経済統計学講義ノート No.12

回帰分析 III

蛭川雅之

2023 年 11 月 7 日

目　次

1 回帰係数に関する t 検定

1.1 t 検定：概要

- 重回帰モデル

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon$$

において、特定の係数 β_j $(j = 0, 1, \ldots, k)$ がある値 b であるかどうかを検定する問題を考える。

  - 帰無仮説は

$$H_0 : \beta_j = b$$

である。

- 多くの場合 $b = 0$、即ち帰無仮説は

$$H_0 : \beta_j = 0$$

である。

  - これは

「他の条件が一定の下で、説明変数X_jは効いている
（＝被説明変数Y に対して影響を及ぼしている）かどうか」

に関する検定である。

- しばしば、$b = 1$ も用いられる。
  - 需要の価格弾力性
  - CAPM...

- β_j の最小 2 乗推定値 $\hat{\beta}_j$ およびその標準誤差 $SE\left(\hat{\beta}_j\right)$ から、検定統計量として t 統計量

$$t_j = \frac{\hat{\beta}_j - b}{SE\left(\hat{\beta}_j\right)} \ (j = 0, 1, \ldots, k)$$

が定義される。

  - 標準的な統計パッケージの回帰分析結果では、$b = 0$ の場合の t 値

$$t_j = \frac{\hat{\beta}_j}{SE\left(\hat{\beta}_j\right)} \ (j = 0, 1, \ldots, k)$$

が報告される。

- 帰無仮説 H_0 が正しい場合、$t_j \sim t\,(n-k-1)$（＝自由度 $n-k-1$ の t 分布）に従う。
  - 観測値の個数 n が十分大きい場合は、$t_j \overset{A}{\sim} N\,(0,1)$ として差し支えない。
- 帰無仮説 H_0 と対立仮説 H_1 の組がどのように設定されているかにより、棄却域は異なる。
  - 有意水準を $\alpha\,(=0.05, 0.01\ldots)$ とすると、
    1. $H_1 : \beta_j \neq b$ の場合、棄却域は $t\,(n-k-1)$ の両側 $100\alpha\%$ 点より外側の領域になる（＝両側検定）。
    2. $H_1 : \beta_j < b$ の場合、棄却域は $t\,(n-k-1)$ の左側 $100\alpha\%$ 点より小さい部分の領域になる（＝左側検定）。
    3. $H_1 : \beta_j > b$ の場合、棄却域は $t\,(n-k-1)$ の右側 $100\alpha\%$ 点より大きい部分の領域になる（＝右側検定）。

1.2 P 値

- 標準的な統計パッケージの回帰分析結果では、t 値に加え P 値も報告される。
- P 値は帰無仮説 $H_0 : \beta_j = 0$ が真である場合に

  絶対値が t 値以上の値が観測される確率

  を表す。
- P 値はこの H_0 を両側検定で棄却できる最小の有意水準と解釈することもできる。
  - P 値があらかじめ定めた有意水準を下回っていれば、両側検定で帰無仮説 $H_0 : \beta_j = 0$ を棄却できる。

1.3 信頼区間

- 回帰係数 β_j の最小 2 乗推定値 $\hat{\beta}_j$ およびその標準誤差 $SE\left(\hat{\beta}_j\right)$、さらには $t\left(n-k-1\right)$ の両側 $100\alpha\%$ 点 $t_{\alpha/2}\left(n-k-1\right)$ を用いて、β_j に関する信頼係数 $100\left(1-\alpha\right)\%$ の信頼区間

$$\left[\hat{\beta}_j - t_{\alpha/2}\left(n-k-1\right)SE\left(\hat{\beta}_j\right), \hat{\beta}_j + t_{\alpha/2}\left(n-k-1\right)SE\left(\hat{\beta}_j\right)\right]$$

  を計算できる。
- この信頼区間は帰無仮説 $H_0 : \beta_j = b$ の両側検定に利用可能である。
  - 信頼区間が b を含まない場合、H_0 を有意水準 $100\alpha\%$ で棄却できる。

2 回帰係数の線形結合に関する検定

- コブ＝ダグラス型生産関数

$$Y = AK^{\alpha}L^{\beta} \ (K, L > 0; A, \alpha, \beta > 0)$$

に関する仮説検定を考える。

  - パラメータ (A, α, β) を推定するため、両辺に自然対数をとった回帰式

$$\begin{aligned} \ln Y &= \ln A + \alpha \ln K + \beta \ln L + \epsilon \\ &= \beta_0 + \beta_1 \ln K + \beta_2 \ln L + \epsilon \end{aligned}$$

