mwp-042

anova - 分散分析/共分散分析　【 評価版 】

anova は分散分析/共分散分析の機能を提供する汎用的なコマンドです。これに対し oneway は一元配置 ANOVA に特化し、多重比較機能も包含するなど、より使いやすさを追求したコマンドです（[R] **oneway** (*mwp-190*) 参照)。また多変量の分散分析/共分散分析機能については manova コマンドが用意されています（[MV] **manova** (*mwp-103*) 参照)。

1. 分散分析

1.1 分散分析と多重比較

平均値が等しいと言えるかどうかを検定する場合、対象とする標本の数が 2 つ以下の場合には通常 ***t* 検定**が用いられます（[R] **ttest** (*mwp-041*) 参照)。それでは 3 つの標本 A, B, C が与えられたとき、t 検定を繰返し使用したら何が悪いのでしょうか？今、検定の有意水準 α を 5% とすると

(1) A-B 間の比較で有意差が検出されない確率 $= 0.95$
(2) A-C 間の比較で有意差が検出されない確率 $= 0.95$
(3) B-C 間の比較で有意差が検出されない確率 $= 0.95$

ですから、3 回の検定を通じて有意差が検出されない確率は $0.95^3 = 0.86$ となります。逆に言えば (1), (2), (3) のいずれかで有意差が検出される確率は $1 - 0.86 = 0.14$ となり、正しい帰無仮説を棄却してしまう過誤（**第 1 種過誤**）の確率が本来の 5% よりも大きく増大するといった問題が生じます。

このため標本数が 3 以上の場合には**分散分析** (ANOVA: analysis of variance) という手法が用いられます。これはすべての平均値間に差がないことを ***F* 検定**によって確認しようとするものです。しかしこの仮説が棄却された場合に、どの平均とどの平均の間に有意差が認められるかについて、分散分析自体は何ら情報をもたらしません。このため、p 値の補正を伴う**多重比較** (multiple comparison) 検定を併用することが通常行われます。

1.2 分散分析の前提条件

分散分析の実行に際しては次の要件が満たされていることが前提となります。

(1) 従属変数（応答変数）は**量的（区間尺度）データ**であること
(2) 従属変数（応答変数）は**正規分布**に従うこと
(3) 観測データは互いに**独立**であること
(4) 各グループの分散は**均一**であること

反復測定 (repeated-measures) ANOVA の場合には独立性に関する前提条件が成り立たなくなります。

2. **一元配置** ANOVA

▷ Example 1: **一元配置** ANOVA

ここでは Example データセット `apple.dta` を用いて一元配置 ANOVA の用例を紹介します。

```
. use http://www.stata-press.com/data/r16/apple.dta [*1]
(Apple trees)
```

このデータセット中にはある化学肥料の濃度がリンゴの生育にどう影響するかについてのデータが記録されています。4 種類の濃度について比較が行われたわけですが、その際の実験デザインは次の通りです。

- それぞれの濃度につき 3 箇所の果樹園をアサインし、収穫された果実の平均重量を計測する。
- それぞれの果樹園には 12 本のリンゴの木が存在する。

合計で 12 箇所の果樹園が用意されていたわけですが、うち 2 箇所の果樹園は途中で破壊されてしまったため、データの構成としてはバランスの取れていないものとなってしまっています。データセット中には `treatment` と `weight` という 2 つの変数が存在します。`treatment` は肥料の濃度を表し、1 から 4 までの整数値を取ります。一方、`weight` は収穫された果実の平均重量を示す実数値（単位は g）です。参考までにデータの内容を示しておくと次のようになります。

```
. list, abbreviate(9) sepby(treatment) [*2]
```

	treatment	weight
1.	1	117.5
2.	1	113.8
3.	1	104.4
4.	2	48.9
5.	2	50.4
6.	2	58.9
7.	3	70.4
8.	3	86.9
9.	4	87.7
10.	4	67.3

[*1] メニュー操作： File ▷ Example Datasets ▷ Stata 16 manual datasets と操作、Base Reference Manual [R] の anova の項よりダウンロードする。

