回帰分析I

12. 仮説検定と信頼区間

イントロダクション

Source	SS	df	MS
Model	7789.39296	1	7789.39296
Residual	144312.057	418	345.244156
Total	152101.45	419	363.010621

Number of obs =	420
F(1, 418) =	22.56
Prob > F =	0.0000
R-squared =	0.0512
Adj R-squared =	0.0489
Root MSE =	18.581

score	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
stratio	-2.279063	.4798082	-4.75	0.000	-3.2222	-1.335925
_cons	698.9222	9.467187	73.83	0.000	680.3129	717.5314

- この講義ノートを理解すると、これらの数字の意味が分かるようになります。

イントロダクション

Source	SS	df	MS
Model	7789.39296	1	7789.39296
Residual	144312.057	418	345.244156
Total	152101.45	419	363.010621

Number of obs =	420
F(1, 418) =	22.56
Prob > F =	0.0000
R-squared =	0.0512
Adj R-squared =	0.0489
Root MSE =	18.581

score	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
stratio	-2.279063	.4798082	-4.75	0.000	-3.2222	-1.335925
_cons	698.9222	9.467187	73.83	0.000	680.3129	717.5314

- この講義ノートを理解すると、これらの数字の意味が分かるようになります。

- ちなみにこれは残差の二乗和（後ほど使います）。

$$SSR = \sum_{i=1}^{N} \hat{u}_i^2$$

イントロダクション

Source	SS	df	MS
Model	7789.39296	1	7789.39296
Residual	144312.057	418	345.244156
Total	152101.45	419	363.010621

Number of obs =	420
F(1, 418) =	22.56
Prob > F =	0.0000
R-squared =	0.0512
Adj R-squared =	0.0489
Root MSE =	18.581

score	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
stratio	-2.279063	.4798082	-4.75	0.000	-3.2222	-1.335925
_cons	698.9222	9.467187	73.83	0.000	680.3129	717.5314

- この講義ノートを理解すると、これらの数字の意味が分かるようになります。

- これはモデルで説明される平方和

$$SSE = \sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2$$

イントロダクション

Source	SS	df	MS
Model	7789.39296	1	7789.39296
Residual	144312.057	418	345.244156
Total	152101.45	419	363.010621

Number of obs	420
F(1, 418)	22.56
Prob > F	0.0000
R-squared	0.0512
Adj R-squared	0.0489
Root MSE	18.581

score	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
stratio	-2.279063	.4798082	-4.75	0.000	-3.2222	-1.335925
_cons	698.9222	9.467187	73.83	0.000	680.3129	717.5314

- この講義ノートを理解すると、これらの数字の意味が分かるようになります。
- これは全変動

$$SSE = \sum_{i=1}^{N} (y_i - \bar{y})^2$$

5

イントロダクション

Source	SS	df	MS
Model	7789.39296	1	7789.39296
Residual	144312.057	418	345.244156
Total	152101.45	419	363.010621

Number of obs	420
F(1, 418)	22.56
Prob > F	0.0000
R-squared	0.0512
Adj R-squared	0.0489
Root MSE	18.581

score	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
stratio	-2.279063	.4798082	-4.75	0.000	-3.2222	-1.335925
_cons	698.9222	9.467187	73.83	0.000	680.3129	717.5314

- この講義ノートを理解すると、これらの数字の意味が分かるようになります。

$$R^2 = \frac{SSE}{SST} = \frac{7789.39}{152101} = 0.051$$

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{144312}{152101} = 0.051$$

$$\bar{R}^2 = 1 - \frac{N-1}{N-k-1}\frac{SSR}{SST} = 1 - \frac{420-1}{420-1-1}\frac{144312}{152101} = 0.049$$

6

注意

- 以下では次の単回帰モデルを考えます。

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

そこでは、

> - 母集団からのランダムサンプル$\{(X_i, Y_i): i = 1, \dots, N\}$がある。
> - Uの条件付期待値はゼロ： $E(U|X) = 0$ → 従って$E(Y|X) = \beta_0 + \beta_1 X$
> - 誤差項UのXについての条件つき分散は一定： $VAR(U|X) = \sigma^2$

が成り立っている前提で話を進めます。

> - なので「バイアス」の話はここでは関係ありません。

7

一つの係数に関する仮説検定：t検定

- 「変数Xは変数Yに影響を与える」という仮説の実証を試みました。
- データを使って、$Y = \beta_0 + \beta_1 X + U$を最小二乗法で推定、次の結果を得ました

$$\hat{Y} = 2 + 4X$$

すなわち$\hat{\beta}_0$=2, $\hat{\beta}_1$=4ですね。

- しかし、この結果をもって、仮説はデータによりサポートされた、とは言いきれません。
- 理由：
  - ランダムサンプルの実現値によって、係数の推定値がいろんな値を取りうるのはシミュレーションで見ましたね。
  - ってことは、極端な話、β_1の真の値が0（$\beta_1 = 0$）だったとしても、β_1の推定値が4（$\hat{\beta}_1$=4）となることはあり得るわけです。
  - ということで、$\hat{\beta}_1$=4だけでは、仮説のサポートとしては弱いですね。

8

一つの係数に関する仮説検定: *t* 検定

- ではどうすれば、、、、ここで出てくるのが「仮説検定」です。
- 帰無仮説を

「説明変数Xは従属変数Y（の期待値）に影響を与えない」

と設定、それを棄却することができれば、

「XがYに影響を与える」

という統計的な証拠になります。

- 従って、帰無仮説・対立仮説を

$H_0: \beta_1 = 0$ （帰無仮説）
$H_1: \beta_1 \neq 0$ （両側の対立仮説）

のように設定して、これを検定すればいいですね。

一つの係数に関する仮説検定: *t* 検定

- より一般的には

$H_0: \beta_1 = \alpha$ （帰無仮説）
$H_1: \beta_1 \neq \alpha$ （両側の対立仮説）

（ここでαは特定の数）のような仮説を検定をしたいこと、あります。

- これらは一つの係数に関する仮説で、「t検定」とよばれるもので検定できます。
- ここではこの「t検定」について説明します。

- t-統計量は一般的に以下のように表現できます:

$$t = \frac{\text{推定量} - \text{仮説の値}}{\text{推定量の標準誤差}}$$

- 平均の差の検定を思い出しましょう:
  - $H_0: E(Y) = \mu_{Y,0}$（帰無仮説） $H_1: E(Y) \neq \mu_{Y,0}$（両側対立仮説）
  - このときのt-統計量は

$$t = \frac{\bar{Y} - \mu_{Y,0}}{SE(\bar{Y})}$$

でした。

- 回帰分析における一つの係数についてのt検定も基本的には同じです。
- 例えば、β_1の値についての検定であれば、
  - $H_0: \beta_1 = \alpha$ （帰無仮説） $H_1: \beta_1 \neq \alpha$（両側の対立仮説）
  - t-統計量は

$$\boxed{t = \frac{\hat{\beta}_1 - \alpha}{SE(\hat{\beta}_1)}}$$

となります。

t 検定の手順

- 一つの係数についてのt検定の手順は、母集団の平均に関する検定と同じです。
- ステップ0: 帰無仮説・対立仮説の設定

$H_0: \beta_1 = \alpha$ （帰無仮説） $H_1: \beta_1 \neq \alpha$（両側の対立仮説）

αの値は分析者が検定したい仮説によります。

- ステップ1: $\hat{\beta}_1$の標準誤差$SE(\hat{\beta}_1)$を求める（講義ノート9の73ページ）。
- ステップ2： サンプルの下でのt-統計量の実現値t^{act}（t値）を計算する。

$$t^{act} = \frac{\hat{\beta}_1^{act} - \alpha}{SE(\hat{\beta}_1)}$$

$\hat{\beta}_1^{act}$は最小二乗推定値（実際に推定された値）です。

- ステップ3：p値を求めます。
  - 帰無仮説が正しいというという前提の下、、、
  - t-統計量が、ステップ2で計算された実現値t^{act}よりも、起こりづらい値をとる確率

$$p\text{値} = Pr\left(|\hat{\beta}_1 - \alpha| > |\hat{\beta}_1^{act} - \alpha|\right) = Pr\left(\left|\frac{\hat{\beta}_1 - \alpha}{SE(\hat{\beta}_1)}\right| > \left|\frac{\hat{\beta}_1^{act} - \alpha}{SE(\hat{\beta}_1)}\right|\right) = Pr(|t| > |t^{act}|)$$

