統計学
回帰分析、最小二乗法5
F 検定

担当:　　長倉　大輔

（ながくらだいすけ）

F 検定

- F 検定

重回帰モデル:

$$Y_i = \alpha + \beta_1 X_{1,i} + \ldots + \beta_K X_{K,i} + u_i \;\;,\quad i = 1, \ldots, n \,,$$

に対して、t 検定はある 1 つのパラメータ β_k に関する仮説を検定するためのものであった。

それに対して、複数の係数についての仮説を検定したい時には ***F* 検定**と呼ばれる検定がよく用いられる。

F 検定

- F 検定の帰無仮説

複数の係数についての仮説とは、例えば

$$\text{(a)}\quad \mathrm{H}_0: \beta_1 = \beta_2 = \ldots = \beta_K = 0\,,$$

や

$$\text{(b)}\quad \mathrm{H}_0: \beta_1 = \beta_2\,, \qquad \text{(c)}\quad \mathrm{H}_0: 2\beta_1 + 3\beta_2 = \beta_3\,,$$

といったような仮説である。

このような仮説の検定は 1 つの係数に関する検定である t 検定ではできない。

F 検定

- F 検定統計量の計算

F 検定統計量の値は以下のように計算される。

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

ここで SSR_R は**帰無仮説を制約条件として**計算した最小二乗推定値のもとでの残差平方和、SSR_{UR} は**何の制約条件もない**時の最小二乗推定値のもとでの残差平方和である。q は**帰無仮説の下での制約の数** (つまり制約の中の "$=$" の数) である。例えば、先ほどの例では

(a) $q = K$,　(b) $q = 1$,　(c) $q = 1$ である。

F 検定

- 制約付き最小二乗推定値とその残差平方和

以下の重回帰モデルを考えよう。

$$Y_i = \alpha + \beta_1 X_{1,i} + \beta_2 X_{2,i} + u_i\,,\quad i = 1, \ldots, n,$$

さらに、H_0: $\beta_1 = \beta_2$ という帰無仮説を検定したいとしよう。この時、もしこの H_0 が正しいのであれば、上記の回帰モデルにおいて

$$\begin{aligned} \beta_1 X_{1,i} + \beta_2 X_{2,i} &= \beta_1 X_{1,i} + \beta_1 X_{2,i} \quad (\beta_2 = \beta_1 \text{ なので}) \\ &= \beta_1 (X_{1,i} + X_{2,i}) \end{aligned}$$

である。これより、先ほどの回帰式は

F 検定

$$Y_i = \alpha + \beta_1 Z_i + u_i\,, \quad Z_i = X_{1,i} + X_{2,i}\,,$$

と書き直せる。この回帰モデルに対する β_1 の最小二乗推定値は、もと回帰モデルにおいて $\mathrm{H}_0 : \beta_1 = \beta_2$ という制約を満たしているので、H_0 を制約条件とした**制約付最小二乗推定値**と呼ばれる。

$\mathrm{H}_0 : \beta_1 = \beta_2$ を検定する F 検定の計算式の中の SSR_R はこの回帰モデルを推定した時の残差平方和である。

F 検定

例題 1

重回帰モデル:

$$Y_i = \alpha + \beta_1 X_{1,i} + \beta_2 X_{2,\,i} + \beta_3 X_{3,\,i} + u_i\,, \quad i = 1, \ldots, n\,,$$

において、先ほどの例の

(c) H_0: $2\beta_1 + 3\beta_2 = \beta_3$

を制約条件とする制約付残差平方和 SSR_R を計算するには、どのような回帰モデルを推定すべきか?

F 検定

- F 検定統計量の分布

F 検定統計量は、**誤差項 u_i が正規分布に従うという仮定の下で**、自由度 $(q\,,\ n-K-1)$ の **F 分布に従う**
(この証明は行列の知識が必要で難しいので略)。

この場合、この分布は**観測値の数 (n)、制約の数(q)**、および無制約の重回帰モデルにおける(定数項を除いた)**説明変数の数(K)**に依存して決まる。

F 値 (F 検定統計量の実現値)がF 分布表から得られる臨界値より大きければ帰無仮説を棄却する事になる。

F 検定

- F 検定の例

以下の重回帰モデルを考えよう。

$$Y_i = \alpha + \beta_1 X_{1,i} + \beta_2 X_{2,i} + u_i\ ,\ \ u_i \sim i.i.d.N(0, \sigma^2),$$
$$i = 1,\ldots,10.$$

