回帰分析I

15.　不均一分散

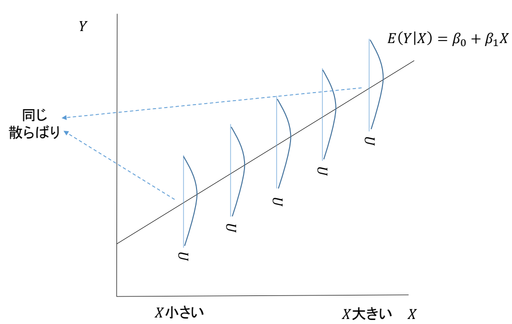
イントロダクション

- この講義ノートで扱うトピックは「誤差項の分散」です。
- これまで誤差項の分散について、「均一分散」仮定を置いてきました（講義ノート9、10参照）。
- すなわち、説明変数がいかなる値でも、誤差項の分散は一定：

$$VAR(U|X_1, \dots, X_k) = \sigma^2$$

どういう意味？

均一分散というのは、ざっくり言うと、
説明変数（X）の値に関わらず、誤差項の散らばりが同じということ



- この均一分散の仮定は、最小二乗推定量の不偏性や一致性には必要ありませんでした。
- それでは、この仮定は何のためにしたのでしょうか？
- この仮定は最小二乗推定量$\hat{\beta}_j$の分散を導出する際に使いました。
- 話を簡単にするために単純回帰モデルを考えます。以下は講義ノート9の復習です。

$$Y = \beta_0 + \beta_1 X + U$$

- 最小二乗推定量$\hat{\beta}_1$は

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{N}(X_i - \bar{X})U_i}{\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

と表すことができます。

- これを使って最小二乗推定量$\hat{\beta}_1$の分散を導出します。

$$VAR(\hat{\beta}_1|X_1,\dots,X_N)$$

$$= VAR\left(\beta_1 + \frac{\sum_{i=1}^N (X_i-\bar{X})U_i}{\sum_{i=1}^N (X_i-\bar{X})^2}\middle|X_1,\dots,X_N\right)$$

$$= \left(\frac{1}{\sum_{i=1}^N (X_i-\bar{X})^2}\right)^2 VAR\left(\sum_{i=1}^N (X_i-\bar{X})U_i\,|X_1,\dots,X_N\right)$$

$$= \left(\frac{1}{\sum_{i=1}^N (X_i-\bar{X})^2}\right)^2 \left(\sum_{i=1}^N (X_i-\bar{X})^2 \boxed{VAR(U_i|X_1,\dots,X_N)}\right)$$

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

ここで誤差項の均一分散の仮定を使いました

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

$$\boxed{= \frac{\sigma^2}{\sum_{i=1}^N (X_i-\bar{X})^2}}$$

ちなみに、これは中間試験の問題になってました

5

$$VAR(\hat{\beta}_1|X_1,\dots,X_N)$$

$$= VAR\left(\beta_1 + \frac{\sum_{i=1}^N (X_i-\bar{X})U_i}{\sum_{i=1}^N (X_i-\bar{X})^2}\middle|X_1,\dots,X_N\right)$$

$$= \left(\frac{1}{\sum_{i=1}^N (X_i-\bar{X})^2}\right)^2 VAR\left(\sum_{i=1}^N (X_i-\bar{X})U_i\,|X_1,\dots,X_N\right)$$

$$= \left(\frac{1}{\sum_{i=1}^N (X_i-\bar{X})^2}\right)^2 \left(\sum_{i=1}^N (X_i-\bar{X})^2 \boxed{VAR(U_i|X_1,\dots,X_N)}\right)$$

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

ここで誤差項の均一分散の仮定を使いました

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

でも、もし誤差項が
均一分散ではなかったら？

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

6

$$VAR(\hat{\beta}_1|X_1,\dots,X_N)$$

$$= VAR\left(\beta_1 + \frac{\sum_{i=1}^N (X_i-\bar{X})U_i}{\sum_{i=1}^N (X_i-\bar{X})^2}\middle|X_1,\dots,X_N\right)$$

$$= \left(\frac{1}{\sum_{i=1}^N (X_i-\bar{X})^2}\right)^2 VAR\left(\sum_{i=1}^N (X_i-\bar{X})U_i\,|X_1,\dots,X_N\right)$$

$$= \left(\frac{1}{\sum_{i=1}^N (X_i-\bar{X})^2}\right)^2 \left(\sum_{i=1}^N (X_i-\bar{X})^2 \boxed{VAR(U_i|X_1,\dots,X_N)}\right)$$

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

ここで誤差項の均一分散の仮定を使いました

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

でも、もし誤差項が
均一分散ではなかったら？

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

最小二乗推定量$\hat{\beta}_1$の分散は
$\sigma^2/\sum_{i=1}^N (X_i-\bar{X})^2$ではない！

均一分散の仮定が正しくなかったら？

- もし均一分散の仮定が正しくなかったら、

$$VAR(\hat{\beta}_1|X_1,\dots,X_N) = \frac{\sigma^2}{\sum_{i=1}^N (X_i-\bar{X})^2}$$