- ステップ4：有意水準を決めます。そして、p値を使い帰無仮説を棄却するかどうかを判断。

有意水準10%の場合：p値が10%以下なら、帰無仮説を棄却
有意水準5%の場合　：p値が5%以下なら、帰無仮説を棄却
有意水準1%の場合　：p値が1%以下なら、帰無仮説を棄却

- ノート：有意水準を先に決定、対応する棄却域（臨界値）を求める、そしてt-統計量の実現値t^{act}が棄却域に入るかどうかをみる、というやり方でももちろんいいです。

13

計量のパッケージでレポートされるt

- 計量のパッケージでレポートされるこの「t」は、、、

```
      Source |       SS           df       MS      Number of obs   =       420
-------------+----------------------------------   F(1, 418)       =     22.56
       Model |  7789.39296         1  7789.39296   Prob > F        =    0.0000
    Residual |  144312.057       418  345.244156   R-squared       =    0.0512
-------------+----------------------------------   Adj R-squared   =    0.0489
       Total |   152101.45       419  363.010621   Root MSE        =    18.581

------------------------------------------------------------------------------
       score |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     stratio |  -2.279063   .4798082    -4.75   0.000      -3.2222   -1.335925
       _cons |   698.9222   9.467187    73.83   0.000     680.3129    717.5314
------------------------------------------------------------------------------
```

- $H_0: \beta_1 = 0$を検定する際のt値です。すなわち

$$t^{act} = \frac{\hat{\beta}_1^{act}}{SE(\hat{\beta}_1)}$$

- 得られた係数（Coef.）を推定量の標準誤差（Std.Err.）で割ったもの、 t値（t）になっていること、確認してください。

$$-\frac{2.2791}{0.4798} = -4.75$$

14

```
      Source |       SS           df       MS      Number of obs   =       420
-------------+----------------------------------   F(1, 418)       =     22.56
       Model |  7789.39296         1  7789.39296   Prob > F        =    0.0000
    Residual |  144312.057       418  345.244156   R-squared       =    0.0512
-------------+----------------------------------   Adj R-squared   =    0.0489
       Total |   152101.45       419  363.010621   Root MSE        =    18.581

------------------------------------------------------------------------------
       score |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     stratio |  -2.279063   .4798082    -4.75   0.000      -3.2222   -1.335925
       _cons |   698.9222   9.467187    73.83   0.000     680.3129    717.5314
------------------------------------------------------------------------------
```

- この値が今のt値に対応するp値です。

$$p\text{値} = Pr(|t| > |-4.75|)$$

- この値、ここでは非常に小さいです、小数点3桁まではゼロですね。
- ですから、有意水準が10%であれ、5%であれ、1%であれ、帰無仮説$H_0: \beta_1 = 0$は棄却できるということになります。
- この例では、「変数*stratio*の係数はゼロ」という帰無仮説が棄却され、「*stratio*は*score*（ここでの従属変数）に影響を与える」ということについて統計的なサポートを得ることができました。

15

- 計量のパッケージでレポートされる「t」は、、、

```
      Source |       SS           df       MS      Number of obs   =       420
-------------+----------------------------------   F(1, 418)       =     22.56
       Model |  7789.39296         1  7789.39296   Prob > F        =    0.0000
    Residual |  144312.057       418  345.244156   R-squared       =    0.0512
-------------+----------------------------------   Adj R-squared   =    0.0489
       Total |   152101.45       419  363.010621   Root MSE        =    18.581

------------------------------------------------------------------------------
       score |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     stratio |  -2.279063   .4798082    -4.75   0.000      -3.2222   -1.335925
       _cons |   698.9222   9.467187    73.83   0.000     680.3129    717.5314
------------------------------------------------------------------------------
```

- この値は、$H_0: \beta_0 = 0$を検定する際のt値です。
  - 「定数項はゼロ」が帰無仮説です。
- p値は非常に小さいです、小数点3桁まではゼロですね。
- ですから、有意水準が10%であれ、5%であれ、1%であれ、帰無仮説$H_0: \beta_0 = 0$は棄却できるということになります。
  - ちなみに、実際の分析においては、定数項がゼロかどうかはあまり興味の対象にはなりませんが。

16

- 仮説検定のステップを説明しましたが、実際はほとんどの部分、計量パッケージがやります。
  - ステップ1－3については計量パッケージがやります。
- ステップ4についてですが、実際の研究では、一つの有意水準を決めて、、、というふうにはしません。
- というのも、推定結果は次のページのように報告するのが一般的だからです。
  - 結果の報告の仕方についてはこの講義の後半で詳しく説明しますが、ここで簡単に紹介しておきます。

Table 8. Estimation Results for Electricity Expenditure

推定された係数

標準誤差

変数名

	(1)	(2)	(3)
ems	-0.059***	-0.077***	-0.065***
	(0.018)	(0.017)	(0.019)
Area-average temperature in January 2014	0.002	0.001	0.001
	(0.010)	(0.008)	(0.008)
All-electric house		0.315***	0.363***
		(0.033)	(0.031)
Photovoltaic system			-0.207***
			(0.049)
Male	0.045**	0.051***	0.050***
	(0.019)	(0.019)	(0.019)
Age	0.006***	0.007***	0.007***
	(0.001)	(0.001)	(0.001)

以下省略

Note: Standard errors are presented in parentheses. ***, **, and * correspond to the one, five, and ten percent levels of significance, respectively

- このように星（＊）をつけて報告します。
- この論文では、$H_0: \beta_j = 0$が有意水準10%の下で棄却できるのなら星を一つ（＊）、5%なら星を二つ（＊＊）、1%なら星を三つ（＊＊＊）つけています。
- なので、ステップ4は、実際のところ、p値を見て「星」をつける作業と言えます。

t統計量の分布

- ここまで一つの係数に関する t 検定のやり方について説明してきました。
- ただし一つだけ説明していないことが有ります。
- それはt統計量の分布です。

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

- この分布が分からなければ、p値は計算できません、よって仮説検定もできません。
  - できませんは、いいすぎました。
  - というのも、計量パッケージにはこの分布がすでにプログラムされていて、p値を自動的に計算してくれるからです（$\hat{\beta}_1$の推定値、$\hat{\beta}_1$の推定量の標準誤差もですが）。
  - しかし、自分が何をやっているのかを、ざっくりでも知っていることは大切です。

t統計量の分布

- この分布を知るためには、最小二乗推定量$\hat{\beta}_1$の分布を知る必要があります。
- この分布について、結果から先に言います。
- サンプルサイズが大きいのならば、最小二乗推定量$\hat{\beta}_1$は近似的に

$$\hat{\beta}_1 \sim N(\beta_1, \sigma^2/N\sigma_X^2)$$

に従います。ただし$\sigma_X^2 = VAR(X)$です。ですから、$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2/N\sigma_X^2}} \sim N(0,1)$となります。

- 従って、$H_0: \beta_1 = \alpha$の下では$\frac{\hat{\beta}_1 - \alpha}{\sigma/\sqrt{N\sigma_X^2}} \sim N(0,1)$となります。
- この分母は講義ノート9で導出した$SE(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2}}$で推定でき、それに置き換えたもの、$\frac{\hat{\beta}_1 - \alpha}{SE(\hat{\beta}_1)}$の分布は変わらず$N(0,1)$です。

t統計量の分布

- ながなが説明してきましたが、結局、、、、

サンプルが大きいのなら、帰無仮説$H_0: \beta_1 = \alpha$のもとで、

$$t = \frac{\hat{\beta}_1 - \alpha}{SE(\hat{\beta}_1)} \sim N(0,1)$$

と近似できるということです。

- 従って、$H_0: \beta_1 = \alpha$（帰無仮説） $H_1: \beta_1 \neq \alpha$（両側の対立仮説）の場合、棄却域は以下の通りです。

有意水準10%なら、棄却域は絶対値で1.65以上。

有意水準5%なら、棄却域は絶対値で1.96以上。

有意水準1%なら、棄却域は絶対値で2.58以上。

t検定の例

```
      Source |       SS       df       MS              Number of obs =     420
-------------+------------------------------           F(  1,   418) =   22.56
       Model |  7789.39296     1  7789.39296           Prob > F      =  0.0000
    Residual |  144312.057   418  345.244156           R-squared     =  0.0512
-------------+------------------------------           Adj R-squared =  0.0489
       Total |   152101.45   419  363.010621           Root MSE      =  18.581

------------------------------------------------------------------------------
       score |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     stratio |  -2.279063   .4798082    -4.75   0.000      -3.2222   -1.335925
       _cons |   698.9222   9.467187    73.83   0.000     680.3129    717.5314
------------------------------------------------------------------------------
```