ここで、私たちは **H_0: $\beta_1 = \beta_2$** という帰無仮説を検定したいとする。この検定を F 検定統計量を用いて行う。

ここで私たちは u_i が正規分布に従っている事がわかっているとする。

F 検定

- F 検定の例

(手順1: 無制約の残差平方和の計算)

まず先ほどの重回帰モデルを何の制約もおかずに普通に最小二乗法によって推定する。その残差平方和の値を SSR_{UR} とする。

(手順2: 制約付の残差平方和の計算)

次に制約を満たす回帰式:

$$Y_i = \alpha + \beta_1 Z_i + u_i\ ,\ \ Z_i = X_{1,i} + X_{2,i}\ ,\ i = 1,\ldots,10.$$

を推定し、その残差平方和の値を SSR_R とする。

F 検定

- F 検定の例

(手順3: F 値の計算)

H_0: $\beta_1 = \beta_2$ であるので $q = 1$ 。また無制約の回帰式の説明変数の数は 2 であるので K =2。さらに観測数 n =10 もわかっているので、公式より F 値を計算することができる。

その値が ***F* = 4.95** であったとしよう。

F 検定

- F 検定の例

(手順4: 帰無仮説の検定)

私たちは u_i が正規分布に従うと想定している。よってこの場合、帰無仮説のもとで F(検定)統計量は**自由度(1, 7) のF 分布**に従う事になる。

有意水準 5 % で検定すると、今 F = 4.95 であるので、この値と F(1, 7)の分布表から導かれる臨界値($F_{1,\,7,\,0.05}$の値) を比べ、4.95がその値よりも**大きければ$H_0: \beta_1 = \beta_2$ を棄却する**。分布表より $F_{1,\,7,\,0.05}$ = 5.59 であるので棄却できない、つまり採択となる。

F 検定

例題 2:

以下の重回帰モデルを考えよう

$$Y_i = \alpha + \beta_1 X_{1,i} + \beta_2 X_{2,i} + u_i\ ,\ \ u_i \sim i.i.d.N(0, \sigma^2),$$
$$i = 1, \ldots, n.$$

上記の重回帰モデルに対して、帰無仮説 $\mathbf{H_0: \beta_1 = \beta_2 = 0}$. を検定したいとする。この時、この帰無仮説に対する F 値は F = 4.15 であったとする。n = 25 とする。

この問題に対する、5% および 1% 有意水準での F 検定の結果はどのようになるか?

F 検定

- F 検定のP値

F検定に対してP値は以下のように定義される。

F検定統計量が従うF分布に従う確率変数を F とすると、

$$F\text{検定統計量のP値} = \Pr(F > F\text{統計量の実現値})$$

である。この値が0.05より小さければそれはそのような実現値が出る確率は5%以下という事なので、帰無仮説を有意水準5%で棄却という事になる(1%の場合も同様)。

F 検定

- エクセルによるF 検定

1. エクセルを含むたくさんの統計ソフトは、通常、t 検定の場合と同様、F 検定に対しても自動的に F 値 とその **P 値** を計算する。この P 値によって F 検定を行ってもよい。
2. ただし、この場合の F 検定は、通常、**定数項以外の全ての説明変数の係数が 0 である、**という帰無仮説を検定するものである (ソフトウェアによっては違う場合があるかもしれないが、少なくとも エクセルではそう)。対立仮説は**それらの係数のうち少なくとも 1 つは 0でない**となる。

エクセルを用いた F 検定

- Excel を用いたF検定

以前に分析した 身長、体重データを用いて Excel で F 検定を行う。以下の線形回帰モデルを推定する。

$$Y_i = \alpha + \beta_1 X_{1,i} + \beta_2 X_{2,i} + u_i\,, \quad u_i \sim N(0, \sigma^2),$$

ここで

Y_i ： 男性の体重,

$X_{1,i}$： 男性の身長, $X_{2,i}$: 父親の身長

である。このモデルを Excel で推定すると、帰無仮説

$$\mathrm{H}_0\text{:}\ \beta_1 = \beta_2 = 0$$

に対する F 値を自動的に出力してくれる。

エクセルを用いた F 検定

- Excel を用いたF検定

このモデルの推定結果の出力は以下のようになる。

概要

回帰統計	
重相関 R	0.7579
重決定 R2	0.5745
補正 R2	0.5679
標準誤差	4.9692
観測数	133

分散分析表

	自由度	変動	分散	観測された分散比	有意 F
回帰	2	4333.6123	2166.8061	87.7490	7.59919E-25
残差	130	3210.1185	24.6932		
合計	132	7543.7307			

