logistic-regression

March 23, 2022

[1]:
```
# math
import numpy as np

# dataframes
import pandas as pd

# plotting
import matplotlib.pyplot as plt
import seaborn

# linear regression two ways
import statsmodels.api as sm
from sklearn.linear_model import LinearRegression

# for choosing covariates to include in model
from patsy import dmatrices

# model selection
from sklearn.model_selection import train_test_split
from sklearn.feature_selection import SelectFromModel
from sklearn.linear_model import LassoCV
```

[3]:
```
diabetes = pd.read_csv('diabetes.csv')
diabetes.head()
```

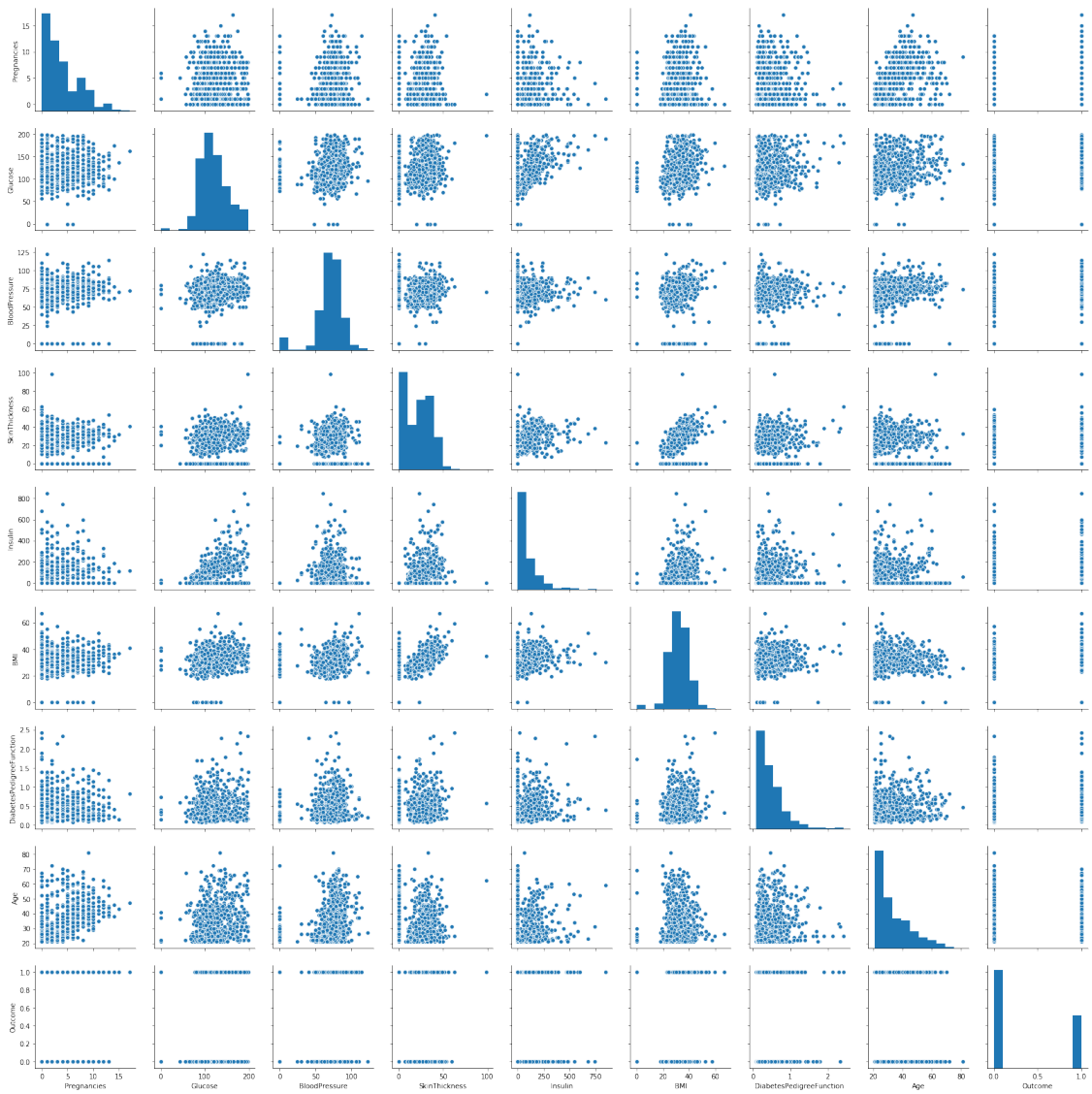
[3]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1			

```
4                    2.288    33          1
```

[4]:
```
seaborn.pairplot(diabetes)
```

[4]:
```
<seaborn.axisgrid.PairGrid at 0x7f89187572d0>
```



[20]:
```
y = diabetes['Outcome']
features = diabetes.columns[:-1]
X = sm.add_constant(diabetes[features])
X.head()
```