- このアウトプットのt値は$H_0: \beta_1 = 0$ を検定するためのt値でした。
- ここでは$H_0: \beta_1 = -2$ $H_1: \beta_1 \neq -2$ を検定してみましょう。
- サンプルが大きいのなら、帰無仮説$H_0: \beta_1 = \alpha$のもとで、$t = \frac{\hat{\beta}_1 - \alpha}{SE(\hat{\beta}_1)} \sim N(0,1)$ です。
- 従って$H_0: \beta_1 = -2$ の下では、$t = \frac{\hat{\beta}_1 + 2}{SE(\hat{\beta}_1)} \sim N(0,1)$となります。
- この実現値は$t^{act} = \frac{-2.2791+2}{0.4798} = \frac{0.2791}{0.4798} = 0.582$となります。
- よって$H_0: \beta_1 = -2$は有意水準10%でも棄却できない、ということになります。

t統計量の分布

- ちなみに、サンプルサイズが大きいのならば、最小二乗推定量$\hat{\beta}_1$は近似的に

$$\hat{\beta}_1 \sim N(\beta_1, \sigma^2 / N\sigma_X^2)$$

というのは、大数の法則と中心極限定理を使って証明できます。

> 証明は、例えば、Wooldridge (2006) *Introductory Econometrics*, Third Edition を参照してください。

> 参考までにざっくり証明を次ページに与えておきます（興味のある人だけ見てください）。

- 多重回帰モデルでもやることは基本的には同じです。
- 少し違うのは$SE(\hat{\beta}_j)$の計算の仕方（講義ノート9の$VAR(\hat{\beta}_1)$とノート10の$VAR(\hat{\beta}_j)$の違いより）。
- ですがパッケージを使う分には気にしなくていいです、パッケージが計算しますから。

- $\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^N (X_i - \bar{X}) U_i}{\sum_{i=1}^N (X_i - \bar{X})^2}$
- $\hat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^N (X_i - \bar{X}) U_i}{\sum_{i=1}^N (X_i - \bar{X})^2}$
- $\sqrt{N}(\hat{\beta}_1 - \beta_1) = \left[\frac{1}{N}\sum_{i=1}^N (X_i - \bar{X})^2\right]^{-1} \sqrt{N}\frac{1}{N}\sum_{i=1}^N (X_i - \bar{X}) U_i$
- 第一項の$\frac{1}{N}\sum_{i=1}^N (X_i - \bar{X})^2$は大数の法則より$\sigma_X^2 = VAR(X)$に確率収束。
- $\sqrt{N}\frac{1}{N}\sum_{i=1}^N (X_i - \bar{X}) U_i = \sqrt{N}\frac{1}{N}\sum_{i=1}^N (X_i - \mu_X) U_i + (\mu_X - \bar{X})[\sqrt{N}\frac{1}{N}\sum_{i=1}^N U_i]$ ただし$\mu_X = E(X)$
- これの第二項は、、、

$plim(\mu_X - \bar{X}) = 0$ （大数の法則）

$\sqrt{N}\frac{1}{N}\sum_{i=1}^N U_i = \sqrt{N}\bar{U} \approx N(0, \sigma^2)$ (中心極限定理）

よって$(\mu_X - \bar{X})[\sqrt{N}\frac{1}{N}\sum_{i=1}^N U_i]$は$N$が十分に大きいときは無視してOK。

- $\sqrt{N}\frac{1}{N}\sum_{i=1}^N (X_i - \mu_X) U_i$を考える。

$E\big((X_i - \mu_X)U_i\big) = E(X_i U_i) - \mu_X E(U_i) = E(X_i)E(U_i) - \mu_X E(U_i) = 0$
$VAR\big((X_i - \mu_X)U_i\big) = \sigma^2 \sigma_X^2$

だから中心極限定理より$\sqrt{N}\frac{1}{N}\sum_{i=1}^N (X_i - \mu_X) U_i \approx N(0, \sigma^2 \sigma_X^2)$

- 先の結果と合わせると、Nが十分に大きいとき、

$$\sqrt{N}(\hat{\beta}_1 - \beta_1) \approx \left[\frac{1}{N}\sum_{i=1}^N (X_i - \bar{X})^2\right]^{-1} \sqrt{N}\frac{1}{N}\sum_{i=1}^N (X_i - \mu_X) U_i \approx N(0,\ \sigma^2 \sigma_X^2 / \{\sigma_X^2\}^2) = N(0,\ \sigma^2 / \sigma_X^2)$$

t検定に関して知っておくべきこと

- ここで扱ったt検定ですが、誤差項Uについて特定の分布は仮定していません。
- 例えば、「誤差項は正規分布する」などという仮定は置いていないということです。
- 必要なのは、「サンプルサイズが大きければ、、、」です。
  - これは、ここで扱ったt検定が、中心極限定理による近似に基づいているためです。
- しかし、「大きければ、、、」というのは、必ずしもN = 10000ぐらいなければだめだ、という意味ではありません。
- 中心極限定理のシミュレーションでは、N = 30程度でも分布の近似はそれなりだったし、N = 100もあれば近似はかなり正確でした。
- なので、異常にサンプルサイズが小さい、ということでもない限り、ここで紹介した方法を使って問題はないです。

t検定に関して知っておくべきこと

- 初歩の回帰分析や計量経済学の教科書では、「誤差項は正規分布する」という仮定の下でのt検定について説明してあることが多いです。
- その仮定の下では、t統計量はt分布と呼ばれるものに従います。
- しかし、実際のデータでは誤差項が正規分布していることはまれで、分析では「誤差項は正規分布する」という仮定は置かないのが普通です。
- 従って、この講義では、初歩の教科書で説明されているが実際の分析では使わないt検定ではなくて、実際に分析で使われているt検定を紹介しました。
- 統計学系や経済学系の大学院にいくのなら「誤差項は正規分布する」という仮定の下でのt検定は知っておいた方がいいですが、実証分析をする上ではこの授業で扱ったt検定で十分です。

信頼区間

```
      Source |       SS           df       MS      Number of obs   =       420
-------------+----------------------------------   F(1, 418)       =     22.56
       Model |  7789.39296         1  7789.39296   Prob > F        =    0.0000
    Residual |  144312.057       418  345.244156   R-squared       =    0.0512
-------------+----------------------------------   Adj R-squared   =    0.0489
       Total |   152101.45       419  363.010621   Root MSE        =    18.581

------------------------------------------------------------------------------
       score |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     stratio |  -2.279063   .4798082    -4.75   0.000      -3.2222   -1.335925
       _cons |   698.9222   9.467187    73.83   0.000     680.3129    717.5314
------------------------------------------------------------------------------
```

- 次はこれ行きましょう、「**95%信頼区間**（confidence interval）」と呼ばれるものです。
- この言葉だけ聞くと、

「この係数の真の値が、95%の確率で[-3.222, -1.335]の範囲に含まれる」

という風に思うかもしれません。

- でもこれはおかしなステートメントです。

- というのは、係数の真の値は決まった値であって、確率変数ではありません。
- 従って、係数の真の値は[-3.222, -1.335]の範囲に入っているか、入っていないかですね。
- 決まった値（動かない値）が、95%の確率で[-3.222, -1.335]の範囲に含まれる、ってのはちょっとおかしいわけです。
- それでは**信頼区間**とはなんでしょうか？

- 先述の通り、サンプルサイズが大きいなら、近似的に、

$$\frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \sim N(0,1)$$

が成り立ちます。

  - ここで$\hat{\beta}_1$は最小二乗推定量です（確率変数）、推定値（実現値）ではありません。

- $\frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)}$は標準正規確率変数ですから、

$$\Pr\left(-1.96 \leq \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \leq 1.96\right) = 0.05$$

が成り立ちます（「1.96」については講義ノート11参照）。

- ちょっといじると、次の式が得られます。

$$Pr\left(\hat{\beta}_1 - 1.96 \cdot SE(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + 1.96 \cdot SE(\hat{\beta}_1)\right) = 0.05$$

- $Pr\left(\hat{\beta}_1 - 1.96 \cdot SE(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + 1.96 \cdot SE(\hat{\beta}_1)\right) = 0.05$

- これ、どういう意味でしょうか？

- 繰り返します、ここで確率変数は$\hat{\beta}_1$です。

- この式が意味するのは、次の通りです。

  - ランダムサンプルを一回引きます。それを使って最小二乗法でβ_1を推定します。
  - それを$\hat{\beta}_1^{(1)}$とします（実現値）。
  - これを使って上の区間を計算します$\left[\hat{\beta}_1^{(1)} - 1.96 \cdot SE\left(\hat{\beta}_1^{(1)}\right),\ \ \hat{\beta}_1^{(1)} + 1.96 \cdot SE\left(\hat{\beta}_1^{(1)}\right)\right]$