[*2] メニュー操作：Data ▷ Describe data ▷ List data

よりわかりやすいテーブル形式に直すと次のようになります。データが unbalanced なものである点に注意してください。

Fertilizer	Measurements 1	2	3
1	117.5	113.8	104.4
2	48.9	50.4	58.9
3	70.4	86.9	
4	87.7	67.3	

この場合、anova に対する従属変数は `weight` であり、モデルを規定するカテゴリ変数は `treatment` ということになります。

- Statistics ▷ Linear models and related ▷ ANOVA/MANOVA
  ▷ Analysis of variance and covariance と操作
- Model タブ: Dependent variable: `weight`
  Model: `treatment`

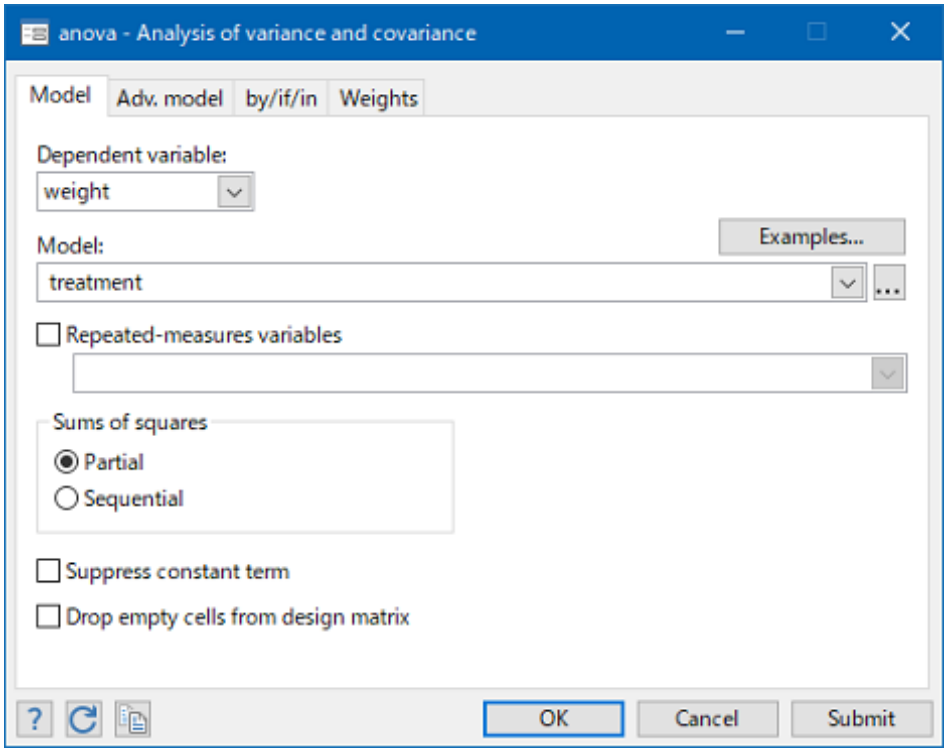

図 1　anova ダイアログ- Model タブ

```
. anova weight treatment

                          Number of obs =         10    R-squared     =  0.9147
                          Root MSE      =    9.07002    Adj R-squared =  0.8721

                   Source | Partial SS         df         MS        F     Prob>F
                ----------+----------------------------------------------------
                    Model |  5295.5443          3   1765.1814     21.46  0.0013
                          |
                treatment |  5295.5443          3   1765.1814     21.46  0.0013
                          |
                 Residual |  493.59167          6   82.265278
                ----------+----------------------------------------------------
                    Total |   5789.136          9   643.23733
```

p 値は 0.0013 であるため、有意水準を 1% としたとしても濃度 (`treatment`) の違いによる差は有意であると言えます。

ANOVA 表の上部に示されているのはベースとなった回帰の結果を集約したものです。フィットに用いられた観測データの数は 10 であり、平均二乗誤差の平方根 (root MSE: root mean squared error) は 9.07 とレポートされています。またモデルの R^2 値は 0.9147、調整済み R^2 値は 0.8721 であることがわかります。

