E\left[\left(\sqrt{N_i}u_i\right)^2\right] = N_i E[u_i^2]$$

また、$E[u_i^2] = \frac{\sigma^2}{N_i}$であるから、次式のように誤差項$u_i^*$の分散は$\sigma^2$で一定である。

$$N_i E[u_i^2] = N_i \frac{\sigma^2}{N_i} = \sigma^2$$

練習問題 5

ここで、$h_i = Z_i$であるため、$Y_i = \alpha + \beta X_i + u_i$の両辺に$1/\sqrt{h_i} = 1/\sqrt{Z_i}$を掛けると、

$$\underbrace{\frac{Y_i}{\sqrt{Z_i}}}_{=Y_i^*} = \alpha \underbrace{\frac{1}{\sqrt{Z_i}}}_{=X_{1i}^*} + \beta \underbrace{\frac{X_i}{\sqrt{Z_i}}}_{=X_{2i}^*} + \underbrace{\frac{u_i}{\sqrt{Z_i}}}_{=u_i^*}$$

となる。新しい誤差項u_i^*の分散は、

$$E[u_i^{*2}] = E\left[\left(\frac{u_i}{\sqrt{Z_i}}\right)^2\right] = \frac{1}{Z_i}E[u_i^2] = \frac{1}{Z_i}cZ_i = c$$

となるため、均一分散を満たす。被説明変数をY_i^*、説明変数をX_{1i}^*、X_{2i}^*としたOLS推定をすればWLS推定量となる。

練習問題 6

ここで、$h_i = X_i^2$であるため、$Y_i = \alpha + \beta X_i + u_i$の両辺に$1/\sqrt{h_i} = 1/|X_i|$を掛けると、

$$\underbrace{\frac{Y_i}{|X_i|}}_{=Y_i^*} = \alpha \underbrace{\frac{1}{|X_i|}}_{=X_{1i}^*} + \beta \underbrace{\frac{X_i}{|X_i|}}_{=X_{2i}^*} + \underbrace{\frac{u_i}{|X_i|}}_{=u_i^*}$$

となる(X_iは負の値をとる可能性があるため、ここでは絶対値をとっている)。新しい誤差項u_i^*の分散は、

$$E[u_i^{*2}] = E\left[\left(\frac{u_i}{|X_i|}\right)^2\right] = \frac{1}{X_i^2}E[u_i^2] = \frac{1}{X_i^2}cX_i^2 = c$$

となるため、均一分散を満たす。被説明変数をY_i^*、説明変数をX_{1i}^*、X_{2i}^*としたOLS推定をすればWLS推定量となる。

練習問題 7

WLS推定について考えてみよう。仮にσ_iが分かっているならば、元の式をσ_iで割ると、

$$\underbrace{\frac{Y_i}{\sigma_i}}_{=Y_i^*} = \alpha \underbrace{\frac{1}{\sigma_i}}_{=X_{1i}^*} + \beta \underbrace{\frac{X_i}{\sigma_i}}_{=X_{2i}^*} + \underbrace{\frac{u_i}{\sigma_i}}_{=u_i^*}$$

となり、新しい誤差項u_i^*の分散は、

$$E[u_i^{*2}] = E\left[\left(\frac{u_i}{\sigma_i}\right)^2\right] = \frac{1}{\sigma_i^2}E[u_i^2] = \frac{1}{\sigma_i^2}\sigma_i^2 = 1$$

