2024/04/12

藪友良

第 4 章の答え

練習問題 1

帰無仮説は棄却することに意味がある。帰無仮説を採択しても、帰無仮説が正しいのか、対立仮説が正しいのか、どちらともいえない。帰無仮説を採択しても、帰無仮説が正しいと、誤って解釈しないように注意が必要である。

ハーバード大学元学長 L・サマーズ(Lawrence Summers)は、「統計学の授業で学生が何度も注意をうけるように、帰無仮説を棄却できないということは帰無仮説の正しさを意味していない」と述べている。これは当然のことだが、勘違いしやすい点なので注意したい。

練習問題 2

第 1 種の過誤は、帰無仮説H_0が正しいとき、帰無仮説H_0を誤って棄却し、対立仮説H_1を採択することをいう。第 2 種の過誤は、対立仮説H_1が正しいとき、帰無仮説H_0を誤って採択することをいう(詳しくは、4.2.2 節参照)。

練習問題 3

有意水準を低く設定する場合として、医薬品の開発やドーピング検査などがある。医薬品開発では、帰無仮説（「医薬品の効果がない」）を誤って棄却して効果のない薬(または有害な薬)を市場に出すデメリットは大きいため有意水準は低く設定される。また、ドーピング検査では、帰無仮説(「ドーピングをしていない」)を誤って棄却して選手のキャリアを傷つけるコストは大きいためやはり有意水準は低く設定される。

逆に、有意水準を低く設定しないものとして、人間ドックなどの簡易検査がある。帰無仮説(「病気にかかっていない（陰性）」)を誤って棄却しても、精密検査を行えば良いだけなので、大きな問題はない。逆に、帰無仮説を誤って採択し、病気を放置してしまうコストは大きい。

練習問題 4

有意水準 1%とした帰無仮説H_0: $\beta = \beta_0$の採択域$(-c < t_{\hat{\beta}} < c)$は、$c = t_{n-2,0.01}$とすることで、次のようになる。

$$-t_{n-2,0.01} < \frac{\hat{\beta} - \beta_0}{s_{\hat{\beta}}} < t_{n-2,0.01}$$

これを書き換えると(両辺に$-s_{\hat{\beta}}$を掛けてから両辺に$\hat{\beta}$を足すと)、次の関係式が得られる。

$$\hat{\beta} - t_{n-2,0.01}s_{\hat{\beta}} < \beta_0 < \hat{\beta} + t_{n-2,0.01}s_{\hat{\beta}}$$

上式左辺はまさに 99%信頼区間の下限$\hat{\beta} - t_{n-2,0.01}s_{\hat{\beta}}$であり、右辺は 99%信頼区間の上限$\hat{\beta} + t_{n-2,0.01}s_{\hat{\beta}}$と一致している。このため、係数$\beta$の99%信頼区間の中に$\beta_0$が含まれる場合には、有意水準 1%で帰無仮説$H_0$は採択され、99%信頼区間の外に$\beta_0$がある場合には、有意水準 1%で帰無仮説$H_0$は棄却される。

練習問題 5

4.4.2 節で学習したとおり、有意性は次のように判断される。

p値≤ 0.01ならば、　有意水準 1%で帰無仮説H_0は棄却される

$0.01 < p$値≤ 0.05 ならば、　有意水準 5%で帰無仮説H_0は棄却される

$0.05 < p$値≤ 0.1 ならば、　有意水準 10%で帰無仮説H_0は棄却される

$0.1 < p$値ならば、　有意水準 10%でも帰無仮説H_0は棄却されない

ここで、p値$= 0.09$であるため、有意水準 10%ならば帰無仮説は棄却されるが、有意水準 5%や 1%では棄却されない。また、p値$= 0.15$なら、有意水準 10%であっても、帰無仮説は棄却されない。

練習問題 6

t 値は$t^* = -0.95$である。t 分布は標準正規分布と同じと仮定すると、

$$p\text{値} = P\{0.95 < |Z|\}$$

となる。ただし、Zは標準正規確率変数となる。

これは標準正規分布表を用いることで計算できる。まず、$P\{Z < 0.95\} = 0.8289$となるため、確率の和は 1 から、

$$P\{Z > 0.95\} = 1 - P\{Z < 0.95\} = 0.1711$$

となる。また、標準正規分布は 0 を中心に左右対称なので、$P\{Z < -0.95\} = 0.1711$となり、よって、p値は次のようになる。

$$P\{0.95 < |Z|\} = P\{Z > 0.95\} + P\{Z < -0.95\} = 0.1711 \times 2 = 0.3422$$

Excel を用いるなら、$P\{Z < -0.95\}$は、「=NORM.S.DIST(-0.95,TRUE) 」と入力すれば計算できる。これは 0.171056 となり、これを 2 倍すると 0.3421 となる。

練習問題 7

第 1 の理由は、推定値と標準誤差がわかれば、簡単に 95%信頼区間を計算できるからである。第 2 の理由は、カッコ内に t 値を掲載することによって、仮説検定の結果(有意性)を強調しすぎてしまうからである(4.6.3 節参照)。

練習問題 8

たとえば、女性ダミーをF_iとしよう(女性ダミーは、女性なら 1 となり、男性なら 0 となるダミー変数である)。このとき、標本平均の分子$\sum_{i=1}^{n} F_i$は、データにおける女性の総数であり、それをサンプルサイズnで割ることで、データにおける女性の割合を求めることができる。つまり、

$$\frac{1}{n}\sum_{i=1}^{n} F_i = \frac{\text{女性の総数}}{\text{サンプルサイズ}} = \text{データにおける女性の割合}$$

