重回帰分析

青木繁伸

2020 年 3 月 17 日

1　目的

重回帰分析を行う。

2　使用法

```
import sys
sys.path.append("statlib")
from multi import mreg
mreg(dat, tolerance = False, make_dummy=False, verbose=True)
```

予測値，標準化残差，偏回帰係数などをプロットする。

```
import sys
sys.path.append("statlib")
from multi import mreg_plot
mreg_plot(obj, type="p", color="blue", alpha=0.3)
```

2.1　引数

dat	従属変数を最後の列に置いたデータフレーム
tolerance	torelance を出力する場合に True にする デフォルトでは VIF を出力する
make_dummy	ダミー変数への変換が必要な場合には True を指定する。デフォルトは False
verbose	必要最小限のプリント出力をする（デフォルトは True）
mute	デフォルト（True）では予測値はプリント表示しない。プリント表示が必要な場合は False にする。
obj	mreg() の戻り値
type	"p"（デフォルト）の場合は実測値と予測値のプロット。 "e" の場合は実測値と標準化残差のプロット。 "c" の場合は偏回帰係数のプロット。 "b" の場合は標準化偏回帰係数のプロット。
color	点または棒の色

alpha　　アルファチャネル

2.2 戻り値の名前

"result"	偏回帰係数，標準誤差，t 値，p 値, 標準化偏回帰係数，VIF
"anova"	回帰の分散分析
"stderr"	残差標準誤差
"dfe"	残差標準誤差の自由度
"R"	重相関係数
"R2"	決定係数（重相関係数の 2 乗）
"R2s"	自由度調整済み重相関係数の二乗
"F"	回帰の分散分析の F 値
"dfr"	第 1 自由度（第 2 自由度は "dfe"
"p"	F の p 値
"loglik"	対数尤度
"AIC"	AIC
"beta"	標準化偏回帰係数
"B"	偏回帰係数
"const"	定数項
"name"	独立変数の名前
"observedvalues"	実測値
"fittedvalues"	予測値
"residuals"	残差
"stdres"	標準化残差

3 使用例

```
import pandas as pd

dat = pd.read_csv("data/mreg.csv")
print(dat.head())
```

```
     X1    X2    X3    X4    X5    X6    X7    X8    X9   X10   X11   X12  \
0  71.5  39.5  71.6  46.1  41.9  65.6  36.7  50.5  46.4  63.4  68.5  42.7
1  38.4  40.9  47.2  50.9  53.9  31.8  33.1  47.4  44.6  58.4  55.7  43.0
2  49.8  54.2  41.7  27.0  72.8  50.3  43.9  36.3  28.6  29.6  48.0  40.2
3  53.0  42.3  71.9  64.7  65.2  59.9  48.2  48.2  43.2  61.0  66.2  81.5
4  31.2  26.9  33.9  54.0  68.3  38.4  45.1  71.0  54.5  51.0  43.8  33.7

      Y
0  75.7
1  32.5
```

```
2  12.4
3  72.9
4  11.4
```

3.1 VIF を出力する

```
import sys
sys.path.append("statlib")
from multi import mreg

a = mreg(dat)
```

```
Coefficients
           Estimate  Std. Error   t value Pr(> |t|)      beta      VIF
X1         0.412004    0.054423  7.570410  < 0.0001   0.26804  3.44236
X2        -0.111885    0.041141 -2.719546    0.0079  -0.07279  1.96704
X3         0.319443    0.050622  6.310382  < 0.0001   0.20786  2.97944
X4         0.433128    0.058125  7.451685  < 0.0001   0.28190  3.92989
X5        -0.394926    0.043465 -9.086134  < 0.0001  -0.25706  2.19797
X6         0.471159    0.055647  8.466993  < 0.0001   0.30660  3.60081
X7         0.164192    0.043884  3.741505    0.0003   0.10684  2.23915
X8        -0.150200    0.037658 -3.988524    0.0001  -0.09773  1.64871
X9         0.215244    0.041273  5.215104  < 0.0001   0.14004  1.98013
X10        0.081436    0.053246  1.529427    0.1298   0.05298  3.29532
X11        0.368669    0.061393  6.005041  < 0.0001   0.23980  4.37881
X12        0.235303    0.033327  7.060501  < 0.0001   0.15306  1.29054
constant -57.178256    7.045523 -8.115545  < 0.0001

Residual standard error: 2.93338 on 87 degrees of freedom
Multiple R: 0.98403, Multiple R-squared: 0.96832, Adjusted R-squared: 0.96395
F-statistic: 221.584 on 12 and 87 DF,  p-value: < 0.0001
loglik = -242.54617  AIC = 513.09235

Anova Table
            Df        Sum Sq      Mean Sq    F value    Pr(>F)
regression  12  22879.990759  1906.665897  221.58426  < 0.0001
residuals   87    748.608841     8.604699
total       99  23628.599600   238.672723
```