またこれに平方根をとった

$$SD(\hat{\beta}_1|X_1,\dots,X_N) = \frac{\sigma}{\sqrt{\sum_{i=1}^N (X_i-\bar{X})^2}}$$

は正しくないということが分かりました。

- では、これの何が問題なのでしょうか？

8

- $SD\left(\hat{\beta}_1 \middle| X_1, \ldots, X_N\right) = \frac{\sigma}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2}}$ この結果、何に使いましたっけ？

- SDのσを（残差を使って推定した）$\hat{\sigma}$で置き換えたものが、最小二乗推定量$\hat{\beta}_1$の標準誤差です。

$$SE\left(\hat{\beta}_1 \middle| X_1, \ldots, X_N\right) = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2}}$$

- これどこで使いましたっけ？

- t値ですね。

$$t = \frac{\hat{\beta}_1}{SE\left(\hat{\beta}_1 \middle| X_1, \ldots, X_N\right)}$$

Source	SS	df	MS
Model	359065.326	1	359065.326
Residual	750369.903999998		.750371404
Total	1109435.23999999		1.10943634

Number of obs	= 1000000
F(1,999998)	= .
Prob > F	= 0.0000
R-squared	= 0.3236
Adj R-squared	= 0.3236
Root MSE	= .86624

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x	.5992904	.0008663	691.75	0.000	.5975924	.6009884
_cons	2.503237	.0027384	914.13	0.000	2.49787	2.508605

- $VAR\left(\hat{\beta}_1 \middle| X_1, \ldots, X_N\right) = \frac{\sigma^2}{\sum_{i=1}^{N}(X_i - \bar{X})^2}$が正しくないなら、それに基づいて計算される最小二乗推定量$\hat{\beta}_1$の標準誤差$SE\left(\hat{\beta}_1 \middle| X_1, \ldots, X_N\right) = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2}}$も正しくない。

- すると、$t = \frac{\hat{\beta}_1}{SE\left(\hat{\beta}_1 \middle| X_1, \ldots, X_N\right)}$も正しい値ではない。

- なのでp値も正しくない → 仮説検定の結果は信用できない！

- 標準誤差が正しくなければ、信頼区間も正しくない！ですね。

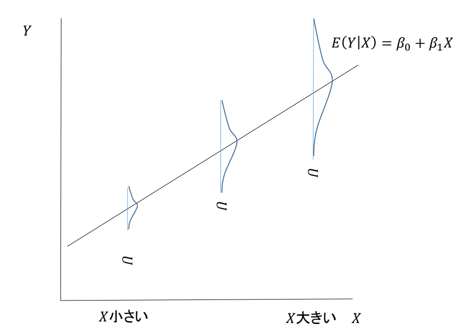
不均一分散

- 均一分散は、誤差項の条件付分散がXの値に依存しないで一定。

- それ以外の場合は、誤差項は「不均一分散」となります。

- 次のページの図は、不均一分散の例。

- ここでは、Xの値が大きいときに、誤差項の散らばりが大きくなるとして図を描いています。



不均一分散の具体例

- 具体例を使って、不均一分散が何を意味するかを説明します。

- 次のモデルを考えましょう。

$$WAGE_i = \beta_0 + \beta_1 MALE_i + U_i$$

- このモデルは

$$WAGE_i = \beta_0 + U_i \quad（女性）$$
$$WAGE_i = \beta_0 + \beta_1 + U_i \quad（男性）$$

ということですね。

- 女性のとき：U_iは、女性iの賃金が、母集団の女性の平均賃金β_0とどれだけ異なるのかを捉える。

- 男性のとき：U_iは、男性iの賃金が、母集団の男性の平均賃金$\beta_0 + \beta_1$とどれだけれだけ異なるのかを捉える。

13

- 従って、、、

- 「U_iの分散が$MALE_i$に依存しない」とは、「賃金の分散が男性と女性で等しい」ということと同じです。

- 言い換えると、この例では、母集団における賃金の分散が、男女で等しければ誤差項は均一分散。

- 等しくなければ誤差項は不均一分散ということになります。

14

均一分散と不均一分散はどちらがより現実的？

- 賃金のモデルについて考えてみましょう。

- 過去のことを言えば、トップクラスの高給の仕事についていたのは、男性がほとんどで女性は少なかった。

- 程度の差はあれ、現在でもある程度はそうと言えるでしょう。

- 一方で、低賃金の仕事に就く男性はいつの時代にも存在します。

- 従って、女性の賃金式の誤差項の分散の方が、男性の賃金式の誤差項の分散よりおそらく小さいと考えられます。

- 従って、不均一分散であると考えるのが妥当でしょう。

15

- 一般的なことをいうと、誤差項は均一分散に従うはず、なんてことを示唆する理論は普通はありません。

  - ➢ 誤差項が均一分散に従うかどうかは、実際のデータしだいですね。

- そのため、最近の実証分析では、「誤差項は不均一分散している」ことを前提にするのが普通です。

- 前提にして何？

- 問題なのは、「誤差項は均一分散している」という仮定に基づいて計算される最小二乗推定量の標準誤差でした。

- よって、「誤差項が不均一分散しているかもしれない」ことを前提にし、それを考慮に入れて最小二乗推定量の標準誤差を計算するのが最近の実証分析の常識となっています。