  - 仮想的ですが、またランダムサンプルを引きます。それを使って最小二乗法でβ_1を推定します。
  - それを$\hat{\beta}_1^{(2)}$とします（実現値）。
  - これを使って上の区間を計算します$\left[\hat{\beta}_1^{(2)} - 1.96 \cdot SE\left(\hat{\beta}_1^{(2)}\right),\ \ \hat{\beta}_1^{(2)} + 1.96 \cdot SE\left(\hat{\beta}_1^{(2)}\right)\right]$

  - 何回も（M回）これを繰り返します。いっぱい区間ができます。

$$\left[\hat{\beta}_1^{(1)} - 1.96 \cdot SE\left(\hat{\beta}_1^{(1)}\right),\ \ \hat{\beta}_1^{(1)} + 1.96 \cdot SE\left(\hat{\beta}_1^{(1)}\right)\right]$$
$$\left[\hat{\beta}_1^{(2)} - 1.96 \cdot SE\left(\hat{\beta}_1^{(2)}\right),\ \ \hat{\beta}_1^{(2)} + 1.96 \cdot SE\left(\hat{\beta}_1^{(2)}\right)\right]$$

……………………………………………………

$$\left[\hat{\beta}_1^{(M)} - 1.96 \cdot SE\left(\hat{\beta}_1^{(M)}\right),\ \ \hat{\beta}_1^{(M)} + 1.96 \cdot SE\left(\hat{\beta}_1^{(M)}\right)\right]$$

このたくさんある区間の中の95%が（別の言い方をすれば0.95M個の区間が）、真の値β_1を含んでいる、

ということです。

- 結果として報告する95%信頼区間は、次のように計算します：

$$\left[\hat{\beta}_1^{act} - 1.96 \cdot SE\left(\hat{\beta}_1^{act}\right),\ \ \hat{\beta}_1^{act} + 1.96 \cdot SE\left(\hat{\beta}_1^{act}\right)\right]$$

ここで$\hat{\beta}_1^{act}$はデータを使って得たβ_1の推定値、$SE\left(\hat{\beta}_1^{act}\right)$は標準誤差です。


```
      Source |       SS           df       MS      Number of obs   =       420
-------------+----------------------------------   F(1, 418)       =     22.56
       Model |  7789.39296         1  7789.39296   Prob > F        =    0.0000
    Residual |  144312.057       418  345.244156   R-squared       =    0.0512
-------------+----------------------------------   Adj R-squared   =    0.0489
       Total |   152101.45       419  363.010621   Root MSE        =    18.581

------------------------------------------------------------------------------
       score |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     stratio |  -2.279063   .4798082    -4.75   0.000      -3.2222   -1.335925
       _cons |   698.9222   9.467187    73.83   0.000     680.3129    717.5314
------------------------------------------------------------------------------
```

- 結果として報告する95%信頼区間は、次のように計算します：

$$\left[\hat{\beta}_1^{act} - 1.96 \cdot SE\left(\hat{\beta}_1^{act}\right),\ \ \hat{\beta}_1^{act} + 1.96 \cdot SE\left(\hat{\beta}_1^{act}\right)\right]$$

ここで$\hat{\beta}_1^{act}$はデータを使って得たβ_1の推定値（Coef.）、$SE\left(\hat{\beta}_1^{act}\right)$は標準誤差（Std.Err.）です。

- なっているかどうか確認してみて下さい。

- 信頼区間について最後に。

- 自分が推定して得た95%信頼区間は、

$$\left[\hat{\beta}_1^{(1)} - 1.96 \cdot SE\left(\hat{\beta}_1^{(1)}\right),\ \ \hat{\beta}_1^{(1)} + 1.96 \cdot SE\left(\hat{\beta}_1^{(1)}\right)\right]$$

$$\left[\hat{\beta}_1^{(2)} - 1.96 \cdot SE\left(\hat{\beta}_1^{(2)}\right),\ \ \hat{\beta}_1^{(2)} + 1.96 \cdot SE\left(\hat{\beta}_1^{(2)}\right)\right]$$

……………………………………………………

$$\left[\hat{\beta}_1^{(M)} - 1.96 \cdot SE\left(\hat{\beta}_1^{(M)}\right),\ \ \hat{\beta}_1^{(M)} + 1.96 \cdot SE\left(\hat{\beta}_1^{(M)}\right)\right]$$

これの中のどれか一つです（普通はランダムサンプル何回もひけませんから）。

- で、この区間の中の95%は真の値を含んでいる。逆に言うと、この中の5%の区間は真の値を含んでいません。

- ということは、自分が得た95%信頼区間は、真の値β_1を含んでいる、とは必ずしも限らないということです。

33

結合仮説のテスト:F検定

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + U$$

- 先ほどやったt-検定は、一つの係数についての検定でした。
  - 例えば、「X_kがYに影響を与えるかどうか」であれば、H_0: $\beta_k = 0$ （帰無仮説）H_1: $\beta_k \neq 0$（両側の対立仮説）をt-検定する。

- ここでは二つ以上の係数に関する仮説検定を考えます。次の例を考えましょう。野球の勝率のモデルです。

$$勝率 = \beta_0 + \beta_1 打率 + \beta_2 ホームラン数 + \beta_3 防御率 + U$$

- 「野球はピッチャーがよければ勝てる。バッターは全く関係ないよ。」

- この仮説を検定したいとします。帰無仮説は、$\beta_1 = \beta_2 = 0$ですね。

- 対立仮説は、$\beta_1 \neq 0$ and/or $\beta_2 \neq 0$ です。

34

- これは「**結合仮説**」と呼ばれるもの例。

- この例では、帰無仮説は二つの係数に制約を与えます。

- これを、モデルに2つの制約（$\beta_1 = 0$と$\beta_2 = 0$）を課す、と言います。

- 一般的に言うと、「結合仮説」とは、

モデルの係数に二つかそれ以上の制約を課す仮説のこと

- 上のモデルにおける結合帰無仮説と対立仮説の別の例は、

H_0: $\beta_1 = \beta_2 = \beta_3 = 0$　　H_1: $\beta_1 = 0$ and/or $\beta_2 = 0$ and/or $\beta_3 = 0$

  - この結合帰無仮説は、「説明変数はすべて勝率とは関係ない」。
  - 対立仮説は、どれか一つ（または一つ以上）の等式が誤り。
  - 制約の数は3つ（$\beta_1 = 0, \beta_2 = 0, \beta_3 = 0$）。

35

個別の係数を一つずつテストすればいいんじゃない？

- 結合帰無仮説が$\beta_1 = \beta_2 = 0$のとき、個別の係数を一つずつテストすればいいんじゃない？って思うかもしれません。

- t-検定を2回やると。
  - H_0: $\beta_1 = 0$ H_1: $\beta_1 \neq 0$ で t-検定
  - H_0: $\beta_2 = 0$ H_1: $\beta_2 \neq 0$ でもう一回 t-検定する。
  - いずれかの仮説を棄却出来たら結合帰無仮説を棄却する。

- これは残念ながらあまり信頼できるアプローチではありません。

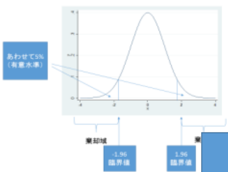
- なぜかというと、仮説検定の「サイズ」が誤ったものになってしまう可能性があるからです。

36

- 仮説検定の「**サイズ**」とは

帰無仮説が正しいにも関わらず、それを棄却してしまう確率

- 単純なt-検定を考えます。



仮説検定の考え方：

検定統計量が棄却域に入ったら、

「そんなに都合よくレアなこと起きないでしょ。てことは、帰無仮説が正しいっていう前提そのものが間違っているんじゃないの。この仮説は棄て棄て」

でした。

しかし本当にレアなことが起きていた、
ことはありえるわけです。

従って、仮説検定では、有意水準の確率で、帰無仮説が正しいのにも関わらずそれを棄却していることになります
（これを第一種の過誤といいます）。

まあ、それを知った上で仮説検定するわけですが。

- H_0: $\beta_1 = \beta_2 = \beta_3 = 0$、この結合帰無仮説を有意水準5%で検定したいとします。
- すなわち「サイズ」は5%にしたいと。
- 「仮説が正しいときに5%の確率で棄却してしまう、まあ5%は仕方ない」って考えているわけです。
- ではt-検定を2回やってみましょう。それぞれ有意水準5%で。
- H_0: $\beta_1 = 0$、H_0: $\beta_2 = 0$、どちらか一つでも棄却したら、結合帰無仮説を棄却します。
- このとき「サイズ」はいくらになるでしょうか？
- 単純化のために、二つのt-統計量は独立とします。