となるため、均一分散を満たす。現実には、分析者はσ_iの値を知らないため、予

測値$\hat{\sigma}_i$を用いた FWLS を行う。

まず、$Y_i = \alpha + \beta X_i + u_i$とした OLS 推定によって残差$\hat{u}_i$を求める。残差$\hat{u}_i$は誤差項$u_i$の推定量である。ここで、$\sigma_i^2 = E[u_i^2] = c_0 + c_1 Z_i$から、被説明変数を$\hat{u}_i^2$とし、説明変数を$Z_i$とした OLS 推定によって、パラメータ$(c_0、\ c_1)$を推定でき、分散の予測値$\hat{\sigma}_i^2 = \hat{c}_0 + \hat{c}_1 Z_i$を求めることができる。次に、元のモデルを予測値$\hat{\sigma}_i$で割ることで、次の式が得られる。

$$\underbrace{\frac{Y_i}{\hat{\sigma}_i}}_{=Y_i^*} = \alpha \underbrace{\frac{1}{\hat{\sigma}_i}}_{=X_{1i}^*} + \beta \underbrace{\frac{X_i}{\hat{\sigma}_i}}_{=X_{2i}^*} + \underbrace{\frac{u_i}{\hat{\sigma}_i}}_{=u_i^*}$$

サンプルサイズが十分に大きければ、推定量$\hat{\sigma}_i$は真の値σ_iとなるため、新しい誤差項u_i^*は$u_i/\hat{\sigma}_i$で均一分散を満たす。このため、被説明変数をY_i^*、説明変数をX_{1i}^*、X_{2i}^*とした OLS 推定は、FWLS 推定量となる。

練習問題 8、9

ウェブサイトから、データと再現に必要な STATA と R のコードをダウンロードできる。

＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊

初版(第 1 刷)には含まれていない新しい練習問題

＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊

10. ★　均一分散が正しい状況を考えよう$(E[u_i^2] = \sigma^2)$。ただし、誤差項は互いに独立とする$(E[u_i u_j] = 0)$。この問題では、不均一分散に対して頑健な分散の推定量の性質を調べ、推定量を改善する方法を提示する。

(a) 不均一分散に対して頑健な分散の推定量$s_{\hat{\beta}}^2$は、真の分散より小さくなること、つまり、次式が正しいことを示せ。

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

Hint: 練習問題 3.13(b)から　$E[\hat{u}_i^2] = \sigma^2(1 - h_{ii})$、$h_{ii}$はレバレッジである。

(b) Stata では、不均一分散に対して頑健な分散の推定量として、次の推定量$\hat{V}^{HC1}$を用いている。推定量$\hat{V}^{HC1}$の是非を述べよ。

$$\hat{V}^{HC1} = \frac{n}{n-2}\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2\hat{u}_i^2}{\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^2}$$

Hint: 練習問題 3.13(c)から$\sum_{i=1}^{n} h_{ii} = 2$となる(よって、平均は$2/n$)。

(c) 標準化残差$\tilde{u}_i = (1 - h_{ii})^{-1/2}\hat{u}_i$を用いた次の推定量$\hat{V}^{HC2}$を考えよう。推定量$V^{HC2}$の是非を述べよ。

$$\hat{V}^{HC2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2\tilde{u}_i^2}{\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^2}$$

Hint: 練習問題 3.13(d)から$E[\tilde{u}_i^2] = \sigma^2$となる。

(d) これらの推定量のうち、どれが望ましいといえるか。

11. ★ レバレッジが大きくなるケースを考えよう。回帰モデルは$Y_i = \alpha + \beta D_i + u_i$とし、$D_i$はダミー変数とする。なお、$\sum_{i=1}^{n} D_i = 1$、$E[u_i^2] = \sigma^2$、$E[u_i u_j] = 0$とする。$\sum_{i=1}^{n} D_i = 1$とは、ある$i$でだけ$D_i = 1$となるが、それ以外はすべて 0 となることを意味する(『入門　実践する統計学』12.4.1 節の一時的ダミーを参照)。

(a) 説明変数の偏差 2 乗和は、次のようになることを示せ。

$$\sum_{i=1}^{n}(D_i - \bar{D})^2 = \frac{n-1}{n}$$

(b) $\hat{\beta}$の分散は次のようになることを示せ(均一分散であることに注意)。

$$V(\hat{\beta}) = \sigma^2\frac{n}{n-1}$$

(c) 推定量$\hat{V}^{HC1}$の期待値は次のようになることを示せ。

$$E[\hat{V}^{HC1}] = E\left[\frac{n}{n-2}\frac{\sum_{i=1}^{n}(D_i - \bar{D})^2\hat{u}_i^2}{\left(\sum_{i=1}^{n}(D_i - \bar{D})^2\right)^2}\right] = \frac{V(\hat{\beta})}{n-1}$$