	係数	標準誤差	t	P-値	下限 95%	上限 95%
切片	-98.0196	12.4624	-7.8653	1.2453E-12	-122.6749	-73.3643
X 値 1	1.0085	0.0783	12.8856	5.4420E-25	0.8537	1.1633
X 値 2	-0.0636	0.0320	-1.9880	0.0489	-0.1268	-0.00031

エクセルを用いた F 検定

- Excel を用いたF検定

F値とP値が出力されている。

概要

回帰統計	
重相関 R	0.7579
重決定 R2	0.5745
補正 R2	0.5679
標準誤差	4.9692
観測数	133

F 値

F 検定の P 値

分散分析表

	自由度	変動	分散	観測された分散比	有意 F
回帰	2	4333.6123	2166.8061	87.7490	7.59919E-25
残差	130	3210.1185	24.6932		
合計	132	7543.7307			

	係数	標準誤差	t	P-値	下限 95%	上限 95%
切片	-98.0196	12.4624	-7.8653	1.2453E-12	-122.6749	-73.3643
X 値 1	1.0085	0.0783	12.8856	5.4420E-25	0.8537	1.1633
X 値 2	-0.0636	0.0320	-1.9880	0.0489	-0.1268	-0.00031

エクセルを用いた F 検定

F値は 87.749 である。

帰無仮説が正しいのであれば、これは $F(2, 130)$ からの実現値である。

しかしながら 対応する P値を見ると F 検定統計量が $F(2, 130)$ に従う時に 87.749 より大きくなる確率は 7.59919×10^{-25} より小さい (E–25 は 10^{-25} を意味する)

よって有意水準 1% で棄却となる。

エクセルを用いた F 検定

- より一般的な制約に対する F 検定

Excel で自動的に出力される F 値は

「 H_0: 定数項以外の係数が全て 0 」

という帰無仮説に対応している。

これ以外のより一般的な帰無仮説も Excel を用いて F 検定を行う事ができるが、これは自動的にはやってくれないので公式を用いて行う必要がある。

エクセルを用いた F 検定

ここで F 検定統計量の公式を確認しておこう。

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

SSR_R: 帰無仮説のもとでの残差平方和
SSR_{UR}: 対立仮説(無制約)のもとでの残差平方和
q: 帰無仮説の下での制約の数
n: 標本数
K: (帰無仮説のもとで)定数項を除いた説明変数の和

である。SSR_R、SSR_{UR}、q、n、K を求めればそこから計算できることがわかる。

エクセルを用いた F 検定

ここで先ほどの身長、体重データの線形回帰モデル

$$Y_i = \alpha + \beta_1 X_{1,i} + \beta_2 X_{2,i} + u_i\,, \quad u_i \sim N(0, \sigma^2),$$

に対して、帰無仮説

$$\mathrm{H}_0: \quad \beta_1 = \beta_2$$

を検定するための F 値を計算してみよう。
上記より、q =1, K= 2 はすぐさまわかる事に注意。

まず、手順1

「無制約の残差平方和の計算」

を実行する。無制約モデルの推定結果は

エクセルを用いた F 検定

以下のようになる。ここで SSR_{UR}は

概要

回帰統計	
重相関 R	0.7579
重決定 R2	0.5745
補正 R2	0.5679
標準誤差	4.9692
観測数	133

分散分析表

	自由度	変動	分散	観測された分散比	有意 F
回帰	2	4333.6123	2166.8061	87.7490	7.59919E-25
残差	130	3210.1185	24.6932		
合計	132	7543.7307			

	係数	標準誤差	t	P-値	下限 95%	上限 95%
切片	-98.0196	12.4624	-7.8653	1.2453E-12	-122.6749	-73.3643
X 値 1	1.0085	0.0783	12.8856	5.4420E-25	0.8537	1.1633
X 値 2	-0.0636	0.0320	-1.9880	0.0489	-0.1268	-0.00031

エクセルを用いた *F* 検定

以下の青丸で囲まれた値である。

概要

回帰統計	
重相関 R	0.7579
重決定 R2	0.5745
補正 R2	0.5679
標準誤差	4.9692
観測数	133

SSR_{UR}

分散分析表

	自由度	変動	分散	観測された分散比	有意 F
回帰	2	4333.6123	2166.8061	87.7490	7.59919E-25
残差	130	3210.1185	24.6932		
合計	132	7543.7307			

	係数	標準誤差	t	P-値	下限 95%	上限 95%
切片	-98.0196	12.4624	-7.8653	1.2453E-12	-122.6749	-73.3643
X 値 1	1.0085	0.0783	12.8856	5.4420E-25	0.8537	1.1633
X 値 2	-0.0636	0.0320	-1.9880	0.0489	-0.1268	-0.00031