- 以下では、その方法について説明します。

16

不均一分散に対して頑健な標準誤差

- ここでは話を簡単にするために単純回帰モデルを考えます。

$$Y_i = \beta_0 + \beta_1 X_i + U_i$$

- 添え字iは以下のことを強調するためにつけました。

- 誤差項は不均一分散を仮定します。

$$\boxed{VAR(U_i|X_i) = \sigma_i^2}$$

- ここのポイントは、σ^2についている添え字のiです。

- これは、誤差項の分散が、X_iの値に依存し一定ではないことを意味します。

17

- 最小二乗推定量$\hat{\beta}_1$の分散は

$VAR\left(\hat{\beta}_1\middle|X_1,\dots,X_N\right)$

$= VAR\left(\beta_1 + \frac{\sum_{i=1}^N (X_i-\bar{X})U_i}{\sum_{i=1}^N (X_i-\bar{X})^2}\middle|X_1,\dots,X_N\right)$

$= \frac{1}{\left[\sum_{i=1}^N (X_i-\bar{X})^2\right]^2} VAR\left(\sum_{i=1}^N (X_i-\bar{X})U_i\,\middle|X_1,\dots,X_N\right)$

$= \frac{1}{\left[\sum_{i=1}^N (X_i-\bar{X})^2\right]^2}\left(\sum_{i=1}^N (X_i-\bar{X})^2 VAR(U_i|X_1,\dots,X_N)\right)$

ここまでは均一分散のときと同じです。

18

- 最小二乗推定量$\hat{\beta}_1$の分散は

$VAR\left(\hat{\beta}_1\middle|X_1,\dots,X_N\right)$

$= VAR\left(\beta_1 + \frac{\sum_{i=1}^N (X_i-\bar{X})U_i}{\sum_{i=1}^N (X_i-\bar{X})^2}\middle|X_1,\dots,X_N\right)$

$= \frac{1}{\left[\sum_{i=1}^N (X_i-\bar{X})^2\right]^2} VAR\left(\sum_{i=1}^N (X_i-\bar{X})U_i\,\middle|X_1,\dots,X_N\right)$

$= \frac{1}{\left[\sum_{i=1}^N (X_i-\bar{X})^2\right]^2}\left(\sum_{i=1}^N (X_i-\bar{X})^2 \boxed{VAR(U_i|X_1,\dots,X_N)}\right)$

ここが均一分散の場合とは違う点です。

$= \frac{1}{\left[\sum_{i=1}^N (X_i-\bar{X})^2\right]^2}\left(\sum_{i=1}^N (X_i-\bar{X})^2 \sigma_i^2\right)$

19

- すなわち不均一分散 − $VAR(U_i|X_i) = \sigma_i^2$ − の仮定の下では、最小二乗推定量$\hat{\beta}_1$の分散は

$$\boxed{VAR\left(\hat{\beta}_1\middle|X_1,\dots,X_N\right) = \frac{\sum_{i=1}^N (X_i-\bar{X})^2\sigma_i^2}{\left[\sum_{i=1}^N (X_i-\bar{X})^2\right]^2}}$$

- ちなみに均一分散の仮定の下では、

$$VAR\left(\hat{\beta}_1\middle|X_1,\dots,X_N\right) = \frac{\sigma^2}{\sum_{i=1}^N (X_i-\bar{X})^2}$$

- 不均一分散の時は、σ_i^2が定数ではないので、$\sum$の前に出ない。

- これにより、均一分散と不均一分散のときの最小二乗推定量$\hat{\beta}_1$の分散が一般的には異なることになります。

- 従って、もし誤差項が不均一分散なら、誤差項が均一分散の仮定の下で導出された分散の式は正しくありません。

20

$$VAR(\hat{\beta}_1|X_1,\dots,X_N)=\frac{\sum_{i=1}^N (X_i-\bar{X})^2\sigma_i^2}{\left[\sum_{i=1}^N (X_i-\bar{X})^2\right]^2}$$

- この分散の推定量として、次の式が使われます：

残差の二乗

$$\widehat{VAR}(\hat{\beta}_1|X_1,\dots,X_N)=\frac{\sum_{i=1}^N (X_i-\bar{X})^2\hat{U}_i^2}{\left[\sum_{i=1}^N (X_i-\bar{X})^2\right]^2}$$

- この推定量はWhite（1980）が開発したもので、サンプルサイズが十分に大きい場合、いかなる形の不均一分散にも妥当な推定量であることが証明されています。

White, H. (1980) "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," Econometrica, 48, 817-838.