Hint: $D_i = 1$のとき、レバレッジは 1 となり、それ以外では$1/(n-1)$となる。

練習問題 10 の答え

(a)

均一分散のもとで、$E[\hat{u}_i^2] = \sigma^2(1 - h_{ii})$となる。これを用いると、

$$E\left[s_{\hat{\beta}}^2\right] = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 E[\hat{u}_i^2]}{\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^2} = \frac{\sigma^2\sum_{i=1}^{n}(X_i - \bar{X})^2(1 - h_{ii})}{\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^2}$$

$$= \frac{\sigma^2 \sum_{i=1}^{n}(X_i - \bar{X})^2}{\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^2} - \frac{\sigma^2 \sum_{i=1}^{n}(X_i - \bar{X})^2 h_{ii}}{\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^2}$$

$$= \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} - \frac{\sigma^2 \sum_{i=1}^{n}(X_i - \bar{X})^2 h_{ii}}{\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^2}$$

となる。また、$1/n < h_{ii}$から、右辺第 2 項はマイナスであり、

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

が成立する。つまり、均一分散が正しいとき、不均一分散に対して頑健な分散の推定量$s_{\hat{\beta}}^2$は、真の分散$\frac{\sigma^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}$を過小評価している。

(b)

不均一分散に対して頑健な分散の推定量$s_{\hat{\beta}}^2$は、真の分散を過小評価するという問題がある。Stata では、不均一分散に対して頑健な分散の推定量$s_{\hat{\beta}}^2$に$\frac{n}{n-2}$を掛けることで、分散を大きめに推定し、こうした過小評価の問題を軽減している。

$\frac{n}{n-2}$という値は、どのように正当化されるのだろうか。これは$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2$の計算と整合的な調整ともいえる。3.4.1 節で確認したとおり、理想的なσ^2の推定量は$\frac{1}{n}\sum_{i=1}^{n}u_i^2$である。しかし、誤差項$u_i$が観察できないため、残差$\hat{u}_i$で置き換えることになる。このとき、$\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2$ではなく、それに$\frac{n}{n-2}$を掛けた$s^2$を用いた。

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

同じ調整を、不均一分散に対して頑健な分散の推定量$s_{\hat{\beta}}^2$に行っているのが、Stata の調整といえる。なお、重回帰分析では、

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

となるから、不均一分散に対して頑健な分散の推定量に$\frac{n}{n-K-1}$を掛けることになる。

では、こうした調整を行うことで、不偏性が満たされるのだろうか。残念ながら、不偏性が満たされるのは特殊ケースになる。この点を確認しよう。レバレッジの平均は、次のようになる。

$$\frac{1}{n}\sum_{i=1}^{n} h_{ii} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right) = \frac{2}{n}$$

ここで、レバレッジh_{ii}は常に同じ値であるとしよう。レバレッジh_{ii}は平均$2/n$と同じであるため、$1 - h_{ii} = 1 - 2/n = (n-2)/n$となる。このとき、

$$E\left[s_{\hat{\beta}}^2\right] = \frac{\sigma^2 \sum_{i=1}^{n}(X_i - \bar{X})^2 (1 - h_{ii})}{\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^2} = \frac{n-2}{n}\frac{\sigma^2 \sum_{i=1}^{n}(X_i - \bar{X})^2}{\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^2} = \frac{n-2}{n}\frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

となる。よって、不均一分散に対して頑健な分散の推定量として、

$$\hat{V}^{HC1} = \frac{n}{n-2}\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 \hat{u}_i^2}{\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^2}$$