に書き直す。

- この生産関数が「規模に関する収穫一定」を満たすかどうかを検定したい。
  - 帰無仮説および対立仮説はそれぞれ

$$H_0 : \alpha + \beta = 1 \Rightarrow H_0 : \beta_1 + \beta_2 = 1$$
$$H_1 : \alpha + \beta \neq 1 \Rightarrow H_1 : \beta_1 + \beta_2 \neq 1$$

    となる。
- 回帰式を普通に最小 2 乗推定すれば検定できるか？
  - 検定統計量は

$$t = \frac{\hat{\beta}_1 + \hat{\beta}_2 - 1}{SE\left(\hat{\beta}_1 + \hat{\beta}_2\right)}$$

    である。
  - 標準的な統計パッケージの回帰分析結果を用いて分子は直ちに計算できるが、分母はどうか？

- 分母（＝標準誤差）を詳しく見てみると、

$$SE\left(\hat{\beta}_1+\hat{\beta}_2\right)=\sqrt{\widehat{Var}\left(\hat{\beta}_1+\hat{\beta}_2\right)}$$
$$=\sqrt{\widehat{Var}\left(\hat{\beta}_1\right)+\widehat{Var}\left(\hat{\beta}_2\right)+2\widehat{Cov}\left(\hat{\beta}_1,\hat{\beta}_2\right)}$$

である。

- $\widehat{Cov}\left(\hat{\beta}_1,\hat{\beta}_2\right)$ は最小 2 乗推定値全体の分散共分散行列から入手できる。
  - 標準的な統計パッケージでは分散共分散行列を出力できる。
  - もっとも、ここから標準誤差を計算するのは少々面倒である。

2.1 より手軽な検定方法

- 以下のようなパラメータ表現を変更する方法を考える。
  1. $\theta = \beta_1 + \beta_2$ とおき、帰無仮説および対立仮説をそれぞれ

$$H_0 : \theta = 1,\ H_1 : \theta \neq 1$$

  と書き改める。
  2. $\beta_2 = \theta - \beta_1$ を回帰式へ代入し、

$$\begin{aligned} \ln Y &= \beta_0 + \beta_1 \ln K + (\theta - \beta_1) \ln L + \epsilon \\ &= \beta_0 + \beta_1 (\ln K - \ln L) + \theta \ln L + \epsilon \end{aligned}$$

  を最小 2 乗推定する。

3. θ の最小 2 乗推定値 $\hat{\theta}$ およびその標準誤差 $SE\left(\hat{\theta}\right)$ から検定統計量

$$t_\theta = \frac{\hat{\theta} - 1}{SE\left(\hat{\theta}\right)}$$

を計算し、t 検定を実行する。

問題 1 データファイル“cd2000.csv”には、米国の製造業 473 業種の 2000 年における総付加価値額 ($vadd$；百万米ドル)、実質総資本ストック (cap；百万米ドル)、雇用総数 (emp；千人）が記録されている。

$$(Y, K, L) = (vadd, cap, emp)$$

として、以下のことを実行せよ。

1. コブ＝ダグラス型生産関数を最小 2 乗推定せよ。
2. この推定結果に基づき、規模に関する収穫一定が成り立っているかどうかを有意水準 5% で検定せよ。

3 複数の回帰係数の同時検定

3.1 帰無仮説および対立仮説

- k 個の説明変数を持つ重回帰モデルにおいて、$q\,(\leq k)$ 個の説明変数の傾きが全てゼロという仮説検定を考える。
  - 簡略化のため、最初の q 個の説明変数の傾きが全てゼロという仮説に対する検定を考える。
- 帰無仮説 H_0 および対立仮説 H_1 はそれぞれ

$$H_0 : \beta_1 = \cdots = \beta_q = 0$$
$$H_1 : H_0 \text{は正しくない}$$

となる。

3.2 検定の手順

1. H_0 が正しい場合の制約付き回帰式

$$Y = \beta_0 + \beta_{q+1} X_{q+1} + \cdots + \beta_k X_k + \epsilon$$

を最小 2 乗推定し、残差 2 乗和 RSS_0 を得る。

2. H_1 が正しい場合の制約なし回帰式

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_q X_q + \beta_{q+1} X_{q+1} + \cdots + \beta_k X_k + \epsilon$$

