回帰分析
重回帰(3)
分散不均一性，多重共線性

内容

- 分散不均一性
  - 分散不均一性の問題点
  - 分散不均一性の検出
  - Heteroskedsticity robust estimator
  - 加重最小二乗法 (Weighted Least Square)
- 誤差項の系列相関
- 多重共線性
- 説明変数の誤差→詳細は「操作変数法」

回帰分析の前提

- $u_i \sim N(0, \sigma^2) \quad i.i.d$
  - 誤差項の期待値は0
  - 誤差項は互いに独立（系列相関は無い）
  - 誤差項の分散は一定（分散均一性）
  - 誤差項は正規分布（t検定，F検定のための前提）
- 説明変数と誤差項は独立
- 説明変数の行列Xはfull rank
  - 説明変数間の多重共線性は存在しない

分散不均一性 (1)

- 分散不均一性　heteroskedasticity
  - 誤差項の分散が一定でないこと
  - 誤差項の分散が一定などの前提 → 最小二乗推定量の確率分布 →t検定やF検定の正しさを保証
- 最小二乗推定量の分布(単回帰の場合)

  最小二乗推定量は次のように求められた(講義資料「回帰分析（単回帰）」reg.pdfを参照せよ）

$$b = \beta + \frac{\sum_i (x_i - \bar{x}) u_i}{S_{xx}}$$

この式から期待値と分散を求めると，$\mathrm{var}(u_i) = \sigma^2$のもとで

$$\mathrm{E}(b) = \beta$$

$$\mathrm{var}(b) = \frac{\sigma^2}{S_{xx}}$$

が導かれた。そして，これから$(b - \beta)/\mathrm{s.e.}(b)$がt分布をすることが導かれた。

分散不均一性 (2)

- 分散不均一性の存在　$\mathrm{var}(u_i) = \sigma_i^2$

$b = \beta + \frac{\sum_i (x_i - \bar{x}) u_i}{S_{xx}}$　の期待値と分散を求めると

ただし，次の性質は成り立つとする

- xと誤差項の独立性
- 誤差項に系列相関は無い

$$\mathrm{E}(b) = \beta + \frac{\sum_i (x_i - \bar{x})\mathrm{E}(u_i)}{S_{xx}} = \beta$$

$$\mathrm{var}(b) = E[(b-\beta)^2] = \frac{1}{{S_{xx}}^2} \sum_i (x_i - \bar{x})^2 {\sigma_i}^2$$

となり，推定量の不偏性(期待値が真の値に一致すること）は成り立つが，分散については，$\mathrm{var}(b) = \sigma^2 / S_{xx}$は成り立たない

- bの確率分布の想定が異なるので，t検定，F検定を今までのように使うことはできない

分散不均一性 (3)

- 分散不均一性の例
  - 誤差項の分散がある変数の関数になっている
    - 例）賃金方程式で，高学歴者ほど賃金のバラつきが大きい
- 分散不均一性の検出
  - 説明変数と残差の散布図でチェックする
  - 被説明変数の推定値(=説明変数の1次関数）と残差の散布図でチェックする
  - 分散不均一性のテスト
    - Breusch and Pagan のテスト
    - Whiteのテスト

分散不均一性の検出

- 残差の平方を被説明変数，元の回帰式の説明変数（またはyの予測値）に回帰させ，説明力があるかどうか調べる
  - 説明変数に説明力は無い→ 分散不均一性は検出されなかった
  - 説明変数に説明力がある→ 分散不均一性が存在

なぜ残差の平方か

最小二乗法→ 残差と説明変数は直交する（相関がない）

- 単回帰，重回帰の理論の解説を参照せよ

→残差を，説明変数（またはyの予測値）に回帰してもその係数はゼロ

→残差の平方と，説明変数（またはyの予測値） の間にシステマティックな関係があるかどうかを調べる

分散不均一性の検出(2)

Breusch and Paganのテスト

estimate: $y_i = \alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i} + u_i$
save: $e_i = y_i - a - b_1 x_{1,i} - b_2 x_{2,i} - \cdots - b_k x_{k,i}$
compute: ${e_i}^2$
estimate: ${e_i}^2 = \delta_0 + \delta_1 x_{1,i} + \delta_2 x_{2,i} + \cdots + \delta_k x_{k,i} + v_i$
test H_0 : $\delta_1 = \delta_2 = \cdots = \delta_k = 0$