を用いれば、不偏性が満たされることになる。

以上をまとめると、Stata の$\hat{V}^{HC1}$は、σ^2の推定量s^2と整合的な調整になっているが、不偏性が満たされるのは特殊ケースであり、理論的根拠が十分とはいえない[1]。

(c)

標準化残差$\bar{u}_i = (1 - h_{ii})^{-1/2}\hat{u}_i$を使った場合、$E[\bar{u}_i^2] = \sigma^2$が成立する。このため、$\hat{V}^{HC2}$の期待値は、

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

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

となり、$\hat{V}^{HC2}$は不偏性を満たす[2]。

(d)

理論的には、$\hat{V}^{HC2}$が優れているが、実証分析では、$\hat{V}^{HC1}$がよく用いられている(教科書でも、V^{HC1}を用いた)。通常、どちらを用いても同じような値になるので、どちらを使っても問題はないだろう。しかし、レバレッジh_{ii}が大きいデータがあれば、$\hat{V}^{HC2}$の方が$\hat{V}^{HC1}$よりも大きな値をとる傾向がある。これはレバレッジh_{ii}

[1] Stata では、reg Y X, r とすると、$\hat{V}^{HC1}$の平方根がロバスト標準誤差として計算される。

[2] Stata では、reg Y X, vce(hc2)とすれば、$\hat{V}^{HC2}$の平方根がロバスト標準誤差として計算される。

が 1 に近いと、標準化残差$\bar{u}_i = (1-h_{ii})^{-1/2}\hat{u}_i$が大きくなり、ひいては、$\hat{V}^{HC2}$が大きくなるためである。このときは、$\hat{V}^{HC2}$を用いるほうがよい。なお、サンプルサイズが大きければ、レバレッジh_{ii}は 0 に収束するため、いずれを用いても同じ結果となる(h_{ii}の定義を思い出してほしい)。

自分でデータ分析する際は、$\hat{V}^{HC1}$と$\hat{V}^{HC2}$の両方を計算し、それらの値を比較することが望ましい。両者の差が大きく異なるようなら、レバレッジを計算し、どの観測値で大きくなっているかを確認しよう。かりにそれが外れ値のようなものなら、そのデータを除去することも、選択肢の 1 つとして考えられる。たとえば、小学生のデータを分析したところ、身長が 210cm の生徒がいたとしよう。これは入力間違いの可能性があるし、たとえ正しい情報であっても外れ値と考えられる。

練習問題 11 の答え

$Y_i = \alpha + \beta D_i + u_i$とし、$D_i$はダミー変数である。ダミー変数は 0 もしくは 1 をとり、1 をとる回数が計n_1あるとしよう。n_1が小さい、もしくは$n - n_1$が小さい場合、ダミー変数D_iはスパース(sparse)という。このとき、レバレッジh_{ii}は大きくなり、$\hat{V}^{HC1}$は真の分散を過小評価してしまうため、$\hat{V}^{HC2}$を用いることが推奨される。

(a)

単純化のため、$D_1 = 1$、それ以外のD_iは 0 としよう。説明変数D_iの平均は

$$\bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i = \frac{1}{n}$$

となる。このため、偏差は、

$$D_i - \bar{D} = \begin{cases} 1 - \dfrac{1}{n} = \dfrac{n-1}{n} & \text{if } i = 1 \\ 0 - \dfrac{1}{n} = -\dfrac{1}{n} & \text{if } i > 1 \end{cases}$$

となり、偏差 2 乗和は次のようになる。

$$\sum_{i=1}^{n}(D_i - \bar{D})^2 = (D_1 - \bar{D})^2 + \sum_{i=2}^{n}(D_i - \bar{D})^2$$

$$= \left(\frac{n-1}{n}\right)^2 + (n-1)\left(-\frac{1}{n}\right)^2$$

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

(b)

誤差項は均一分散であるため、$\hat{\beta}$の分散は次のようになる。

$$V(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^{n}(D_i - \bar{D})^2} = \frac{n}{n-1}\sigma^2$$