- 結合仮説が正しいのにも関わらず棄却してしまう確率は、、、
- 個別の帰無仮説が二つとも棄却されない確率を求めて、それを1から引けば出ますね。
- H_0: $\beta_1 = 0$ H_1: $\beta_1 \neq 0$ で t-検定。棄却されない確率は0.95
- H_0: $\beta_2 = 0$ H_1: $\beta_2 \neq 0$ でt-検定。棄却されない確率は0.95
- 従って2回とも棄却しない確率は、0.95*0.95 = 90.25%
- ということは、このアプローチは、結合仮説があっているときに、100－90.25 = 9.75%の確率で棄却します（サイズは9.75%です）。
- このケースでは、結合帰無仮説を過剰に棄却してしまいます。
- 二つのt-統計量が独立でない場合も、サイズは誤ったものになってしまいます。

結合仮説のテスト：F検定の手順

- それでは結合仮説のテストである F検定を説明します。

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + U$$

- H_0: $\beta_1 = \beta_2 = 0$ とします。ここでは制約は2つです。
- まず制約あり（restricted）のモデルを考えます。制約が正しい下では、

$$Y = \beta_0 + \beta_3 X_3 + U$$

このモデルを推定したとして、その残差の二乗和ををSSR_Rとします。

- 次に制約なし（unrestricted）モデルを考えます。

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + U$$

このモデルを推定したとして、その残差の二乗和をSSR_{UR}とします。

- そして次の統計量を考えます：

$$F = \frac{(SSR_R - SSR_{UR})/q}{SSR_{UR}/(N-k-1)}$$

ここでqは制約の数、kは説明変数の数です。

- ➢ 従って、今の例では$q = 2, k = 3$です。

- これはF統計量と呼ばれるものです。
- サンプルサイズが大きいとき、帰無仮説の下で、この統計量は$F_{q,\infty}$分布に従うことが知られています（導出は難しいのしません）。
- あとは、有意水準を決めて、$F_{q,\infty}$分布に基づき、棄却域を決定。
- 実際に制約ありモデル、制約なしモデルを推定、 F統計量の実現値を計算。それが棄却域に入れば、結合帰無仮説を棄却します。

41

カイ二乗（χ^2）分布

- 新しい分布、F分布とはどのような分布でしょうか？
- F分布を理解するためには、もう一つ別の分布を理解する必要があります。
- それは「カイ二乗分布」です。

42

カイ二乗（χ^2）分布

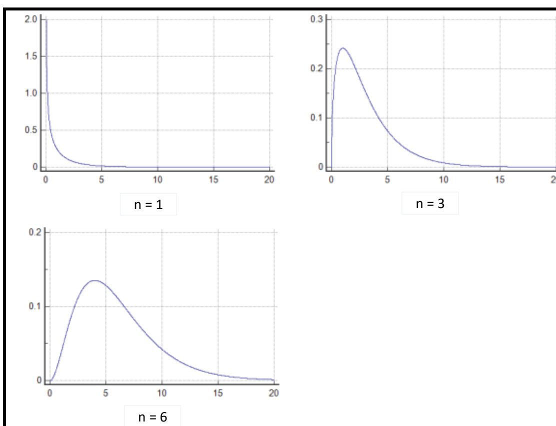
- カイ二乗分布は、

標準正規分布に従うn個の独立した確率変数をそれぞれ二乗して合計した値が従う分布

に当たります。

- この分布はn に依存し、 n はカイ二乗分布の「自由度」と呼ばれます。
- 自由度nのカイ二乗分布をχ_n^2と書くこともあります。
- 例えば、Z_1, Z_2, Z_3を独立した標準正規確率変数とすると、$Z_1^2 + Z_2^2 + Z_3^2$は自由度3のカイ二乗分布に従います。

43



44

カイ二乗分布表 上側確率

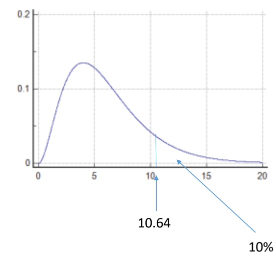
n（自由度）	10%	5%	1%
1	2.71	3.84	6.63
2	4.61	5.99	9.21
3	6.25	7.81	11.34
4	7.78	9.49	13.28
5	9.24	11.07	15.09
6	10.64	12.59	16.81
7	12.02	14.07	18.48
8	13.36	15.51	20.09
9	14.68	16.92	21.67
10	15.99	18.31	23.21
11	17.28	19.68	24.72
12	18.55	21.03	26.22
13	19.81	22.36	27.69
14	21.06	23.68	29.14
15	22.31	25.00	30.58
16	23.54	26.30	32.00
17	24.77	27.59	33.41
18	25.99	28.87	34.81
19	27.20	30.14	36.19
20	28.41	31.41	37.57
21	29.62	32.67	38.93
22	30.81	33.92	40.29
23	32.01	35.17	41.64
24	33.20	36.42	42.98
25	34.38	37.65	44.31
26	35.56	38.89	45.64
27	36.74	40.11	46.96
28	37.92	41.34	48.28
29	39.09	42.56	49.59
30	40.26	43.77	50.89

45

カイ二乗分布表 上側確率

n（自由度）	10%	5%	1%
1	2.71	3.84	6.63
2	4.61	5.99	9.21
3	6.25	7.81	11.34
4	7.78	9.49	13.28
5	9.24	11.07	15.09
6	10.64	12.59	16.81
7	12.02	14.07	18.48
8	13.36	15.51	20.09
9	14.68	16.92	21.67
10	15.99	18.31	23.21
11	17.28	19.68	24.72
12	18.55	21.03	26.22
13	19.81	22.36	27.69
14	21.06	23.68	29.14
15	22.31	25.00	30.58
16	23.54	26.30	32.00
17	24.77	27.59	33.41
18	25.99	28.87	34.81
19	27.20	30.14	36.19
20	28.41	31.41	37.57
21	29.62	32.67	38.93
22	30.81	33.92	40.29
23	32.01	35.17	41.64
24	33.20	36.42	42.98
25	34.38	37.65	44.31
26	35.56	38.89	45.64
27	36.74	40.11	46.96
28	37.92	41.34	48.28
29	39.09	42.56	49.59
30	40.26	43.77	50.89

例えば、χ_6^2で上側確率10%なら



$Pr(\chi_6^2 \geq 10.64) = 0.1$

46

Q1: $X \sim \chi_4^2$ このとき$Pr(X \leq 7.78)$は？

Q2: $X \sim \chi_{10}^2$ このとき$Pr(X > 18.31)$は？

47

F分布

- W: 自由度mのカイ二乗変数
  V: それと独立した自由度nのカイ二乗変数

このとき

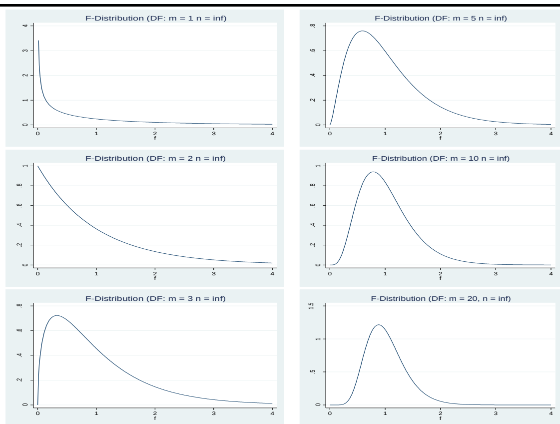
$$\frac{W/m}{V/n}$$

は

は、分子の自由度m、分母の自由度n のF分布に従います。

- 以下、$F_{m,n}$と表します。この確率変数は負の値は取りません。
- F分布の特別なケースとして重要なのは、分母の自由度nが十分に大きいケースで、その結果$F_{m,n}$分布が$F_{m,\infty}$分布として近似できる場合です。

