数量分析と確率・統計

証券アナリスト

Takayuki Suzuki

This is institude of the author

回帰分析目標

回帰分析目標は、Excel で出力される値のすべてが説明できること
および単回帰分析が手で計算できること。(下記は出力例)

概要

回帰統計	
重相関 R	0.910478
重決定 R2	0.828971
補正 R2	0.815815
標準誤差	1.637172
観測数	15

分散分析表

	自由度	変動	分散	測された分散	有意 F
回帰	1	168.889	168.889	63.01051	2.44E-06
残差	13	34.84431	2.680331		
合計	14	203.7333			

	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
切片	11.84117	0.789411	15	1.38E-09	10.13575	13.54659	10.13575	13.54659
x	1.15047	0.144934	7.937916	2.44E-06	0.83736	1.46358	0.83736	1.46358

以下のスライドで、それぞれの項目に関連する説明を行う。

単回帰分析

単回帰分析については、分散、平均、共分散などの基本統計量から、
単回帰係数が計算できればよいだろう。
推定方法として最小二乗法 (OLS) を用いた単回帰の場合、回帰係数は共分散を説明変数の分散で割ったものである。
ところで、相関係数は 共分散を二つの変数の標準偏差で割ったものであるから、
回帰係数は相関係数を説明変数の標準偏差で割り、目的変数の標準偏差をかけたものに等しい

単回帰係数 $\hat{\beta}_1 = \dfrac{S_{x,y}}{S_x^2}$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}\bar{x}$$

相関係数 $\rho = \dfrac{S_{x,y}}{S_x S_y}$

したがって

$$\hat{\beta}_1 = \rho \cdot \frac{S_y}{S_x}$$

分散分析

分散分析 (Analysis Of Variance, ANOVA)
は、データの分散を、回帰式によって説明できる部分と、説明できない部分（しばしばノイズとして扱われる）に分けてその割合を分析する方法のこと。
SST (sum of squared total) 元データの分散のこと
SSR (sum of squared regression) ほぼ、予測値の分散のこと
SSE (sum of squared errors) は、残差の二乗和。

> Point :
> 全平方和は、回帰平方和と残差平方和の合計である。
> それぞれの分散の大小関係を比較することで、モデルの当てはまりの良さを評価できる。

とくに、SSE と SSR の大小関係は、後で出てくる回帰係数の仮説検定で重要になる。
もし $SSE >> SSR$ なら、ノイズの方が大きくてモデルの当てはまりが悪いといえる。
逆に $SSE << SSR$ であれば、ノイズは小さく、モデルの当てはまりが良いと言えそうだ。

回帰の標準誤差

SSE を自由度 (n - k -1) で割ったものの平方根を、残差分散という。

$$
\begin{aligned}
\text{残差分散} &= \mathrm{SSE}/(n-k-1) \\
&= \sum(\hat{y}-y)^2/(n-k-1) \\
&\text{(ただし } \hat{y} \text{ は y の予測値)}
\end{aligned}
$$

普通の不偏分散に似ているが、k (説明変数の数。単回帰なら k=1) を引いている点が異なる.
回帰の標準誤差は、残差分散の平方根である。

$$
\text{回帰の標準誤差} = \sqrt{\text{残差分散}}
$$

これは回帰の誤差項の分散の推定値である。
回帰の標準誤差（または誤差項の分散の推定値）は、後で回帰係数の検定を行う時に利用する。

決定係数

決定係数は、よくでてくる。相関係数の二乗でもある。

$$\text{決定係数 } R^2 = \frac{SSR}{SST}$$

相関係数は、決定係数のルートに等しい。
つまり、回帰式で説明可能な変動と、総変動の比である。
回帰式で説明可能な変動の割合が大きい＝相関が大きい。
Excel では「重相関」というが、まあ、相関係数のことだろう。
自由度調整済み決定係数というのもある。 決定係数を説明変数の数で割り引いたもの。