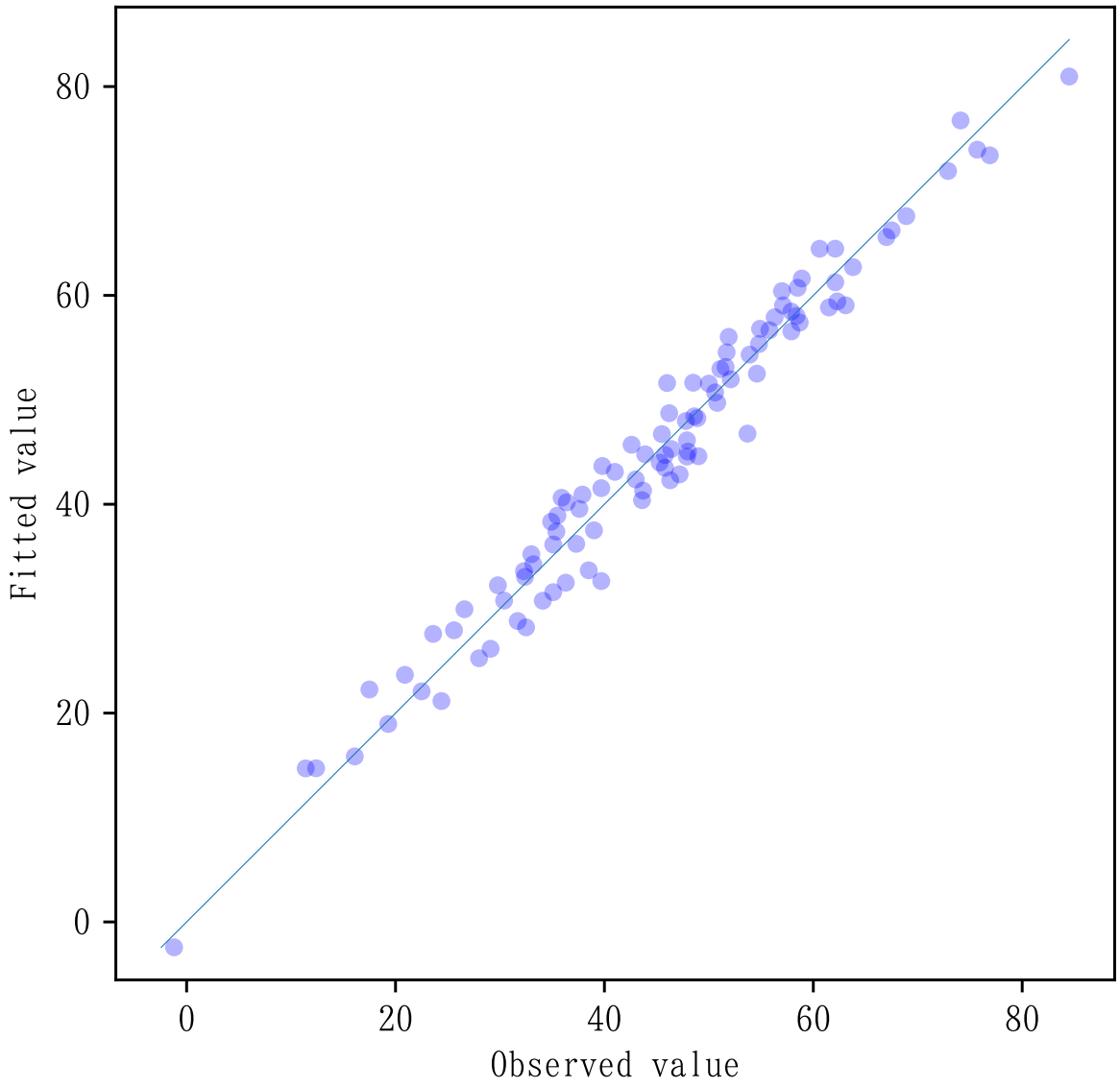
3.2 tolerance を出力する

```
a = mreg(dat, tolerance=True)
```

```
Coefficients
           Estimate  Std. Error   t value Pr(> |t|)      beta tolerance
X1         0.412004    0.054423  7.570410  < 0.0001   0.26804   0.29050
X2        -0.111885    0.041141 -2.719546    0.0079  -0.07279   0.50838
X3         0.319443    0.050622  6.310382  < 0.0001   0.20786   0.33563
X4         0.433128    0.058125  7.451685  < 0.0001   0.28190   0.25446
X5        -0.394926    0.043465 -9.086134  < 0.0001  -0.25706   0.45497
X6         0.471159    0.055647  8.466993  < 0.0001   0.30660   0.27772
X7         0.164192    0.043884  3.741505    0.0003   0.10684   0.44660
X8        -0.150200    0.037658 -3.988524    0.0001  -0.09773   0.60653
X9         0.215244    0.041273  5.215104  < 0.0001   0.14004   0.50502
X10        0.081436    0.053246  1.529427    0.1298   0.05298   0.30346
X11        0.368669    0.061393  6.005041  < 0.0001   0.23980   0.22837
X12        0.235303    0.033327  7.060501  < 0.0001   0.15306   0.77487
constant -57.178256    7.045523 -8.115545  < 0.0001

Residual standard error: 2.93338 on 87 degrees of freedom
Multiple R: 0.98403, Multiple R-squared: 0.96832, Adjusted R-squared: 0.96395
F-statistic: 221.584 on 12 and 87 DF,  p-value: < 0.0001
loglik = -242.54617  AIC = 513.09235

Anova Table
            Df        Sum Sq      Mean Sq    F value    Pr(>F)
regression  12  22879.990759  1906.665897  221.58426  < 0.0001
residuals   87    748.608841     8.604699
total       99  23628.599600   238.672723
```