エクセルを用いた F 検定

また、標本数 n は、以下の赤丸で囲まれた値である。

概要

回帰統計	
重相関 R	0.7579
重決定 R2	0.5745
補正 R2	0.5679
標準誤差	4.9692
観測数	133

n

分散分析表

	自由度	変動	分散	観測された分散比	有意 F
回帰	2	4333.6123	2166.8061	87.7490	7.59919E-25
残差	130	3210.1185	24.6932		
合計	132	7543.7307			

	係数	標準誤差	t	P-値	下限 95%	上限 95%
切片	-98.0196	12.4624	-7.8653	1.2453E-12	-122.6749	-73.3643
X 値 1	1.0085	0.0783	12.8856	5.4420E-25	0.8537	1.1633
X 値 2	-0.0636	0.0320	-1.9880	0.0489	-0.1268	-0.00031

エクセルを用いた F 検定

これで、先ほどの F 値の公式において、すでに SSR_{UR}、q、n、Kの値がわかった。次に手順2

「制約付の残差平方和の計算」

を行う(これは SSR_R の計算)。

先程みたように制約を満たす回帰式は

$$Y_i = \alpha + \beta_1 Z_i + u_i \,, \quad Z_i = X_{1,i} + X_{2,i} \,,$$

と書けるので、これを推定する。まず Z_i を作成する。

エクセルを用いた F 検定

「父親の身長」の右横に一列挿入し、データ名を Zi (ここは任意の名前でよい)とし、データ名の下のセルに、Ziの計算式を入力する。

(男性のデータ)			
身長	父親の身長	Zi	
172.720	176.910	=B5+C5	
174.360	158.100		
175.660	170.650		
171.720	178.850		
163.110	171.910		
171.410	170.910		
180.400	178.470		

エクセルを用いた F 検定

Enter キーを押して計算を確定し、ドラッグして「相対参照」でコピーをする(相対参照についてわからない場合は、補足資料参照)。

(男性のデータ)		
身長	父親の身長	Z_i
172.720	176.910	349.630
174.360	158.100	332.460
175.660	170.650	346.310
171.720	178.850	350.570
163.110	171.910	335.020
171.410	170.910	
180.400	178.470	