$$\text{自由度調整済み決定係数 } \bar{R}^2 = 1 - \frac{n-1}{n-k-1}\left(1 - R^2\right)$$

回帰係数の検定

良くある問題として、回帰係数が 0 と有意に異なるか否か、を検定しろというものがある。
下記の式を用いて t 値を計算し、それを閾値と比べて判定する。

$$t = \frac{\text{推定値 - 仮説値}}{\text{推定値の標準誤差}}$$

回帰係数が 0 と有意に異なるかを t 検定する場合は下記の計算になる。

$$\text{t 値} = \frac{\text{回帰係数} - 0}{\text{回帰係数の標準誤差}}$$

係数の標準誤差は、ふつう問題で与えられているはず。(または t 値から逆算する)
回帰係数の標準誤差は、残差分散を説明変数の偏差二乗和で割って、平方根を取ったものである。

回帰係数の標準誤差の算出

上記の通り、回帰係数の標準誤差は、ふつう問題文で与えられているはず。
念のため計算方法を記載する。

$$\text{回帰係数の標準誤差} = \sqrt{\frac{\text{残差分散}}{\text{x の偏差二乗和}}}$$

F 検定

F 検定は、二つの分布の分散が等しいか否かを検定する方法である。

point:
自由度 d_1,d_2 の二つのデータの不偏分散の比 s_1/s_2 は、自由度 (d_1, d_2) の F 分布 $F(d_1, d_2)$ に従う。ふつう、二つの分散が等しいという帰無仮説に対する検定を行うために使われる。検定の閾値は、F 分布表で自由度 (d_1, d_2) に対応する箇所を探して使う。 問題で与えられている場合もある。

Note:

- 正確には二つの分布が正規分布である仮定が必要だが、
  後の「回帰分析の仮定が成り立たない場合」の項でも出てこないので忘れてよい
- たまに、解説にカイ 2 乗分布がどうのこうのと書いている場合があるが、これも忘れてよい。

二つの期間のボラティリティが等しいか検定する問題 (2016 年午前第 8 問問 5) では、ボラティリティは標準偏差であるので、F 検定を適用するにはまず不偏分散を求める。不変分散はボラティリティを二乗して自由度で割ればよい。自由度はデータ数-1 である。

2016 年午前第 8 問問題 5

問題文により、状態１の時のデータ数２３２、ボラティリティ１６．１%、
状態２の時のデータ数が１２，ボラティリティ３１．７%である。
これを当てはめる。分散はボラティリティの二乗なので（ボラは標準偏差なので)、

$$F = \frac{0.161^2}{0.317^2}$$

これを自由度 $(231, 11)$ の F 分布の棄却域の下側閾値、および上側閾値と比較して回答すればよい。

回帰係数の仮設検定と F 検定

回帰分析では、しばしば、回帰式を作成してもその回帰式の検定を行うことがある。
単回帰の場合や、重回帰でも個別の係数の検定を行う場合は、回帰係数をその標準誤差 SE で割って t 値を得て、それを t 値の閾値と比較する、t 検定を行えばよい。（上記で説明済み）

一方、たまに、重回帰のすべての係数がゼロであるか否か（つまり、重回帰式は全く無意味か否か）という検定をする場合がある。この場合は、個別の係数に対する検定とは全くことなり、前記の分散分析と、F 検定を用いた検定を行う。

回帰係数の仮設検定と F 検定

戦略:
予測値の分散と、残差の分散に差があるかないか、調べる。
予測値の分散はそのま母分散 ($=$SSR$=(\hat{y}-\bar{y})^2$)
残差分散は前記で示したもの（$=$SSE$/(n-k-1)$)
この二つはどちらも分散なので、F 検定を用いて、等しいか否かを検定することができる。
自由度は、予測値は説明変数の数 k, 残差は n-k-1 である。

$$\mathrm{F} = \frac{SSR/k}{SSE/(n-k-1)}$$

この値を自由度 $(k, n-k-1)$ の F 分布の値と比較して、検定を行う。