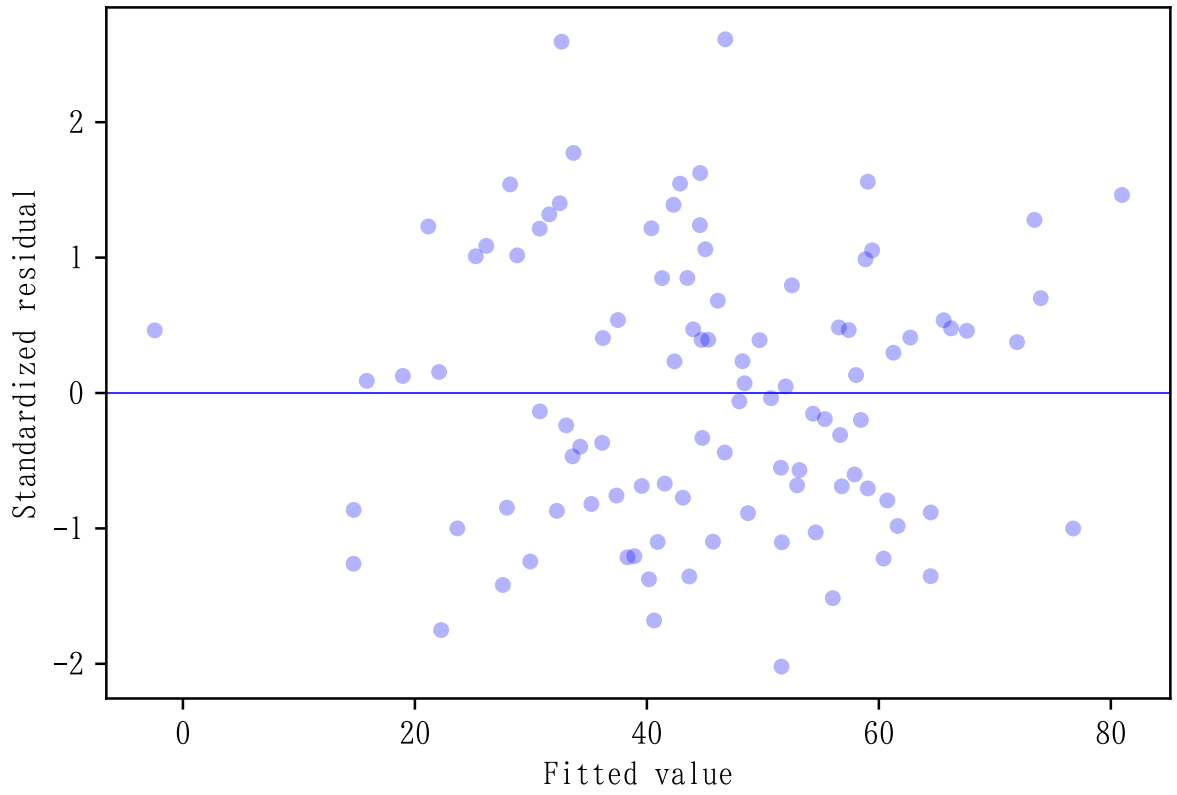
3.3 実測値と予測値のプロット

```
import sys
sys.path.append("statlib")
from multi import mreg_plot

mreg_plot(a)
```