- これに平方根をとったものを、「**不均一分散に対して頑健（ロバスト）な標準誤差**」と呼びます。単に「**ロバスト標準誤差**」と呼ぶこともあります。

$$Robust\ SE(\hat{\beta}_1|X_1,\dots,X_N)=\sqrt{\widehat{VAR}(\hat{\beta}_1|X_1,\dots,X_N)}$$

- 誤差項の均一分散は、誤差項の不均一分散の特殊ケース。
- 従って、たとえ誤差項が均一分散であったとしても、サンプルサイズが十分に大きい場合、ロバスト標準誤差は妥当な推定量です。
- **不均一分散に対してロバストなt**検定は、この標準誤差を用いて行われます。
- 例えば、$H_0: \beta_1=\alpha, H_1: \beta_1\neq\alpha$なら、$t$統計量は

$$t=\frac{\hat{\beta}_1-\alpha}{Robust\ SE}$$

- これを「**不均一分散に対してロバストなt統計量**」と呼びます。
- 検定の仕方はこれまでと全く同じです。

自分で実証論文を書くとき

- たとえ誤差項が均一分散であったとしても、サンプルサイズが十分に大きい場合、ロバスト標準誤差は妥当な推定量。
- そのため最近の実証分析では、ロバスト標準誤差、そしてそれに基づいた検定結果を報告するのが普通です。
- ほとんどの論文で普通の標準誤差は報告されていません。
- 普通の標準誤差「のみ」を報告すると、論文の読み手が推定結果を全く信用しないことすらあります。
- 従って、皆さんが実証論文を書く場合は、ロバスト標準誤差を使うのが無難だと思います。

使って見ましょう

- データ：hprice1_practice.dta
  - Wooldridge (2006)　Introductory Econometrics: A Modern Approach, Thomson, South-Western からのデータで一部改変したもの。
- 変数：price（家の価格、単位$1000）
  lotsize（土地面積、単位square feet）
  sqrft（建物面積、単位square feet）

```
. reg pricemy lotsizetb sqrfttb

      Source |       SS           df       MS      Number of obs   =        88
-------------+----------------------------------   F(2, 85)        =     83.67
       Model |    60866850         2    30433425   Prob > F        =    0.0000
    Residual |  30918600.8        85  363748.245   R-squared       =    0.6631
-------------+----------------------------------   Adj R-squared   =    0.6552
       Total |  91785450.8        87  1055005.18   Root MSE        =    603.12

------------------------------------------------------------------------------
     pricemy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   lotsizetb |   .7827759   .2394666     3.27   0.002     .3066522      1.2589
     sqrfttb |   49.39334   4.221075    11.70   0.000     41.00071    57.78597
       _cons |   59.32415   235.1236     0.25   0.801    -408.1645    526.8128
------------------------------------------------------------------------------
```

- ロバスト標準誤差を使う場合は一番最後に robust とつけるだけでいいです。

```
. reg pricemy lotsizetb sqrfttb, robust

Linear regression                               Number of obs     =         88
                                                F(2, 85)          =      32.81
                                                Prob > F          =     0.0000
                                                R-squared         =     0.6631
                                                Root MSE          =     603.12

------------------------------------------------------------------------------
             |               Robust
     pricemy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   lotsizetb |   .7827759   .4457812     1.76   0.083    -.1035566    1.669108
     sqrfttb |   49.39334   6.234889     7.92   0.000     36.99671    61.78997
       _cons |   59.32415   336.7108     0.18   0.861    -610.1472    728.7955
------------------------------------------------------------------------------
```

25

```
. reg pricemy lotsizetb sqrfttb

      Source |       SS           df       MS      Number of obs   =        88
-------------+----------------------------------   F(2, 85)        =     83.67
       Model |    60866850         2    30433425   Prob > F        =    0.0000
    Residual |  30918600.8        85  363748.245   R-squared       =    0.6631
-------------+----------------------------------   Adj R-squared   =    0.6552
       Total |  91785450.8        87  1055005.18   Root MSE        =    603.12

------------------------------------------------------------------------------
     pricemy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   lotsizetb |   .7827759   .2394666     3.27   0.002     .3066522      1.2589
     sqrfttb |   49.39334   4.221075    11.70   0.000     41.00071    57.78597
       _cons |   59.32415   235.1236     0.25   0.801    -408.1645    526.8128
------------------------------------------------------------------------------
```

- ロバスト標準誤差を使う場合は一番最後に robust とつけるだけでいいです。

推定された係数は変わりません。

いずれも最小二乗法
で推定されていますから。

```
                                                Number of obs     =         88
                                                F(2, 85)          =      32.81
                                                Prob > F          =     0.0000
                                                R-squared         =     0.6631
                                                Root MSE          =     603.12

------------------------------------------------------------------------------
             |               Robust
     pricemy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   lotsizetb |   .7827759   .4457812     1.76   0.083    -.1035566    1.669108
     sqrfttb |   49.39334   6.234889     7.92   0.000     36.99671    61.78997
       _cons |   59.32415   336.7108     0.18   0.861    -610.1472    728.7955
------------------------------------------------------------------------------
```