48

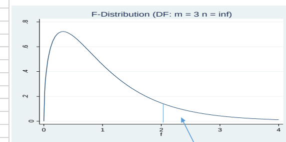


F分布表 上側確率：$F_{m,\infty}$から抜粋

m	10%	5%	1%
1	2.71	3.84	6.64
2	2.30	3.00	4.61
3	2.08	2.61	3.79
4	1.95	2.37	3.32
5	1.85	2.22	3.02
6	1.78	2.10	2.81
7	1.72	2.01	2.64
8	1.67	1.94	2.51
9	1.63	1.88	2.41
10	1.60	1.83	2.32
11	1.57	1.79	2.25
12	1.55	1.75	2.19
13	1.53	1.72	2.13
14	1.51	1.69	2.09
15	1.49	1.67	2.04
16	1.47	1.65	2.00
17	1.46	1.62	1.97
18	1.45	1.61	1.94
19	1.43	1.59	1.91
20	1.42	1.57	1.88

注： EXCELのfinv関数を使って作成

例えば、$F_{3,\infty}$で上側確率10%なら



2.08

10%

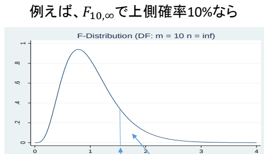
$Pr(F_{3,\infty} \geq 2.08) = 0.1$

F分布表 上側確率：$F_{m,\infty}$から抜粋

m	10%	5%	1%
1	2.71	3.84	6.64
2	2.30	3.00	4.61
3	2.08	2.61	3.79
4	1.95	2.37	3.32
5	1.85	2.22	3.02
6	1.78	2.10	2.81
7	1.72	2.01	2.64
8	1.67	1.94	2.51
9	1.63	1.88	2.41
10	1.60	1.83	2.32
11	1.57	1.79	2.25
12	1.55	1.75	2.19
13	1.53	1.72	2.13
14	1.51	1.69	2.09
15	1.49	1.67	2.04
16	1.47	1.65	2.00
17	1.46	1.62	1.97
18	1.45	1.61	1.94
19	1.43	1.59	1.91
20	1.42	1.57	1.88

注： EXCELのfinv関数を使って作成

例えば、$F_{10,\infty}$で上側確率10%なら



1.60

10%

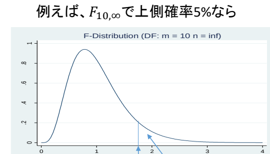
$Pr(F_{10,\infty} \geq 1.60) = 0.1$

F分布表 上側確率：$F_{m,\infty}$から抜粋

m	10%	5%	1%
1	2.71	3.84	6.64
2	2.30	3.00	4.61
3	2.08	2.61	3.79
4	1.95	2.37	3.32
5	1.85	2.22	3.02
6	1.78	2.10	2.81
7	1.72	2.01	2.64
8	1.67	1.94	2.51
9	1.63	1.88	2.41
10	1.60	1.83	2.32
11	1.57	1.79	2.25
12	1.55	1.75	2.19
13	1.53	1.72	2.13
14	1.51	1.69	2.09
15	1.49	1.67	2.04
16	1.47	1.65	2.00
17	1.46	1.62	1.97
18	1.45	1.61	1.94
19	1.43	1.59	1.91
20	1.42	1.57	1.88

注： EXCELのfinv関数を使って作成

例えば、$F_{10,\infty}$で上側確率5%なら



1.83

5%

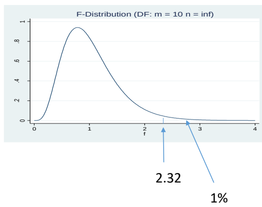
$Pr(F_{10,\infty} \geq 1.83) = 0.05$

F分布表 上側確率：$F_{m,\infty}$から抜粋

m	10%	5%	1%
1	2.71	3.84	6.64
2	2.30	3.00	4.61
3	2.08	2.61	3.79
4	1.95	2.37	3.32
5	1.85	2.22	3.02
6	1.78	2.10	2.81
7	1.72	2.01	2.64
8	1.67	1.94	2.51
9	1.63	1.88	2.41
10	1.60	1.83	$\boxed{2.32}$
11	1.57	1.79	2.25
12	1.55	1.75	2.19
13	1.53	1.72	2.13
14	1.51	1.69	2.09
15	1.49	1.67	2.04
16	1.47	1.65	2.00
17	1.46	1.62	1.97
18	1.45	1.61	1.94
19	1.43	1.59	1.91
20	1.42	1.57	1.88

注： EXCELのfinv関数を使って作成

例えば、$F_{10,\infty}$で上側確率5%なら



$$Pr\left(F_{10,\infty} \geq 2.32\right) = 0.01$$

53

カイ二乗分布と$F_{m,\infty}$分布の関係

- 自由度nのカイ二乗分布は、標準正規分布（Z_i）に従うn個の独立した確率変数をそれぞれ二乗して合計した値が従う分布。

$$Z_1^2 + \cdots + Z_n^2 \sim \chi_n^2$$

- Wを自由度mのカイ二乗変数、Vをそれと独立した自由度nのカイ二乗変数としたとき、$\frac{W/m}{V/n}$は分子の自由度m、分母の自由度n の$F_{m,n}$分布に従う。

- これの分母を丁寧に書くと、

$$\frac{W/m}{(Z_1^2 + \cdots + Z_n^2)/n}$$

  - $F_{m,\infty}$分布はnが十分に大きいときなので、分母は1（Z_i^2の平均、大数の法則より）。
  - よってW/mと等しい。

- これより、

$F_{m,\infty}$分布は自由度mのカイ二乗分布をmで割ったものと等しい

54

カイ二乗分布表 上側確率

n（自由度）	10%	5%	1%
1	2.71	3.84	6.63
2	4.61	5.99	9.21
3	6.25	7.81	11.34
4	7.78	9.49	13.28
5	9.24	11.07	15.09
6	10.64	12.59	16.81
7	12.02	14.07	18.48
8	13.36	15.51	20.09
9	14.68	16.92	21.67
10	15.99	18.31	23.21
11	17.28	19.68	24.72
12	18.55	21.03	26.22
13	19.81	22.36	27.69
14	21.06	23.68	29.14
15	22.31	25.00	30.58
16	23.54	26.30	32.00
17	24.77	27.59	33.41
18	25.99	28.87	34.81
19	27.20	30.14	36.19
20	28.41	31.41	37.57

F分布表 上側確率：$F_{m,\infty}$から抜粋

m	10%	5%	1%
1	2.71	3.84	6.64
2	2.30	3.00	4.61
3	2.08	2.61	3.79
4	1.95	2.37	3.32
5	1.85	2.22	3.02
6	1.78	2.10	2.81
7	1.72	2.01	2.64
8	1.67	1.94	2.51
9	1.63	1.88	2.41
10	1.60	1.83	2.32
11	1.57	1.79	2.25
12	1.55	1.75	2.19
13	1.53	1.72	2.13
14	1.51	1.69	2.09
15	1.49	1.67	2.04
16	1.47	1.65	2.00
17	1.46	1.62	1.97
18	1.45	1.61	1.94
19	1.43	1.59	1.91
20	1.42	1.57	1.88

$\chi_m^2 = m \cdot F_{m,\infty}$が成り立っていることを確認

55

F検定の手順

- まず制約あり（restricted）のモデルを推定。この残差の二乗和をSSR_Rとする。
- 次に制約なし（unrestricted）モデルを推定。残差の二乗和をSSR_{UR}とする。
- F統計量

$$F = \frac{(SSR_R - SSR_{UR})/q}{SSR_{UR}/(N-k-1)}$$

ここでqは制約の数、kは説明変数の数。この統計量はNが十分に大きいとき、$F_{q,\infty}$分布に従う。

- 統計量の実現値をF^{act}とする。
- この値F^{act}が帰無仮説の下ではめったに起こらないほど「大きければ」、帰無仮説を棄却。
- そうでなければ帰無仮説は棄却できない。

56

- めったに起こらないほど「大きければ」＝棄却域

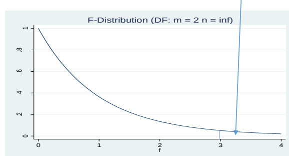
- 棄却域は制約の数（＝自由度）と有意水準に依存。

- 例えば、制約の数が2で有意水準が5%なら、

F分布表 上側確率：$F_{m,\infty}$から抜粋

m	10%	5%	1%
1	2.71	3.84	6.64
2	2.30	3.00	4.61
3	2.08	2.61	3.79
4	1.95	2.37	3.32
5	1.85	2.22	3.02
6	1.78	2.10	2.81
7	1.72	2.01	2.64
8	1.67	1.94	2.51
9	1.63	1.88	2.41
10	1.60	1.83	2.32
11	1.57	1.79	2.25
12	1.55	1.75	2.19
13	1.53	1.72	2.13
14	1.51	1.69	2.09
15	1.49	1.67	2.04
16	1.47	1.65	2.00
17	1.46	1.62	1.97
18	1.45	1.61	1.94
19	1.43	1.59	1.91
20	1.42	1.57	1.88