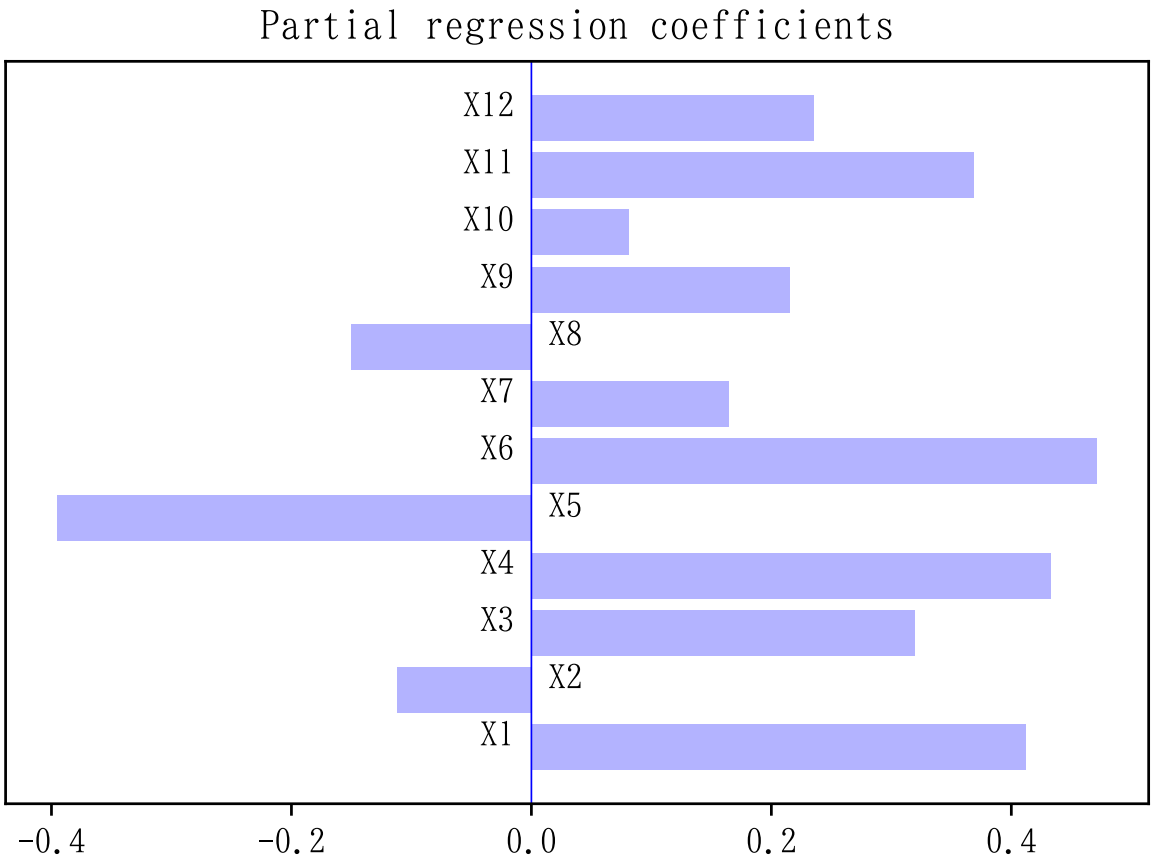
3.4 標準化残差のプロット

```
import sys
sys.path.append("statlib")
from multi import mreg_plot

mreg_plot(a, type="e")
```