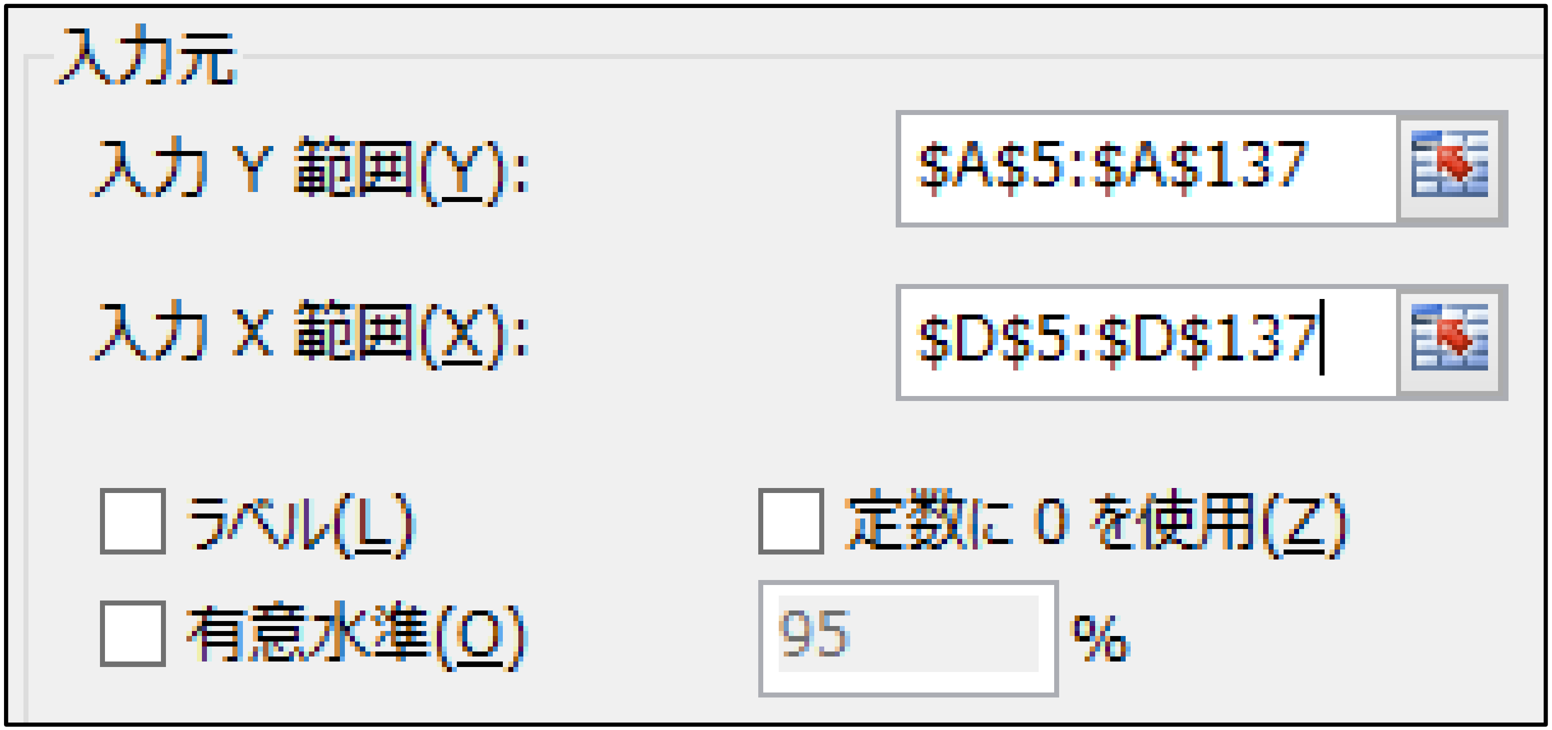
エクセルを用いた F 検定

以下のようなデータができるので、A列のデータをD列のデータに回帰させれば、SSR_Rを計算できる。

A	B	C	D
		20代男女の300人の身長デ	
	(男性のデータ)		
体重	身長	父親の身長	Zi
72.380	172.720	176.910	349.630
71.430	174.360	158.100	332.460
66.500	175.660	170.650	346.310
75.280	171.720	178.850	350.570

エクセルを用いた F 検定

「データ」タブ→「データ分析」→「回帰分析」で推定。



エクセルを用いた F 検定

推定結果、SSR_Rの値は以下のようになる。これは常にSSR_{UR}より大きくなることに注意。

概要

回帰統計	
重相関 R	0.397085
重決定 R2	0.157676
補正 R2	0.151246
標準誤差	6.964612
観測数	133

SSR_R

分散分析表

	自由度	変動	分散	測された分散	有意 F
回帰	1	1189.468	1189.468	24.52217	2.22E-06
残差	131	6354.263	48.50582		
合計	132	7543.731			

	係数	標準誤差	t	P-値	下限 95%
切片	6.324236	11.7096	0.54009	0.590052	-16.8401
X 値 1	0.169394	0.034207	4.951987	2.22E-06	0.101724

エクセルを用いた F 検定

これで必要な値は全てわかったので、公式より計算できる。エクセルで数式を入力して計算すればよい。
例えば以下のように入力して計算する。

=((G4-G3)/G6)/(G3/(G5-G7-1))

F	G	H	I
SSRUR	3210.118		
SSRR	6354.263		
n	133		
q	1		
K	2		
F	=((G4-G3)/G6)/(G3/(G5-G7-1))		

エクセルを用いた F 検定

計算結果は以下のようになる。

K	2	
F	127.3282	

帰無仮説が正しければ、この値は $F(1, 130)$ からの実現値である。F分布表には $F(1, 130)$ の%点は載っていないが、$F_{m, x, \alpha} < F_{m, y, \alpha}$ $(x > y$ に対して) という関係があるので、$F_{1,130, 0.05}$ と $F_{1,130, 0.01}$ はそれぞれ $F_{1,100, 0.05} = 3.94$ と $F_{1,100, 0.01} = 6.90$ より小さい事がわかる。よって、帰無仮説は有意水準 1%でも5%でも棄却されることがわかる。

エクセルを用いた F 検定

また、他にも、例えば先ほどの回帰式で、

$$\mathrm{H}_0:\ \alpha = 0,\ \beta_1 = 0.5,\ \beta_2 = 0$$

のような帰無仮説を検定する場合を考えてみよう
(これは体重は身長の約半分という帰無仮説)。

エクセルを用いた F 検定

この場合、$q = 3$ であり、n, K, SSR_{UR} は同じ値である。また帰無仮説の下では、

$$Y_i = 0.5\,X_{1,i} + u_i,$$

であるので、SSR_Rは 残差平方和は

$$SSR_R = \sum_{i=1}^{133}(Y_i - 0.5X_{1,i})^2$$

となる(最小二乗推定は行う必要がない事に注意)。この場合 F値を計算すると 849.75 となる。帰無仮説は明らかに棄却される。