練習問題 9

個人iの点数Y_iは、次のようになる。

$$Y_i = \mu_f + (\mu_m - \mu_f)M_i + u_i$$

たとえば、生徒iが男子ならば($M_i = 1$)、

$$Y_i = \mu_f + \left(\mu_m - \mu_f\right) + u_i = \mu_m + u_i$$

となる一方、女子ならば($M_i = 0$)、次のようになる。

$$Y_i = \mu_f + u_i$$

上式は、次の単回帰モデルとして表せる。

$$Y_i = \alpha + \beta M_i + u_i$$

ただし、$\alpha = \mu_f$、$\beta = \mu_m - \mu_f$と定義した。この例から、男女の点数差を推定する

ために、女性ダミーか男性ダミーのいずれかを用いればよいとわかる。ただし、どちらのダミー変数を用いるかで、定数項や係数の解釈が異なる点に注意が必要である。

練習問題 10

数学の点数を見てみよう。係数は 0.422 であり、男性は女性より 0.422 点だけ高い。ここで、$\hat{\beta} = 0.422$、$s_{\hat{\beta}} = 0.298$であり、$t_{\hat{\beta}} = 0.422/0.298 = 1.42$となる。ただし、$p$値は 10%より大きく、有意水準 10%でも帰無仮説$H_0: \beta = 0$は棄却できない(つまり、男女の平均点に有意な差があるとは言えない)。定数項は女性の平均点であり、これは 49.79 点となる。

次に、理科の点数を見てみよう。係数は 1.258 であり、男性は女性より 1.258 点だけ高い。ここで、$\hat{\beta} = 1.258$、$s_{\hat{\beta}} = 0.297$であり、$t_{\hat{\beta}} = 1.258/0.297 = 4.23$となる。$p$値は 1%より小さく、有意水準 1%で帰無仮説$H_0: \beta = 0$は棄却される。定数項は女性の平均点であり、これは 49.37 点となる。

例 4-4 では、説明変数を女性ダミーとしている一方、この問題では、説明変数を男性ダミーとしている。これらの推定結果を比較すると、推定結果は本質的に同じとなっていることがわかるだろう。ただし、定数項や係数の解釈は変わる点に注意が必要である。

練習問題 11

F_iの標本平均は$\bar{F} = \sum_{i=1}^{n} F_i / n = n_1/n$となる(女性は計$n_1$人いるため、$\sum_{i=1}^{n} F_i = n_1$となる)。このため、偏差は、

$$F_i - \bar{F} = \begin{cases} 1 - \dfrac{n_1}{n} = \dfrac{n_2}{n} & \text{if } \ F_i = 1 \\ 0 - \dfrac{n_1}{n} = -\dfrac{n_1}{n} & \text{if } \ F_i = 0 \end{cases}$$

となる(式展開では、$n_2 = n - n_1$を用いた)。したがって、偏差 2 乗和は、

$$\sum_{i=1}^{n} (F_i - \bar{F})^2 = \sum_{i=1}^{n_1} (F_i - \bar{F})^2 + \sum_{i=n_1+1}^{n} (F_i - \bar{F})^2$$

$$= n_1 \left(\frac{n_2}{n}\right)^2 + n_2 \left(\frac{n_1}{n}\right)^2$$

$$= \frac{n_1 n_2}{n^2}(n_1 + n_2) = \frac{n_1 n_2}{n}$$

となる(式展開では、$n = n_1 + n_2$を用いた)。

2 章で学習したとおり、説明変数Xの係数βの OLS 推定量$\hat{\beta}$は、

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

となる。2 番目の等号は、次式を用いた。

$$\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{n}(X_i - \bar{X})Y_i - \bar{Y}\underbrace{\sum_{i=1}^{n}(X_i - \bar{X})}_{\text{偏差の和は }0} = \sum_{i=1}^{n}(X_i - \bar{X})Y_i$$

OLS 推定量$\hat{\beta}$の式に、$X_i = F_i$と$\bar{X} = \bar{F}$を代入すると、次のようになる。

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(F_i - \bar{F})Y_i}{\sum_{i=1}^{n}(F_i - \bar{F})^2}$$

$$= \frac{\sum_{i=1}^{n_1}(F_i - \bar{F})Y_i + \sum_{i=n_1+1}^{n}(F_i - \bar{F})Y_i}{\sum_{i=1}^{n}(F_i - \bar{F})^2}$$

ここで、$F_i = 1$なら$F_i - \bar{F} = \frac{n_2}{n}$、$F_i = 0$なら$F_i - \bar{F} = -\frac{n_1}{n}$となること、また、$\sum_{i=1}^{n}(F_i - \bar{F})^2 = \frac{n_1 n_2}{n}$となることに注意すると、上式右辺は次のようになる。

$$\frac{\frac{n_2}{n}\sum_{i=1}^{n_1} Y_i - \frac{n_1}{n}\sum_{i=n_1+1}^{n} Y_i}{\frac{n_1 n_2}{n}} = \frac{1}{n_1}\sum_{i=1}^{n_1} Y_i - \frac{1}{n_2}\sum_{i=n_1+1}^{n} Y_i = \bar{Y}_f - \bar{Y}_m$$

最後の等式は、$\bar{Y}_f$は女性の標本平均、$\bar{Y}_m$は男性の標本平均であること、つまり、以下の関係式を用いた。

$$\bar{Y}_f = \frac{1}{n_1}\sum_{i=1}^{n_1} Y_i \qquad \bar{Y}_m = \frac{1}{n_2}\sum_{i=n_1+1}^{n} Y_i$$

練習問題 12

ウェブサイトから、再現に必要なデータと再現コードをダウンロードできる。