26

```
. reg pricemy lotsizetb sqrfttb

      Source |       SS           df       MS      Number of obs   =        88
-------------+----------------------------------   F(2, 85)        =     83.67
       Model |    60866850         2    30433425   Prob > F        =    0.0000
    Residual |  30918600.8        85  363748.245   R-squared       =    0.6631
-------------+----------------------------------   Adj R-squared   =    0.6552
       Total |  91785450.8        87  1055005.18   Root MSE        =    603.12

------------------------------------------------------------------------------
     pricemy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   lotsizetb |   .7827759   .2394666     3.27   0.002     .3066522      1.2589
     sqrfttb |   49.39334   4.221075    11.70   0.000     41.00071    57.78597
       _cons |   59.32415   235.1236     0.25   0.801    -408.1645    526.8128
------------------------------------------------------------------------------
```

- ロバスト標準誤差を使う場合は一番最後に robust とつけるだけでいいです。

よって回帰直線の標準誤差や決定係数
なども変わりません。

```
                                                Number of obs     =         88
                                                F(2, 85)          =      32.81
                                                Prob > F          =     0.0000
                                                R-squared         =     0.6631
                                                Root MSE          =     603.12

------------------------------------------------------------------------------
             |               Robust
     pricemy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   lotsizetb |   .7827759   .4457812     1.76   0.083    -.1035566    1.669108
     sqrfttb |   49.39334   6.234889     7.92   0.000     36.99671    61.78997
       _cons |   59.32415   336.7108     0.18   0.861    -610.1472    728.7955
------------------------------------------------------------------------------
```

27

```
. reg pricemy lotsizetb sqrfttb

      Source |       SS           df       MS      Number of obs   =        88
-------------+----------------------------------   F(2, 85)        =     83.67
       Model |    60866850         2    30433425   Prob > F        =    0.0000
    Residual |  30918600.8        85  363748.245   R-squared       =    0.6631
-------------+----------------------------------   Adj R-squared   =    0.6552
       Total |  91785450.8        87  1055005.18   Root MSE        =    603.12

------------------------------------------------------------------------------
     pricemy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   lotsizetb |   .7827759   .2394666     3.27   0.002     .3066522      1.2589
     sqrfttb |   49.39334   4.221075    11.70   0.000     41.00071    57.78597
       _cons |   59.32415   235.1236     0.25   0.801    -408.1645    526.8128
------------------------------------------------------------------------------
```

- ロバスト標準誤差を使う場合は一番最後に robust とつけるだけでいいです。

```
. reg pricemy lotsizetb sqrfttb, robust

Linear regression
```

変わるのは標準誤差です。

```
------------------------------------------------------------------------------
             |               Robust
     pricemy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   lotsizetb |   .7827759   .4457812     1.76   0.083    -.1035566    1.669108
     sqrfttb |   49.39334   6.234889     7.92   0.000     36.99671    61.78997
       _cons |   59.32415   336.7108     0.18   0.861    -610.1472    728.7955
------------------------------------------------------------------------------
```

28

```
. reg pricemy lotsizetb sqrfttb

      Source |       SS           df       MS      Number of obs   =        88
-------------+----------------------------------   F(2, 85)        =     83.67
       Model |    60866850         2    30433425   Prob > F        =    0.0000
    Residual |  30918600.8        85  363748.245   R-squared       =    0.6631
-------------+----------------------------------   Adj R-squared   =    0.6552
       Total |  91785450.8        87  1055005.18   Root MSE        =    603.12

------------------------------------------------------------------------------
     pricemy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   lotsizetb |   .7827759   .2394666     3.27   0.002     .3066522      1.2589
     sqrfttb |   49.39334   4.221075    11.70   0.000     41.00071    57.78597
       _cons |   59.32415   235.1236     0.25   0.801    -408.1645    526.8128
------------------------------------------------------------------------------
```

- ロバスト標準誤差を使う場合は一番最後に robust とつけるだけでいいです。

```
. reg pricemy lotsizetb sqrfttb, robust

Linear regression
```

変わるのは標準誤差です。

それに伴い、t値、p値、信頼区間も変わります。

```
------------------------------------------------------------------------------
             |               Robust
     pricemy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   lotsizetb |   .7827759   .4457812     1.76   0.083    -.1035566    1.669108
     sqrfttb |   49.39334   6.234889     7.92   0.000     36.99671    61.78997
       _cons |   59.32415   336.7108     0.18   0.861    -610.1472    728.7955
------------------------------------------------------------------------------
```



```
. reg pricemy lotsizetb sqrfttb

      Source |       SS           df       MS      Number of obs   =        88
-------------+----------------------------------   F(2, 85)        =     83.67
       Model |    60866850         2    30433425   Prob > F        =    0.0000
    Residual |  30918600.8        85  363748.245   R-squared       =    0.6631
-------------+----------------------------------   Adj R-squared   =    0.6552
       Total |  91785450.8        87  1055005.18   Root MSE        =    603.12

------------------------------------------------------------------------------
     pricemy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   lotsizetb |   .7827759   .2394666     3.27   0.002     .3066522      1.2589
     sqrfttb |   49.39334   4.221075    11.70   0.000     41.00071    57.78597
       _cons |   59.32415   235.1236     0.25   0.801    -408.1645    526.8128
```