この場合、係数βを推定するための情報は、$D_1 = 1$だけしか存在せず、サンプルサイズが大きくなっても分散はσ^2となる。つまり、サンプルサイズが大きくなっても、$\hat{\beta}$の推定精度は改善しないことがわかる[3]。

(c)

レバレッジを求めよう。$\sum_{i=1}^{n}(D_i - \bar{D})^2 = (n-1)/n$に注意すると、

$$\frac{(D_i - \bar{D})^2}{\sum_{i=1}^{n}(D_i - \bar{D})^2} = \begin{cases} \dfrac{[(n-1)/n]^2}{(n-1)/n} = \dfrac{n-1}{n} & \text{if } i = 1 \\ \dfrac{(-1/n)^2}{(n-1)/n} = \dfrac{1}{n(n-1)} & \text{if } i > 1 \end{cases}$$

となり、レバレッジh_{ii}は、次のようになる。

$$h_{ii} = \begin{cases} \dfrac{1}{n} + \dfrac{n-1}{n} = 1 & \text{if } i = 1 \\ \dfrac{1}{n} + \dfrac{1}{n(n-1)} = \dfrac{1}{n-1} & \text{if } i > 1 \end{cases}$$

ここで、$h_{11} = 1$となり、D_1は他のデータに比べて大きく異なることがわかる。これはD_1が 1 となり、他のD_iは 0 となることから明らかであろう。

D_iは$i = 1$のとき 1 で、他では 0 となることから、$\hat{\beta}$は$\hat{u}_1$を 0 とするように選ばれる(『入門　実践する統計学』12.4.1 節参照)。$\hat{u}_1 = 0$を用いると、$\hat{V}^{HC1}$の分子は

[3] このケースでは、nが大きくなると、偏差 2 乗和は 1 に収束している。したがって、標準的仮定 2 は満たされない。これはnが大きくなっても、$\hat{\beta}$の分散が 0 にならないことを示唆している。

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

となる。また、$E[\hat{u}_i^2]=(1-h_{ii})\,\sigma^2$であるから、

$$\sum_{i=1}^{n}E[\hat{u}_i^2]=\sum_{i=1}^{n}(1-h_{ii})\sigma^2$$

$$=(1-1)\sigma^2+\sum_{i=2}^{n}\left(1-\frac{1}{n-1}\right)\sigma^2$$

$$=(n-2)\sigma^2$$

となる。これらの結果を用いると、次のように展開できる。

$$E\left[\hat{V}^{HC1}\right]=\frac{n}{n-2}\frac{\sum_{i=1}^{n}(D_i-\bar{D})^2E[\hat{u}_i^2]}{\left(\sum_{i=1}^{n}(D_i-\bar{D})^2\right)^2}$$

$$=\frac{n}{n-2}\frac{\frac{1}{n^2}(n-2)\sigma^2}{\left(\frac{n-1}{n}\right)^2}$$

$$=\frac{n}{(n-1)^2}\sigma^2=\frac{1}{n-1}\left(\frac{n}{n-1}\sigma^2\right)$$

この結果から、たとえば、$n=101$なら、$\hat{V}^{HC1}$は真の分散の 100 分の 1 となってしまうことがわかる。これは、本当は有意でなかったとしても、$\hat{V}^{HC1}$を用いることで､有意であると誤って判断する可能性が高くなることを意味している。この問題は、$n_1=1$とした場合だけでなく、n_1が小さい(1 をとるケースが少ない)、もしくは$n-n_1$が小さい(0 をとるケースが少ない)場合にも生じる。

解決策として、$\hat{V}^{HC1}$の代わりに$\hat{V}^{HC2}$を用いることが挙げられる(ただし、$n_1=1$の場合には、h_{ii}のうち 1 つは 1 となり、標準化残差$\bar{u}_i=(1-h_{ii})^{-1/2}\hat{u}_i$、ひいては$\hat{V}^{HC2}$も計算できない)。1 次的ダミーを用いる場合は、標準誤差を過小評価する問題があることを念頭に置いておこう。