を最小 2 乗推定し、残差 2 乗和 RSS_1 を得る。

3. 検定統計量

$$F = \frac{(RSS_0 - RSS_1)/q}{RSS_1/(n-k-1)}$$

を計算する。

- 右辺の分子・分母を TSS で割り、制約付きおよび制約なし回帰式に対する決定係数 R_0^2 および R_1^2 を使い、検定統計量を

$$F = \frac{\left(R_1^2 - R_0^2\right)/q}{\left(1 - R_1^2\right)/\left(n - k - 1\right)}$$

と表現することもできる。

4. H_0 が正しい場合、検定統計量は自由度 $(q, n-k-1)$ の F 分布（$F_{q,n-k-1}$ と表記する）に従う。
   - $RSS_0 \geq RSS_1 \left(\Leftrightarrow R_1^2 \geq R_0^2\right)$ であるため、$F \geq 0$ である。
   - 直感的には、検定統計量が大きい値をとる場合 H_0 を棄却する。
   - 標準的な統計パッケージには、複数の回帰係数の同時検定（より正確には、回帰係数の線形制約に関する同時検定）を行うコマンドが用意されている。

3.3 応用：回帰式全体の有意性の検定

- k 個の説明変数を持つ重回帰モデル

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon$$

において、説明変数全ての傾きがゼロという仮説検定を考える。

- 帰無仮説 H_0 および対立仮説 H_1 はそれぞれ

$$H_0 : \beta_1 = \cdots = \beta_k = 0$$
$$H_1 : H_0 \text{は正しくない}$$

となる。

- H_0 が正しい場合、切片のみの回帰式 $Y = \beta_0 + \epsilon$ となる。
  - この場合、残差 2 乗和は $RSS_0 = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 = TSS$ である。

- H_1 が正しい場合の回帰式からの残差 2 乗和を $RSS_1 = RSS$ と表記し、検定統計量

$$F = \frac{(RSS_0 - RSS_1)/k}{RSS_1/(n-k-1)} = \frac{(TSS - RSS)/k}{RSS/(n-k-1)} = \frac{ESS/k}{RSS/(n-k-1)}$$

を得る。

  - 検定統計量は、（H_1 が正しい場合の）決定係数 R^2 を用いて

$$F = \frac{R^2/k}{(1-R^2)/(n-k-1)}$$

とも表現できる。

- 標準的な統計パッケージはこの検定統計量の値（F 値）を自動的に出力する。

4 関数型の選択

- 回帰モデルとして、純粋な線形型式のほか半対数型（例：賃金関数）

$$\ln Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon$$

および両対数型（例：コブ＝ダグラス型生産関数）

$$\ln Y = \beta_0 + \beta_1 \ln X_1 + \cdots + \beta_k \ln X_k + \epsilon$$

が広く用いられる。

- 半対数型・両対数型モデルの係数は以下のように解釈される。
  1. 半対数型モデルの場合、

$$\beta_j = \frac{\Delta \ln Y}{\Delta X_j} = \frac{\Delta Y/Y}{\Delta X_j}$$

     - X_j の単位当たり変化に対する Y の伸び率（$=$ % 変化）を表す。
  2. 両対数型モデルの場合、

$$\beta_j = \frac{\Delta \ln Y}{\Delta \ln X_j} = \frac{\Delta Y/Y}{\Delta X_j/X_j}$$

     - X_j の伸び率（$=$ % 変化）対する Y の伸び率（$=$ % 変化）、即ち Y の X_j に対する弾力性を表す。

5 説明変数の選択

5.1 ダミー変数

5.1.1 定数項ダミー

- 性別が賃金に与える影響を考慮した回帰モデル

$$\ln(wage) = \beta_0 + \beta_1 educ + \beta_2 female + \epsilon$$

を考える。

  - ここで、

$$female = \begin{cases} 0 & : \text{男性} \\ 1 & : \text{女性} \end{cases}$$

はダミー変数の一例である。

- 回帰モデルは性別により

$$\ln(wage) = \begin{cases} \beta_0 + \beta_1 educ + \epsilon & : \text{男性} \\ (\beta_0 + \beta_2) + \beta_1 educ + \epsilon & : \text{女性} \end{cases} \tag{1}$$