この例では、土地面積の有意性に変化があります。

普通の標準誤差を使うと土地面積は1%水準でも有意ですが、
ロバスト標準誤差を使った場合は5%水準で有意ではありません。
（10%水準で有意ですが、10%水準だとエビデンスとしては弱いです）

```
             |               Robust
     pricemy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   lotsizetb |   .7827759   .4457812     1.76   0.083    -.1035566    1.669108
     sqrfttb |   49.39334   6.234889     7.92   0.000     36.99671    61.78997
       _cons |   59.32415   336.7108     0.18   0.861    -610.1472    728.7955
------------------------------------------------------------------------------
```



```
. reg pricemy lotsizetb sqrfttb

      Source |       SS           df       MS      Number of obs   =        88
-------------+----------------------------------   F(2, 85)        =     83.67
       Model |    60866850         2    30433425   Prob > F        =    0.0000
    Residual |  30918600.8        85  363748.245   R-squared       =    0.6631
-------------+----------------------------------   Adj R-squared   =    0.6552
```

F値も違っています。

不均一分散が存在する場合、通常のF統計量は妥当では無くなります。

robustのオプションを使うと、
Stataは不均一分散に対してロバストなF統計量を計算します。

- ...いいです。

```
. reg pricemy lotsizetb sqrfttb, robust

Linear regression                               Number of obs     =         88
                                                F(2, 85)          =      32.81
                                                Prob > F          =     0.0000
                                                R-squared         =     0.6631
                                                Root MSE          =     603.12
```

不均一分散に対してロバストなF統計量については、
Wooldridge, J.M. (2010) *Econometric Analysis of Cross Section and Panel Data*, Cambridge, MA: MIT Press を参照してください。

- 今の例のように、均一分散を仮定して標準誤差を計算した場合と、不均一分散の可能性を考慮に入れて標準誤差を計算した場合で、検定の結果が変わってしまうことが有ります。
- 必ずではないのですが、後者の標準誤差の方が大きくなることが多いです。
- よって、変数Aの係数が、、、
  - 普通の標準誤差を使って検定すると、統計的に有意にゼロと異なる。
  - しかし、ロバスト標準誤差を使って検定すると、そうでは無い。

というケースがよく起こります。

- このような場合、ロバスト標準誤差を使って検定するのが一般的ですから、「変数Aは従属変数に影響を与えない」と結論付けるのが普通ですね。

不均一分散の検定

- 「誤差項が不均一分散している」かどうかを検定することができます
  - ロバスト標準誤差を使うのがスタンダードになったため、最近ではこの検定はあまりされなくなりました。
  - とはいえ、いろいろな文脈で誤差項の不均一分散の検定はされることがあります（このコースの範囲外の文脈ですが）。
  - また、回帰分析における有名な検定の一つですので、詳しく解説することにします。
- 次のモデルを考えます：

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + U_i,\ E(U_i | X_{1i}, X_{2i}, \ldots, X_{ki}) = 0$$

- 帰無仮説は均一分散です：

$$H_0: VAR(U_i | X_{1i}, X_{2i}, \ldots, X_{ki}) = \sigma^2$$

33

- $E(U_i | X_{1i}, X_{2i}, \ldots, X_{ki}) = 0$が成り立っているため、

$$VAR(U_i | X_{1i}, X_{2i}, \ldots, X_{ki}) = E(U_i^2 | X_{1i}, X_{2i}, \ldots, X_{ki})$$

- よって均一分散であるという帰無仮説は

$$H_0: E(U_i^2 | X_{1i}, X_{2i}, \ldots, X_{ki}) = \sigma^2$$

と等しくなります。

- また繰り返し期待値の法則より、

$$E(U_i^2) = E\left(E(U_i^2 | X_{1i}, X_{2i}, \ldots, X_{ki})\right) = E(\sigma^2) = \sigma^2$$

なので

$$H_0: E(U_i^2 | X_{1i}, X_{2i}, \ldots, X_{ki}) = E(U_i^2) = \sigma^2$$

ですね。

34

$$\delta H_0: E(U_i^2 | X_{1i}, X_{2i}, \ldots, X_{ki}) = E(U_i^2) = \sigma^2$$

- これは、均一分散を検定するためには、「U_i^2が説明変数と関係していない」、ということを検定すればよいことを示しています。
- いくつかアプローチがありますが、ここでは最も単純なものを取り上げます。
- まず線形関数を仮定します。

$$U_i^2 = \delta_0 + \delta_1 X_{1i} + \delta_2 X_{i2} + \cdots + \delta_k X_{ik} + V_i$$