5%



棄却域

F^{act}が3より大きければ帰無仮説を棄却。
そうでなければ、帰無仮説は棄却できない。

回帰モデルの全体有意性の検定

```
      Source |       SS           df       MS      Number of obs   =       420
-------------+----------------------------------   F(  1,   418)   =     22.56
       Model |  7789.39296         1  7789.39296   Prob > F        =    0.0000
    Residual |  144312.057       418  345.244156   R-squared       =    0.0512
-------------+----------------------------------   Adj R-squared   =    0.0489
       Total |   152101.45       419  363.010621   Root MSE        =    18.581

------------------------------------------------------------------------------
       score |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     stratio |  -2.279063   .4798082    -4.75   0.000      -3.2222   -1.335925
       _cons |   698.9222   9.467187    73.83   0.000     680.3129    717.5314
------------------------------------------------------------------------------
```

- ようやくこれです。これでアウトプット全部の説明が終わります。

- これはF検定の結果です。

- ただし帰無仮説が「すべての説明変数の係数がゼロ」です。
  - この例では説明変数は一つなので H_0: $\beta_1 = 0$です。

- 従って、この帰無仮説は「モデルに説明力は全く無い」というものです。

- 一般的には、

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

がモデルだと、H_0: $\beta_1 = \cdots = \beta_k = 0$で、これを$F$検定したときの検定統計量の実現値が$F$としてリポートされます。

- この例では、説明変数は一つなのでH_0: $\beta_1 = 0$、制約は一つ。

- $F(1, \ldots)$ 最初の数は制約の数（＝説明変数の数）。この例では1

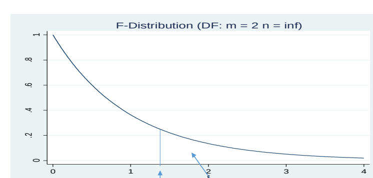
- $F(.,418)$ 次の数は$N - k - 1$。この例では$420-1-1 = 418$。

- $F(1{,}418) = 22.56$ この値が検定統計量の実現値です（検定統計量は$F_{1,\infty}$分布）。

- Prob > F この値はp値です。

- 今の例の場合は$Pr(F_{1,\infty} > 22.56)$が報告されています。ほぼゼロです。

- 従って、「モデルに全く説明力が無い」という帰無仮説は1%の有意水準で棄却されます。



- 一般的には、検定統計量の実現値がここなら、ここの確率がp値になります。

- 例で得られている$F(1,418) = 22.56$を確認してみます。

- 制約ありモデルは

$$SCORE = \beta_0 + U$$

この結果は

```
. reg score

      Source |       SS       df       MS              Number of obs =     420
-------------+------------------------------           F(  0,   419) =    0.00
       Model |           0     0           .           Prob > F      =       .
    Residual |   152101.45   419  363.010621           R-squared     =  0.0000
-------------+------------------------------           Adj R-squared =  0.0000
       Total |   152101.45   419  363.010621           Root MSE      =  19.053

------------------------------------------------------------------------------
       score |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   654.1602   .9296833   703.64   0.000     652.3328    655.9877
------------------------------------------------------------------------------
```

この値が残差の二乗和です。　なので$SSR_R = 152101.45$

- 例で得られている$F(1,418) = 22.56$を確認してみます。

- 制約なしモデルは

$$SCORE = \beta_0 + \beta_1 STRATIO + U$$

この結果は

```
. reg score stratio

      Source |       SS       df       MS              Number of obs =     420
-------------+------------------------------           F(  1,   418) =   22.56
       Model |  7789.39296     1  7789.39296           Prob > F      =  0.0000
    Residual |  144312.057   418  345.244156           R-squared     =  0.0512
-------------+------------------------------           Adj R-squared =  0.0489
       Total |   152101.45   419  363.010621           Root MSE      =  18.581

------------------------------------------------------------------------------
       score |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     stratio |  -2.279063   .4798082    -4.75   0.000      -3.2222   -1.335925
       _cons |   698.9222   9.467187    73.83   0.000     680.3129    717.5314
------------------------------------------------------------------------------
```

$SSR_R = 152101.45$

この値が残差の二乗和です。　なので$SSR_{UR} = 144312.057$

$$\boxed{F = \frac{(SSR_R - SSR_{UR})/q}{SSR_{UR}/(N-k-1)}}$$

$$F^{act} = \frac{(152101.45 - 144312.057\ )/1}{144312.057/(420-1-1)} = \frac{7789.393}{345.244} = 22.56$$

一致しました。

F検定：Stataのコマンドでやってみる

- F検定のためのStataのコマンドは test です。

- 色々な検定ができます。

□ test _b[変数名] = 0（H_0:「変数」の係数は0）

□ test _b[変数名] = 10（H_0:「変数」の係数は10）

□ test _b[変数名1] = _b[変数名2] = 0 （H_0:「変数1」の係数は0 &「変数2」の係数は0）

□ test _b[変数名1] = _b[変数名2] = _b[変数名3] = 0 （H_0:「変数1~3」の係数は0）


```
. reg score stratio

      Source |       SS       df       MS              Number of obs =     420
-------------+------------------------------           F(  1,   418) =   22.56
       Model |  7789.39296     1  7789.39296           Prob > F      =  0.0000
    Residual |  144312.057   418  345.244156           R-squared     =  0.0512
-------------+------------------------------           Adj R-squared =  0.0489
       Total |   152101.45   419  363.010621           Root MSE      =  18.581

------------------------------------------------------------------------------
       score |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     stratio |  -2.279063   .4798082    -4.75   0.000      -3.2222   -1.335925
       _cons |   698.9222   9.467187    73.83   0.000     680.3129    717.5314
------------------------------------------------------------------------------
```

```
. test _b[stratio] = 0

 ( 1)  stratio = 0

       F(  1,   418) =   22.56
            Prob > F =    0.0000
```

```
. test _b[stratio] = 10

 ( 1)  stratio = 10

       F(  1,   418) =  654.93
            Prob > F =    0.0000
```

これまでのおさらい：出生体重の決定要因

- データ： BWGHT_NOMISS.DTA
  - データの出典 Wooldridge, J.M.（2006）*Introductory Econometrics, 3rd edition, Thomson South-western.*
- モデル：$bwght = \beta_0 + \beta_1 cigs + \beta_2 parity + \beta_3 faminc + \beta_4 motheduc + \beta_5 fatheduc + u$

Q1: このモデルに説明力があるかどうかをテストしたいとします。帰無仮説と対立仮説を書きなさい。そしてその帰無仮説を5%の有意水準でテストしなさい。使用する検定統計量、その分布、棄却域も答えなさい。

Q2: 出生体重の動きをこのモデルはどの程度説明していますか？

Q3: $bwght = \beta_0 + \beta_1 cigs + \beta_2 parity + \beta_3 faminc + \varepsilon$ このモデルと先に推定したモデルはどちらの方がデータによりフィットしていますか？

Q4: 他の要因をコントロールしたとき、出生の順番は出生体重に影響を与えるかテストしたいとします？帰無仮説と対立仮説を書きなさい。そしてその帰無仮説を5%の有意水準でテストしなさい。使用する検定統計量、その分布、棄却域も答えなさい。また係数を解釈しなさい。

65

Q5：他の要因をコントロールしたとき、親の教育水準は出生体重に影響を与えるかどうかテストしたいとします？結合帰無仮説と対立仮説を書きなさい。そしてその結合帰無仮説を10%の有意水準でテストしなさい。使用する検定統計量、その分布、棄却域も答えなさい。

Q6：Q5の結合帰無仮説を個別に仮説検定しなさい（ただし有意水準はそれぞれ10%とします）。使用する検定統計量、その分布、棄却域も答えなさい。また、結果についてコメントしなさい。

Q7: $cigs$と$parity$の標本相関係数を計算して下さい。

Q8：Q7で計算した$cigs$と$parity$の標本相関係数を踏まえて次の問いに答えてください。

分析者が、出生の順番が出生体重に与える影響を考慮せずに、以下のモデルを推定したとします。

$$bwght = \beta_0 + \beta_1 cigs + \beta_3 faminc + \beta_4 motheduc + \beta_5 fatheduc + v$$

最小二乗推定量$\hat{\beta}_1$はバイアスする？一番最初に推定したモデルが真のモデルと仮定してコメントしなさい。

Q9：実際にQ8のモデルを最小二乗法で推定し、推定値$\hat{\beta}_1$についてコメントしなさい。

66

Q10: 出生体重に影響を与える要因のひとつに、新生児の性別があります。しかし、最初に推定したモデルには$male$（ = 1 if male child）が含まれていませんでした。