と表現できる。

  - ダミー変数 $female$ は定数項ダミーの一例である。

- $female$ を

$$male = \begin{cases} 1 & : \text{男性} \\ 0 & : \text{女性} \end{cases}$$

に置き換えたらどうなるか？

  - この場合、

$$\begin{aligned} \ln(wage) &= \alpha_0 + \alpha_1 educ + \alpha_2 male + \epsilon \\ &= \begin{cases} (\alpha_0 + \alpha_2) + \alpha_1 educ + \epsilon & : \text{男性} \\ \alpha_0 + \alpha_1 educ + \epsilon & : \text{女性} \end{cases} \end{aligned} \tag{2}$$

となる。

- （1）と（2）は本質的に同じものであるから、

$$\beta_0 = \alpha_0 + \alpha_2, \beta_1 = \alpha_1, \beta_0 + \beta_2 = \alpha_0$$

即ち

$$\alpha_0 = \beta_0 + \beta_2, \alpha_1 = \beta_1, \alpha_2 = -\beta_2$$

が成り立つ。

- 説明変数として $male$ と $female$ を両方含めてはどうか？
  - $male + female \equiv 1$ となり、説明変数間に完全な多重共線性が成り立つ（$\Rightarrow$「よくある失敗例」）。
  - 区分・選択肢が 3 以上ある場合も含め、区分・選択肢より一つ少ないダミー変数を使用する。

5.1.2 係数ダミー

- 性別が $educ$ の係数も変化させ得ると考え、交差項 $female \cdot educ$ を説明変数に加えた次の回帰モデルに修正する。

$$\ln(wage) = \beta_0 + \beta_1 educ + \beta_2 female + \beta_3 female \cdot educ + \epsilon$$

  - ダミー変数 $female$ は定数項ダミーのみならず、係数ダミーとしても使用されている。

- 実際、回帰モデルは性別により

$$\ln(wage) = \begin{cases} \beta_0 + \beta_1 educ + \epsilon & : \text{男性} \\ (\beta_0 + \beta_2) + (\beta_1 + \beta_3) educ + \epsilon & : \text{女性} \end{cases}$$

と表現できる。

5.2 多項式

- しばしば、ある説明変数のべき乗を新たな説明変数として回帰モデルに加えることがある。
  - 所得・賃金はある年齢で頭打ちになる ⇒ 年齢の 2 乗
  - フィリップス曲線 ⇒ 完全失業率の −1 乗（＝逆数）

問題 2 データファイル“wage2.csv”のデータを使用して以下のことを実行せよ。

1. $Y = \ln(wage)$ を被説明変数とし、**(i)** 多項式 $(X_1, X_2, X_3) = \left(educ, age, age^2\right)$、**(ii)** 定数項ダミー $(X_1, X_2, X_3, X_4) = \left(educ, age, age^2, married\right)$、**(iii)** 係数ダミー $(X_1, X_2, X_3, X_4, X_5) = \left(educ, age, age^2, married, educ \cdot married\right)$ を説明変数とする回帰モデルをそれぞれ最小 2 乗推定せよ。
2. **(i)** を制約付き回帰式、**(iii)** を制約なし回帰式と考え、有意水準 5% で回帰係数に関する同時検定を行え。

■表形式による推定結果報告

被説明変数: $\ln(wage)$

説明変数	(i)	(ii)	(iii)
$educ$	0.0601	0.0618	0.0499
	(0.0059)	(0.0058)	(0.0176)
age	0.0397	0.0411	0.0410
	(0.1047)	(0.1034)	(0.1034)
age^2	−0.0003	−0.0003	−0.0003
	(0.0016)	(0.0016)	(0.0016)
$married$	−	0.2083	0.0243
	(−)	(0.0415)	(0.2602)
$educ \cdot married$	−	−	0.0133
	(−)	(−)	(0.0186)
定数項	4.9421	4.7468	4.9108
	(1.7252)	(1.7037)	(1.7194)
n	935	935	935
R^2	0.1249	0.1479	0.1484
$\bar{R}^2$	0.1221	0.1443	0.1438