ここでV_iは誤差項です。

- この回帰の左辺は、ちょっと変な感じがしますが、もとのモデルの「誤差項の二乗」であることに注意して下さい。
- 帰無仮説（均一分散）は、

$$H_0: \delta_1 = \delta_2 = \cdots = \delta_k = 0$$

35

- $\delta_1 = \delta_2 = \cdots = \delta_k = 0$であれば、$U_i^2$がいかなる説明変数とも関係していないわけですから、分散は一定（均一分散）ということですね。
- $H_0: \delta_1 = \delta_2 = \cdots = \delta_k = 0$は、結合帰無仮説ですから、$F$検定をすればよいです。
- $U_i^2 = \delta_0 + \delta_1 X_{1i} + \delta_2 X_{i2} + \cdots + \delta_k X_{ik} + V_i$
- 左辺は元の式の「誤差項の二乗」です。
- 誤差項は見えません、データとしてはありません。
- 実際の検定では、ここを「残差の二乗」で置き換えます。

36

不均一分散の検定のまとめ

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + U_i,\ E(U_i | X_{1i}, X_{2i}, \ldots, X_{ki}) = 0$$

$$H_0: VAR(U_i | X_{1i}, X_{2i}, \ldots, X_{ki}) = \sigma^2$$

- ステップ1: モデルを最小二乗法で推定し残差$\widehat{U}_i$を出す。そして残差の二乗$\widehat{U}_i^2$を計算する。
- ステップ2: $\widehat{U}_i^2$を従属変数、$X_{1i}, X_{2i}, \ldots, X_{ki}$を説明変数としたモデルを最小二乗法を使って推定する。

$$\widehat{U}_i^2 = \delta_0 + \delta_1 X_{1i} + \delta_2 X_{i2} + \cdots + \delta_k X_{ik} + V_i$$

- ステップ3: $H_0: \delta_1 = \delta_2 = \cdots = \delta_k = 0$を$F$検定する。
  - ➢ F検定は通常のものでいい。不均一分散に対してロバストにする必要はない。
  - ➢ H_0 を棄却できない → 均一分散を棄却できるだけの十分な証拠はない（均一分散）。
  - ➢ H_0を棄却できる → 均一分散を棄却。不均一分散と結論付ける。

やってみましょう

```
. reg pricemy lotsizetb sqrfttb

      Source |       SS           df       MS      Number of obs   =        88
-------------+----------------------------------   F(2, 85)        =     83.67
       Model |    60866850         2    30433425   Prob > F        =    0.0000
    Residual |  30918600.8        85  363748.245   R-squared       =    0.6631
-------------+----------------------------------   Adj R-squared   =    0.6552
       Total |  91785450.8        87  1055005.18   Root MSE        =    603.12

------------------------------------------------------------------------------
     pricemy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   lotsizetb |   .7827759   .2394666     3.27   0.002     .3066522      1.2589
     sqrfttb |   49.39334   4.221075    11.70   0.000     41.00071    57.78597
       _cons |   59.32415   235.1236     0.25   0.801    -408.1645    526.8128
------------------------------------------------------------------------------
```

- 残差を計算します。predictコマンドを使えばいいです。
- predict 変数名（自分でつけます）, r　　で残差を計算します。
- ここでは残差の変数名をresidとします。predict resid, r
- これの二乗を計算します。変数名は residsq とします。（ステップ１終わり）

- residsqを従属変数、lotsizetbとsqrfttbを説明変数にしたモデルを推定します（ステップ2）。

```
. reg residsq lotsizetb sqrfttb

      Source |       SS           df       MS      Number of obs   =        88
-------------+----------------------------------   F(2, 85)        =      7.93
       Model |  7.0715e+12         2  3.5358e+12   Prob > F        =    0.0007
    Residual |  3.7888e+13        85  4.4574e+11   R-squared       =    0.1573
-------------+----------------------------------   Adj R-squared   =    0.1375
       Total |  4.4959e+13        87  5.1677e+11   Root MSE        =   6.7e+05

------------------------------------------------------------------------------
     residsq |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   lotsizetb |   729.9321    265.085     2.75   0.007     202.8721    1256.992
     sqrfttb |   10852.95    4672.65     2.32   0.023     1562.471    20143.43
       _cons |  -416489.1   260277.4    -1.60   0.113    -933990.3      101012
------------------------------------------------------------------------------
```

- lotsizetbとsqrfttbの係数がゼロをF検定（ステップ3）。これは
- 帰無仮説は1%の有意水準で棄却できます → 不均一分散あり、ですね。
- よって普通の標準誤差は適切ではない（ロバスト標準誤差が適切）、ということになります。