ANOVA 表の 1 行目はモデルに関する情報を示しています。モデルに関する平方和 (`Partial SS`) は 5295.5 であり、その自由度 (`df`) は 3 と示されています。その結果、平均平方 (`MS`) は $5295.5/3 \approx 1765.2$ と算出されます。対応する F 統計量は 21.46 であり、その有意水準は 0.0013 となります。すなわち、モデルは 0.13% のレベルで有意と言えるわけです。

2 行目はモデル中における最初の項である `treatment` について情報を集約したものです。ここで設定したモデルでは 1 つの項しか存在しないため、2 行目の内容はモデル全般に関する 1 行目の内容と同一となります。

3 行目は残差に関する情報を示しています。残差平方和は 493.59、自由度は 6、平均平方は 82.27 とレポートされています。`Root MSE` としてレポートされている 9.07 という値は $\sqrt{82.27}$ として算出されたものです。

モデルの平方和と残差の平方和を足したものが全平方和 (total sum of squares) であり、ANOVA 表の最下行に 5789.1 とレポートされています。これは平均値を除去した後の `weight` の平方和の合計値を意味します。同様にモデルの自由度と残差の自由度を合計したものが総自由度 (total degrees of freedom) であり、この例では 9 となっています。これは観測データの総数 10 から平均値に対応する 1 を引いた値であるわけです。 ◁

▷ Example 2: **線形回帰モデル**

Example 1 の結果からすると化学肥料の濃度が果実の平均的な重さに対し有意な影響を及ぼしていることが確認できたわけですが、次の段階としてはどの濃度が最も効果的であったかが知りたくなります。それを知る 1 つの方法は ANOVA の基盤をなす回帰係数を調べてみることです。`anova` 実行後に `regress` と入力すればその情報を確認することができます。

```
. regress
```

```
. regress

      Source |       SS           df       MS      Number of obs   =        10
-------------+----------------------------------   F(3, 6)         =     21.46
       Model |  5295.54433         3  1765.18144   Prob > F        =    0.0013
    Residual |  493.591667         6  82.2652778   R-squared       =    0.9147
-------------+----------------------------------   Adj R-squared   =    0.8721
       Total |    5789.136         9  643.237333   Root MSE        =      9.07

------------------------------------------------------------------------------
      weight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   treatment |
           2 |  -59.16667   7.405641    -7.99   0.000    -77.28762   -41.04572
           3 |     -33.25   8.279758    -4.02   0.007    -53.50984   -12.99016
           4 |      -34.4   8.279758    -4.15   0.006    -54.65984   -14.14016
             |
       _cons |       111.9   5.236579    21.37   0.000     99.08655    124.7134
------------------------------------------------------------------------------
```

この情報から基盤となった線形回帰モデルが

```
. regress weight i.treatment
```

であったことがわかります（`i.` 演算子については *mwp-028* を参照ください）。この場合、ベースレベルである `treatment` $= 1$ に比べ、`treatment` $= 2, 3, 4$ では `weight` に対する効果がそれぞれ 59.2, 33.3. 34.4 だけ減ずることがわかります。つまり濃度 1 が最も重い果実を、濃度 2 が最も軽い果実を、濃度 3, 4 がその中間的な果実をもたらしたというわけです。 ◁

▷ Example 3: ANOVA **の再表示**

評価版では割愛しています。

3. **二元配置** ANOVA

評価版では割愛しています。

4. N **元配置** ANOVA

評価版では割愛しています。

5. **重み付きデータ**

評価版では割愛しています。

6. ANCOVA

評価版では割愛しています。

7. **ネスト型デザイン**

評価版では割愛しています。

8. **混合型デザイン**

評価版では割愛しています。

9. **ラテン方格デザイン**

評価版では割愛しています。

10. **反復測定** ANOVA

評価版では割愛しています。

補足１ – 繰返しのない二元配置 ANOVA

評価版では割愛しています。

■