5.3 よくある失敗例

- 以下のような場合、説明変数間に完全な多重共線性が成立し、回帰モデルを推定できないので注意せよ。
  1. 区分・選択肢と同数のダミー変数を使用する。
     - 例：

$$male = \begin{cases} 1 & : \text{男性} \\ 0 & : \text{女性} \end{cases}, female = \begin{cases} 0 & : \text{男性} \\ 1 & : \text{女性} \end{cases}$$

  2. 説明変数間に恒等式が成り立つ。
     - 例：

$$exper \equiv age - educ - 6$$

3. 全てのシェアを使用する。
   - 例：毎月の家計支出が食費、家賃、教育費のみから構成されている場合、

$$S_1 = \frac{(\text{食費})}{(\text{家計支出})}, S_2 = \frac{(\text{家賃})}{(\text{家計支出})}, S_3 = \frac{(\text{教育費})}{(\text{家計支出})}$$

     を全て回帰モデルに含める。
4. ダミー変数の多項式を使用する。
   - 例：

$$male = \begin{cases} 1 & : \text{男性} \\ 0 & : \text{女性} \end{cases} \Rightarrow male^2 = \begin{cases} 1 & : \text{男性} \\ 0 & : \text{女性} \end{cases}$$

6 分散不均一性

6.1 分散不均一性とは

- ガウス＝マルコフ定理によれば、一定の条件の下で、線形回帰モデルに対する最小 2 乗推定量は全ての線形不偏推定量の中で最小の分散を持つ。
  - 「誤差項の分散は均一である」という条件も含まれる。
- 分散不均一性（heteroskedasticity）とは、誤差項の（条件付き）分散が一定でない状況を指す。
  - 例えば、単回帰モデル $Y = \beta_0 + \beta_1 X + \epsilon$ において、

$$Var\left(\epsilon \middle| X\right) = E\left(\epsilon^2 \middle| X\right)$$

が X に依存する場合を指す。

6.2 不均一分散の帰結

- 最小 2 乗推定量の不偏性・一致性は保たれる。
- ただし、以下の問題点が生じる。
  1. 最小 2 乗推定量より分散の小さい線形不偏推定量を作ることができる。
  2. 標準的な統計パッケージの最小 2 乗法コマンドで自動計算される標準誤差は適切でない。

6.3 加重最小 2 乗法

- 最小 2 乗推定量より分散の小さい線形不偏推定量とは？
  - 具体的には、加重最小 2 乗法（Weighted Least Squares；略称 “WLS”）を指す。
  - WLS は一般化最小 2 乗法（Generalized Least Squares；略称 “GLS”）の一例である。
- WLS を単回帰モデル $Y = \beta_0 + \beta_1 X + \epsilon$ へ応用する。
  1. ある関数 $g(\cdot)\,(> 0)$ に対し、$Var(\epsilon | X) \propto g(X)$ であるとする。
  2. WLS 推定量は回帰モデル

$$\frac{Y}{\sqrt{g(X)}} = \beta_0 \left(\frac{1}{\sqrt{g(X)}} \right) + \beta_1 \left(\frac{X}{\sqrt{g(X)}} \right) + \frac{\epsilon}{\sqrt{g(X)}}$$

を最小 2 乗推定して得られる。

- $g(\cdot)$ が既知であることを前提とするため、WLS はあまり使われない。
  - $g(\cdot)$ が未知の場合には、この関数も推定する必要が生じる
  - これは実行可能な GLS（Feasible GLS；略称“FGLS”）の問題である。
  - $g(\cdot)$ の推定には以下の懸念がある。
    1. $g(\cdot)$ の定式化は正しいか？
    2.（正しく定式化されていたとしても）$g(\cdot)$ の推定誤差が WLS 推定量の精度に与える影響を無視できるか？

6.4 標準誤差の修正

- 分散不均一性は最小 2 乗推定量の不偏性・一致性に影響しない。
  - 標準誤差の計算方法を変更すれば十分ではないか？
- Eicker-Huber-White の公式によるロバスト標準誤差はこの発想に基づいて提案されている。
  - Eicker-Huber-White の公式は均一分散・不均一分散いずれの場合にも一致性を持つ。
  - Stata では、最小 2 乗法のコマンド“`regress`”上でオプション“`robust`”を用いることによりロバスト標準誤差が自動的に計算される。

問題 3 不均一分散を前提として問題 1・2 を再度実行せよ。