- 同じことをコマンド一発でできます。モデルを推定後、

estat hettest lotsizetb sqrfttb, fs

です。

```
. reg pricemy lotsizetb sqrfttb

      Source |       SS           df       MS      Number of obs   =        88
-------------+----------------------------------   F(2, 85)        =     83.67
       Model |    60866850         2    30433425   Prob > F        =    0.0000
    Residual |  30918600.8        85  363748.245   R-squared       =    0.6631
-------------+----------------------------------   Adj R-squared   =    0.6552
       Total |  91785450.8        87  1055005.18   Root MSE        =    603.12

------------------------------------------------------------------------------
     pricemy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   lotsizetb |   .7827759   .2394666     3.27   0.002     .3066522      1.2589
     sqrfttb |   49.39334   4.221075    11.70   0.000     41.00071    57.78597
       _cons |   59.32415   235.1236     0.25   0.801    -408.1645    526.8128
------------------------------------------------------------------------------

. estat hettest lotsizetb sqrfttb, fs

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: lotsizetb sqrfttb

         F(2 , 85)    =     7.93
         Prob > F     =   0.0007
```

- これはBreusch-Pagan テストと呼ばれるものの*F*検定バージョンです。
  - ➢ コマンドラインの最後の fs っていうのが、「*F*検定バージョンでお願いします」と頼んでいるところです。
- アイデアは同じですが、別の検定統計量を使うバージョンもあります。
- Stataは、これら以外の不均一分散の検定も数多く取り揃えています。
- 興味のある人は、estate hettestで検索してみて下さい。
  - ➢ ヘルプ → 検索

大切なおまけ

- 不均一分散についての学習は以上で終わりです。
- 実証分析上、大切なことは、

「回帰分析するときは、reg の最後に robust オプションをつける」

そして、

「ロバスト標準誤差を使いました」

と論文で説明する、です。

- ただちょっとしたパズルが残りました。
- 「家の価格に土地面積が関係ない？」です。

- 「家の価格に土地面積が関係ない」なんてことはありません。
- この結果は、「モデルの関数形が適切なものではなかった」ことから生じたものだと考えられます。
- 家の価格、土地面積、建物面積、すべてに自然対数をとってみましょう。

```
. reg lpricemy llotsizetb lsqrfttb, robust

Linear regression                               Number of obs     =         88
                                                F(2, 85)          =      58.51
                                                Prob > F          =     0.0000
                                                R-squared         =     0.6353
                                                Root MSE          =     .18547

------------------------------------------------------------------------------
             |               Robust
    lpricemy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  llotsizetb |   .1684569   .0382696     4.40   0.000     .0923668    .2445471
    lsqrfttb |   .7623692   .0772034     9.87   0.000     .6088682    .9158703
       _cons |   4.024582   .3719377    10.82   0.000      3.28507    4.764094
------------------------------------------------------------------------------
```

- 土地面積は1%水準で有意ですね。

- このように不適切な関数形を使うことで、誤った結論に至ってしまうことが有ります。
- 従って、先行研究をチェックして、どのような関数形が使われているのかを知ることは大切です。
- ちなみに、今回の場合、対数一対数モデルだと、均一分散のようですね。

```
. reg lpricemy llotsizetb lsqrfttb

      Source |       SS           df       MS      Number of obs   =        88
-------------+----------------------------------   F(2, 85)        =     74.04
       Model |    5.093628         2    2.546814   Prob > F        =    0.0000
    Residual |  2.92397431        85  .034399698   R-squared       =    0.6353
-------------+----------------------------------   Adj R-squared   =    0.6267
       Total |  8.01760231        87  .092156348   Root MSE        =    .18547

------------------------------------------------------------------------------
    lpricemy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  llotsizetb |   .1684569   .0384596     4.38   0.000      .091989    .2449249
    lsqrfttb |   .7623692   .0808863     9.43   0.000     .6015457    .9231928
       _cons |   4.024582   .3222772    12.49   0.000     3.383809    4.665356
------------------------------------------------------------------------------

. estat hettest llotsizetb lsqrfttb, fs

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: llotsizetb lsqrfttb

         F(2 , 85)    =      0.78
         Prob > F     =    0.4613
```

- もう一つおまけです。対数を取ると単位が関係なくなります。

- 変数のおおもとの単位は$、square feetでした。これらの変数に対数を取って同じモデルを推定してみましょう。

  generate lprice = ln(price)

  generate llotsize = ln(lotsize)

  generate lsqrft = ln(sqrft)

```
. reg lprice llotsize lsqrft

      Source |       SS           df       MS      Number of obs   =        88
-------------+----------------------------------   F(2, 85)        =     74.04
       Model |  5.09362891         2  2.54681446   Prob > F        =    0.0000
    Residual |  2.92397461        85  .034399701   R-squared       =    0.6353
-------------+----------------------------------   Adj R-squared   =    0.6267
       Total |  8.01760352        87  .092156362   Root MSE        =    .18547

------------------------------------------------------------------------------
      lprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    llotsize |   .1684569   .0384596     4.38   0.000     .0919889    .2449249
      lsqrft |   .7623693   .0808863     9.43   0.000     .6015457    .9231929
       _cons |  -1.640071   .6018805    -2.72   0.008    -2.836771   -.4433717
------------------------------------------------------------------------------
```

- 前のページの結果と定数項を除いてすべて一致しました。

45