H_0の検定：F検定で

$$\frac{(RSS - TSS)/k}{RSS/(n-(k+1))} = \frac{ESS/k}{RSS/(n-(k+1))} \sim F(k, n-(k+1))$$

分散不均一性の検出(3)

Whiteのテスト

- 残差の平方 e^2 を被説明変数
- 説明変数：x_j，x_jの平方，x_jとx_hの交差項
- これらの説明変数の係数が全て0という仮説を検定する
- 簡便な方法
  - yの予測値（説明変数の線形関数），およびその平方を説明変数に加える

Rでの分散不均一性

wage1.lm <- lm(wage ~ educ + exper + tenure)
残差はresid(wage1.lm)で取り出せる
残差の平方を被説明変数として回帰

```
res <- resid(wage1.lm)    # 残差をresに代入
res2 <- res^2             # 残差の平方をres2に代入
wage1_bp.lm <- lm(res2 ~ educ + exper + tenure)
summary(wage1_bp.lm)
```

説明変数の係数が全て0であるという仮説は棄却
→分散不均一性の存在

結果

```
---(途中省略)---
F-statistic: 15.53 on 3 and 522 DF,  p-value: 1.09e-09
```

パーケージ　lmtest の bptest( ) という関数を用いる方法もあり

分散不均一性の検定（まとめ）

- R

  残差は　resid(オブジェクト名）で取り出せる

  ```
  res <- resid(wage1.lm)　# wage1.lm に回帰分析の結果
  res2 <- res^2
  ```

  として，res2を被説明変数として元の回帰式の説明変数に回帰させ，全ての変数の係数=0の検定を行う

- bptest( )を用いる
  - パッケージlmtestが必要
- plot( )を用いて，残差のチェック
  - 回帰分析の結果がwage1.lm に保存されている場合，plot(wage1.lm)とタイプすると，残差のチェックのためのグラフが出力される
    - 被説明変数の予測値と残差の散布図
    - 残差が正規分布にしたがっているかをみるためのグラフ　など
    - 統計的な検定とは言えないが，回帰式が妥当かどうか簡単にチェックできる

問題1

- データセット：wage1.xls (wage1.raw)
- 回帰式
  - 被説明変数： wage
  - 説明変数： educ, exper, tenure, female

1. 分散不均一性のテストを(Breusch and Paganのテストを行え。
2. 被説明変数をlwage（wageの対数値）に変えて回帰分析を行い，分散不均一性が検出されるかどうか確かめよ。

問題2

- データセット：hprice1.xls (hprice.raw)
  - 住宅価格と住宅の属性についてのデータ

1. 次のモデルを推計し，分散不均一性のテストを行え
   - 被説明変数：price(住宅価格)
   - 説明変数：lotsize, sqrft, bdrms
2. 上のモデルを対数形で推計し，分散不均一性のテストを行え
   - 被説明変数： log(price)
   - 説明変数：log(lotsize), log(sqrft), log(bdrms)

Heteroskedasticity Consistent Estimator

- 分散不均一性
  - 係数の推定値は不偏性をもつ
  - しかし，分散の推定値は正しくない →係数の信頼区間， t検定，F検定は正しくない
- 漸近的に正しい統計量

  サンプルサイズが十分に大きいときに一致性を持つ（推定量が真の値に近づいていくという性質）

  Heteroskedasticity robust estimator（頑健な推定量）とも言われる

  OLSの残差をeとして，次のように計算（単回帰の場合）

$$\mathrm{var}(b) = \frac{\sum_i (x_i - \bar{x})^2 e_i^2}{{S_{xx}}^2}$$

- 係数のHC estimator とOLS estimator
  - 係数自体は同じ
  - t値，標準誤差が異なる

RでのHC estimator

- vcov(object名)
  - 回帰分析の係数の分散共分散行列
- vcovHC(object名)
  - OLSの残差をもとに係数の分散共分散行列を修正
- パッケージ sandwitch が必要
  - library(sandwich)
- 回帰分析の結果-->wage1.lm
  - vcov(wage1.lm)で通常の分散共分散行列,
  - vcovHC(wage1.lm)で誤差項の分散不均一性を考慮した分散共分散行列