あなたの推定結果を見た人が、

「$male$がモデルに含まれていないため、最初のモデルの推定結果には欠落変数バイアスの問題生じているはずだ。そのためあなたの推定結果は信用できない。」

とコメントしたとします。このコメントにコメントし返して下さい。

Q11：次のモデルを推定しなさい。

$$bwght = \beta_0 + \beta_1 cigs + \beta_2 parity + \beta_3 faminc + \beta_4 motheduc + \beta_5 fatheduc + \beta_6 male + \omega$$

$male$は$bwght$に影響を与える？有意水準5%で仮説検定しなさい。使用する検定統計量、その分布、棄却域も答えなさい。

Q12: $male$と他の説明変数の相関係数を計算しなさい。

Q13：Q11とQ12の結果は、Q10であなたがしたコメントをサポートするものですか？

Q14: $faminc$が$3,000増えたとします。このとき、この変化によって予想される$bwght$の変化は$\beta_3 \cdot 3$です。$\hat{\beta}_3 \cdot 3$に関する95%信頼区間を、Q11のモデルの結果を使って計算しなさい。

67

複数の係数が関係する制約のテスト

- 一つの制約式に二つ以上の回帰係数がかかわるような仮説をテストしたいことがあります。
- さっきのモデルを考えましょう：

$$bwght = \beta_0 + \beta_1 cigs + \beta_2 parity + \beta_3 faminc + \beta_4 motheduc + \beta_5 fatheduc + \beta_6 male + \omega$$

- 知りたいのは次の仮説の妥当性とします。

「母親の教育水準が出生体重に与える影響は、父親の教育水準が出生体重に与える影響と同じ」

- 帰無仮説は$\beta_4 = \beta_5$、対立仮説は$\beta_4 \neq \beta_5$で検定。
- この検定はSTATAのtestコマンドを使って簡単にできます。
- test _b[motheduc] = _b[fatheduc] とすればいいです。

68

```
. test _b[ motheduc] = _b[ fatheduc]

 ( 1)  motheduc - fatheduc = 0

       F(  1,  1184) =    2.40
            Prob > F =    0.1213
```

- 10%水準でも帰無仮説を棄却できない、っていう結果ですね。
  - 「違う」と結論づけるのに十分な証拠はない、ということです。
  - この結果はQ5の結果と整合的ですね。
- このようにSTATAのtestコマンドを使っていろいろな仮説を検定できます。
  - ここで紹介したのはほんの一部です。
  - 実際に使うときはマニュアルを見てみると良いでしょう。
- ちょっと適当な仮説ですが、「母親の教育水準が出生体重に与える影響は、父親の教育水準が出生体重に与える影響の2倍」を検定してみましょう。
- 帰無仮説は$\beta_4 = 2\beta_5$、対立仮説は$\beta_4 \neq 2\beta_5$ですね。
- test _b[motheduc] = _b[fatheduc]*2 とすればいいです。

69

```
. test _b[ motheduc] = _b[ fatheduc]*2

 ( 1)  motheduc - 2*fatheduc = 0

       F(  1,  1184) =    2.62
            Prob > F =    0.1059
```

- 帰無仮説は10%水準で棄却できません。
- 少し遊んでみましょう。
- 「母親の教育水準が出生体重に与える影響は、父親の教育水準が出生体重に与える影響の3倍」を検定

```
. test _b[ motheduc] = _b[ fatheduc]*3

 ( 1)  motheduc - 3*fatheduc = 0

       F(  1,  1184) =    2.66
            Prob > F =    0.1029
```

- 「母親の教育水準が出生体重に与える影響は、父親の教育水準が出生体重に与える影響の100倍」を検定

```
. test _b[ motheduc] = _b[ fatheduc]*100

 ( 1)  motheduc - 100*fatheduc = 0

       F(  1,  1184) =    2.57
            Prob > F =    0.1092
```

70

- どういうこと？
- これらの結果は極めて整合的です。

```
. test _b[ motheduc] = _b[ fatheduc] = 0

 ( 1)  motheduc - fatheduc = 0
 ( 2)  motheduc = 0

       F(  2,  1184) =    1.34
            Prob > F =    0.2635
```

- ゼロを何倍してもゼロ。
- なので、$H_0: \beta_4 = 2\beta_5$も$H_0: \beta_4 = 3\beta_5$も$H_0: \beta_4 = 100\beta_5$も棄却できなかった、というわけです。

71

複数の係数が関係する制約のテスト

- testコマンドを使わないでやってみましょう。
- いくつかやり方はありますが、ここではF検定でやってみます。
- 帰無仮説は$\beta_4 = 2\beta_5$とします。
- 制約なしモデルは、

$$bwght = \beta_0 + \beta_1 cigs + \beta_2 parity + \beta_3 faminc + \beta_4 motheduc + \beta_5 fatheduc + \beta_6 male + \omega$$

```
      Source |       SS           df       MS      Number of obs   =     1191
-------------+----------------------------------   F(  6,  1184)   =     9.79
       Model |  22815.5605         6  3802.59342   Prob > F        =   0.0000
    Residual |  459931.131      1184  388.455347   R-squared       =   0.0473
-------------+----------------------------------   Adj R-squared   =   0.0424
       Total |  482746.692      1190  405.669489   Root MSE        =   19.709

------------------------------------------------------------------------------
       bwght |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cigs |   -.583468   .1099713    -5.31   0.000    -.7992284   -.3677077
      parity |   1.836455   .6569277     2.80   0.005     .5475826    3.125327
      faminc |    .063363   .0364842     1.74   0.083     -.008218    .1349439
    motheduc |  -.3693236   .3185701    -1.16   0.247    -.9943484    .2557013
    fatheduc |   .4504962   .2815881     1.60   0.110    -.1019711    1.002964
        male |   3.727986   1.146103     3.25   0.001     1.479366    5.976605
       _cons |   112.5272   3.763887    29.90   0.000     105.1426    119.9118
------------------------------------------------------------------------------
```

72

- 制約ありモデルは少しトリッキーです。
- 制約なしモデル

$$bwght = \beta_0 + \beta_1 cigs + \beta_2 parity + \beta_3 faminc + \beta_4 motheduc + \beta_5 fatheduc + \beta_6 male + \omega$$

に$\beta_4 = 2\beta_5$の制約を課します。

$$bwght = \beta_0 + \beta_1 cigs + \beta_2 parity + \beta_3 faminc + 2\beta_5 motheduc + \beta_5 fatheduc + \beta_6 male + \omega$$

- 変形すると

$$bwght = \beta_0 + \beta_1 cigs + \beta_2 parity + \beta_3 faminc + \beta_5 (2motheduc + fatheduc) + \beta_6 male + \omega$$

となり、これが制約ありモデルになります。

- このモデルを推定するために、新しい変数 $2motheduc + fatheduc$ を作ります。
- 新しい変数には適当に名前をつけて下さい（この変数は検定のためのもので、分析上は特に興味はないです）。

73

```
. generate ftest_educ =  2*motheduc +  fatheduc

. reg bwght cigs parity faminc ftest_educ male

      Source |       SS       df       MS              Number of obs =    1191
-------------+------------------------------           F(  5,  1185) =   11.21
       Model |  21798.2927     5  4359.65853           Prob > F      =  0.0000
    Residual |  460948.399  1185  388.985991           R-squared     =  0.0452
-------------+------------------------------           Adj R-squared =  0.0411
       Total |  482746.692  1190  405.669489           Root MSE      =  19.723

------------------------------------------------------------------------------
       bwght |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cigs |  -.5798242   .1100233    -5.27   0.000    -.7956865   -.363962
      parity |    1.89161   .6564909     2.88   0.004     .6035956   3.179624
      faminc |   .0692931   .0363245     1.91   0.057    -.0019744   .1405607
  ftest_educ |   .0217611   .0954615     0.23   0.820    -.1655313   .2090534
        male |   3.760616   1.146708     3.28   0.001     1.510812   6.010421
       _cons |   112.4607   3.766233    29.86   0.000     105.0715     119.85
------------------------------------------------------------------------------
```

- $F = \frac{(SSR_R - SSR_{UR})/q}{SSR_{UR}/(N-k-1)}$
- $SSR_{UR} = 459931.131$
- $SSR_R = 460948.399$
- 制約の数（q）は1。説明変数の数（k）は6。
- $F = \frac{(460948.399 - 459931.131)/1}{459931.131/1184} = 2.618$
- 70ページのF値と一致しました。

74