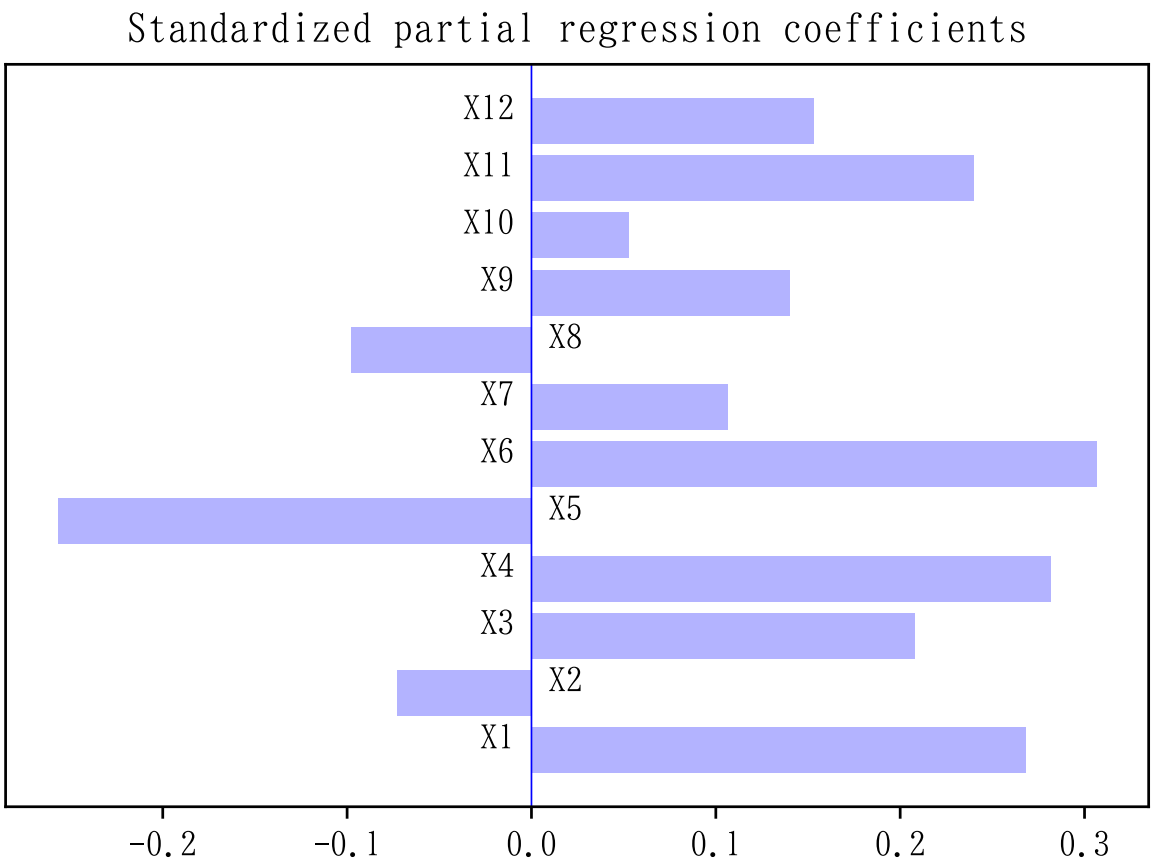
3.5 偏回帰係数の大きさのプロット

```
mreg_plot(a, type="c")
```



3.6 標準化偏回帰係数の大きさのプロット

```
mreg_plot(a, type="b")
```



3.7 ダミー変数を使う

独立変数の中にダミー変数に変換すべきものがある場合には，make_dummy=True を指定することでダミー変数を作る。

iris データセットの最後の列 "sp" は品種を表す 3 種類の文字列になっている。これを 2 個のダミー変数として重回帰分析に組み込む例を示す。

重回帰分析プログラムに渡すデータフレームは，最後の列が従属変数 "sl" になるように作る。

```
import pandas as pd

dat = pd.read_csv("data/iris.csv")
dat = pd.concat([dat.iloc[:,1:], dat.iloc[:,0]], axis=1)
print(dat.head())
```

```
    sw   pl   pw      sp   sl
0  3.5  1.4  0.2  setosa  5.1
1  3.0  1.4  0.2  setosa  4.9
2  3.2  1.3  0.2  setosa  4.7
3  3.1  1.5  0.2  setosa  4.6
4  3.6  1.4  0.2  setosa  5.0
```

```
import sys
```

```
sys.path.append("statlib")
from multi import mreg

a = mreg(dat, make_dummy=True)
```

Coefficients

	Estimate	Std. Error	t value	Pr(> |t|)	beta	VIF
sw	0.495889	0.086070	5.761466	< 0.0001	0.26102	2.22747
pl	0.829244	0.068528	12.100867	< 0.0001	1.76781	23.16165
pw	-0.315155	0.151196	-2.084418	0.0389	-0.29010	21.02140
sp_versicolor	-0.723562	0.240169	-3.012721	0.0031	-0.41329	20.42339
sp_virginica	-1.023498	0.333726	-3.066878	0.0026	-0.58461	39.43438
constant	2.171266	0.279794	7.760227	< 0.0001		

Residual standard error: 0.30683 on 144 degrees of freedom
Multiple R: 0.93130, Multiple R-squared: 0.86731, Adjusted R-squared: 0.86271
F-statistic: 188.251 on 5 and 144 DF, p-value: < 0.0001
loglik = -32.55801 AIC = 79.11602

Anova Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
regression	5	88.611848	17.722370	188.25095	< 0.0001
residuals	144	13.556485	0.094142		
total	149	102.168333	0.685694		