RでのHC estimator （2）

- パッケージ lmtest が必要
  - library(lmtest)
- OLSの結果をobjectとして保存（例えばwage1.lm）
- coeftest(wage1.lm)
  - 係数の推定値，標準誤差，t値，p値などが出力される
- coeftest(wage1.lm, vcov=vcovHC)
  - 分散不均一性を考慮して，標準誤差，t値，p値が修正された結果が出力される
- 係数の推定値自体は，分散不均一性があっても変わらない（不偏性を持つ）ことに注意
- 複数の制約がある場合はwaldtest(制約なしモデル，制約付きモデル)を用いる

加重最小二乗法　Weighted Least Square

- 不均一性のテストは検出のみ
  - どのような方法で対処すべきかは教えてくれない
  - 推定する方程式の関数型を変えることで解決する場合もある
- 誤差項の分散がある変数に比例していることがわかっている場合（実際にはほとんど無いが）

  →Weighted Least Square 加重最小二乗法
- WLS : 次の式を最小化するように係数を決定

$$\sum_{i=1}^{n} w_i \left(y_i - a - b_1 x_{1,i} - \cdots b_k x_{k,i}\right)^2$$

w_i ： weight

Weighted Least Square

次のモデルを考える

$$y_i = \alpha + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i} + u_i$$

$$\mathrm{var}(u_i) = \sigma_i^2 = h_i \sigma^2$$

- 誤差項の分散が変数hに比例→ 分散不均一性

このとき次のように式変換すれば

$$\frac{y_i}{\sqrt{h_i}} = \alpha \frac{1}{\sqrt{h_i}} + \beta_1 \frac{x_{1,i}}{\sqrt{h_i}} + \cdots + \beta_k \frac{x_{k,i}}{\sqrt{h_i}} + \frac{u_i}{\sqrt{h_i}}$$

$$\mathrm{var}\left(\frac{u_i}{\sqrt{h_i}}\right) = \sigma^2 \rightarrow \text{分散は均一}$$

Weighted Least Square (2)

- 前ページの2番めの式をもとに係数を推定

→次の式の最小化

$$\sum_{i=1}^{n}\left[\frac{y_i}{\sqrt{h_i}}-a\frac{1}{\sqrt{h_i}}-b_1\frac{x_{1,i}}{\sqrt{h_i}}-\cdots-b_k\frac{x_{k,i}}{\sqrt{h_i}}\right]^2$$
$$=\sum_{i=1}^{n}\frac{1}{h_i}\left[y_i-a-b_1x_{1,i}-\cdots-b_kx_{k,i}\right]^2$$
$$=\sum_{i=1}^{n}w_i\left(y_i-a-b_1x_{1,i}-\cdots b_kx_{k,i}\right)^2$$

- 元のモデルの誤差項の分散がhに比例する→weight変数を$1/h$にする

RでのWLS

- wls：　lm(y~x1+x2+x3,weights=w)　で実行(w=1/h)
- 例）賃金方程式で誤差項の分散が教育年数(educ)に比例する場合，weights=1/educ とし

```
wage.wls <- lm(lwage ~educ + expre + tenure, weights=1/educ, subset=(educ>0))
```

summary(wage.wls)で結果を取り出す

注意）weight変数に0があるとエラーが出ます(自動的に除外してくれない）。lm( ) 関数のoptionでsubset=( )を指定すると，(　)内の条件を満たすようなデータについてのみの回帰を行うことができる

誤差項の系列相関

- 回帰分析の前提：誤差項は互いに独立
- 誤差項に系列相関
  - 誤差項u_iとu_jに相関がある（独立でない）
  - 例）ある年に景気が良いと，次の年の景気も良い
- 誤差項に系列相関があると回帰係数bの分散が$\sigma^2(X'X)^{-1}$にならない
  - t検定，F検定が正しくない
- クロスセクションデータ
  - 通常，オブザベーションの並びに意味は無いので，誤差項の系列相関は問題にならない
  - オブザベーションの並び方が，隣接した地域や人の順番になっている場合には意味がある場合あり。
  - 時系列データの場合には系列相関の問題は無視できない

Durbin Watson検定

1階の系列相関を調べる検定

$$DW = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} {e_t}^2} = \frac{\sum_{t=2}^{T} {e_t}^2 + \sum_{t=1}^{T-1} {e_t}^2 - 2\sum_{t=1}^{T-1} e_t e_{t-1}}{\sum_{t=1}^{T} {e_t}^2} \cong 2(1-\rho)$$

- DW比は多くの統計パッケージでは自動的に出力される
- Stata
  - 時系列データのみ　コマンドラインで　estat dwatson
- Rではdwtest( )関数を用いる（パッケージlmtestが必要）
- 経済データでは，ρ>0のケースが普通　（ρは1階の相関係数）
- 大雑把なルールではDW比が1に近いと系列相関あり
- 現在では，誤差項はもっと一般的にAR(p)過程に従うとして推計。また，時系列データの分析では，変数が定常過程か非定常過程かの区別が重要

多重共線性 multicolinearity

- 説明変数間の相関が高いこと
  - 回帰分析において，個々の変数の影響を分離して推計することができなくなる
  - 単相間だけで判断してはいけない
  - 変数間の単相間は低くても，ある説明変数が別の複数の説明変数の線形結合でかなり説明できる場合もある
- 多重共線性が存在すると
  - 回帰式全体では当てはまりが良いが，個々の説明変数の係数が有意でない（s.e.が大きい）という現象が生じる
  - 実験データ→個々の変数の影響が十分に分離できるように実験計画を立てる
  - 経済データ→上のようなことは不可能→分析方法の再検討

多重共線性の検出

OLSにおいて説明変数x_jの係数の分散は$\mathrm{var}(b_j) = \sigma^2 / S_{xx}^j$で与えられた。$S_{xx}^j$ は説明変数 x_j 「固有」の平方和で

$$S_{xx}^j = S_{jj}(1 - R_j{}^2)$$

で与えられる。ここで

S_{jj} は説明変数x_jの平均値の回りの平方和

$R_j{}^2$: 説明変数x_jを他の説明変数に回帰した場合のR^2（決定係数）

である。このような意味でS_{xx}^j は説明変数 x_j 「固有」の平方和である（これについての正確な議論は行列の知識が必要で難しい）

- 多重共線性が存在すると

  x_j が他の説明変数で説明される

  →$R_j{}^2$が高い

  →b_jの分散が大きくなる

- 結局，回帰式全体としては当てはまりがよくても，個々の係数の標準誤差が大きくなる

多重共線性の検出(2)

- 多重共線性の検出には，VIFという指標を用いるのが便利
- VIF (Variance inflation factor　分散増幅因子)

$$VIF(b_j) = \frac{1}{1 - R_j^2}$$

- VIF の意味は次の式から明らか

$$\mathrm{var}(b_j) = \sigma^2 / S_{xx}^j = \sigma^2 / \left[S_{jj} \left(1 - {R_j}^2 \right) \right]$$

- R : vif(回帰分析のobjec名）で出力される。パッケージcarが必要。

多重共線性の例

- 地方政府の支出活動
  - 説明変数
    - 地域の財政状況（債務残高，税収，国からの補助金，交付税額）
    - 地域の属性（山間地，豪雪地帯,..）
    - 所得，面積等
  - 国からの補助金は，その地域属性によって決まる
    - 所得が低い，中山間地，….
  - → 財政状況と地域属性の間の多重共線性
  - 個々の変数の効果が捉えられない
- MLBプレイヤーの年俸の決定要因の分析
  - HR数と打点数に非常に強い単相間
  - HR数の効果と打点の効果を分離できない

説明変数の誤差

真のモデル

$$y_i = \alpha + \beta x_i^* + u_i$$

説明変数x_i*は観察できない：そのかわりx_iが観察できる

$$x_i = x_i^* + v_i$$
$$\mathrm{E}(v_i) = 0, \mathrm{cov}(u_i, v_j) = 0 \quad \text{for all } i \neq j$$

このとき

$$y_i = \alpha + \beta(x_i - v_i) + u_i = \alpha + \beta x_i + (u_i - \beta v_i)$$
$$\equiv \alpha + \beta x_i + w_i$$

となり，誤差項w_iの期待値は0，分散は一定だが，w_iとx_iには相関があり，OLSの前提は満たされない

この問題は「操作変数法」で改めて説明