[20]:
```
   const  Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin   BMI  \
0    1.0            6      148             72             35        0  33.6
```

```
1    1.0            1       85             66             29        0  26.6
2    1.0            8      183             64              0        0  23.3
3    1.0            1       89             66             23       94  28.1
4    1.0            0      137             40             35      168  43.1

   DiabetesPedigreeFunction  Age
0                     0.627   50
1                     0.351   31
2                     0.672   32
3                     0.167   21
4                     2.288   33
```

[24]:
```
model = sm.Logit(y,X).fit()
model.summary()
```

```
Optimization terminated successfully.
         Current function value: 0.470993
         Iterations 6
```

[24]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                           Logit Regression Results
==============================================================================
Dep. Variable:                Outcome   No. Observations:                  768
Model:                          Logit   Df Residuals:                      759
Method:                           MLE   Df Model:                            8
Date:                Mon, 05 Apr 2021   Pseudo R-squ.:                  0.2718
Time:                        12:12:04   Log-Likelihood:                -361.72
converged:                       True   LL-Null:                       -496.74
Covariance Type:            nonrobust   LLR p-value:                 9.652e-54
==============================================================================
                                 coef    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------
const                          -8.4047      0.717    -11.728      0.000      -9.809
-7.000
Pregnancies                     0.1232      0.032      3.840      0.000       0.060
0.186
Glucose                         0.0352      0.004      9.481      0.000       0.028
0.042
BloodPressure                  -0.0133      0.005     -2.540      0.011      -0.024
-0.003
SkinThickness                   0.0006      0.007      0.090      0.929      -0.013
0.014
Insulin                        -0.0012      0.001     -1.322      0.186      -0.003
```

```
0.001
BMI                          0.0897      0.015      5.945      0.000       0.060
0.119
DiabetesPedigreeFunction     0.9452      0.299      3.160      0.002       0.359
1.531
Age                          0.0149      0.009      1.593      0.111      -0.003
0.033
==============================================================================
============
"""
```

[25]:
```
model.pvalues
```

[25]:
```
const                       9.161475e-32
Pregnancies                 1.229642e-04
Glucose                     2.509132e-21
BloodPressure               1.107208e-02
SkinThickness               9.285152e-01
Insulin                     1.860652e-01
BMI                         2.758957e-09
DiabetesPedigreeFunction    1.579980e-03
Age                         1.111920e-01
dtype: float64
```

[26]:
```
np.max(model.pvalues)
```

[26]:
```
0.9285152151977185
```

[28]:
```
variables = X.columns
variables
```

[28]:
```
Index(['const', 'Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness',
       'Insulin', 'BMI', 'DiabetesPedigreeFunction', 'Age'],
      dtype='object')
```

[31]:
```
maxp = np.max(model.pvalues)
newvariables = variables[model.pvalues < maxp]
newvariables
```

[31]:
```
Index(['const', 'Pregnancies', 'Glucose', 'BloodPressure', 'Insulin', 'BMI',
       'DiabetesPedigreeFunction', 'Age'],
      dtype='object')
```

0.1 model selection on training set, then model fit on test set

[21]:
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.5,␣
↪random_state=0)
```

[36]:
```
def minAIC(X,y):
    variables = X.columns
    model = sm.Logit(y,X[variables]).fit()
    while True:
        maxp = np.max(model.pvalues)
        newvariables = variables[model.pvalues < maxp]
        newmodel = sm.Logit(y,X[newvariables]).fit()
        if newmodel.aic < model.aic:
            model = newmodel
            variables = newvariables
        else:
            break
    return model,variables
```

[37]:
```
# select on training set, fit on test set
model,variables = minAIC(X_train, y_train)
model = sm.Logit(y_test,X_test[variables]).fit()
model.summary()
```

```
Optimization terminated successfully.
         Current function value: 0.487101
         Iterations 6
Optimization terminated successfully.
         Current function value: 0.487423
         Iterations 6
Optimization terminated successfully.
         Current function value: 0.488511
         Iterations 6
Optimization terminated successfully.
         Current function value: 0.489768
         Iterations 6
Optimization terminated successfully.
         Current function value: 0.491894
         Iterations 6
Optimization terminated successfully.
         Current function value: 0.501024
         Iterations 6
Optimization terminated successfully.
         Current function value: 0.462374
         Iterations 6
```

[37]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                           Logit Regression Results
==============================================================================
Dep. Variable:                Outcome   No. Observations:                  384
Model:                          Logit   Df Residuals:                      379
Method:                           MLE   Df Model:                            4
Date:                Mon, 05 Apr 2021   Pseudo R-squ.:                  0.2796
Time:                        13:51:13   Log-Likelihood:                -177.55
converged:                       True   LL-Null:                       -246.45
Covariance Type:            nonrobust   LLR p-value:                 8.376e-29
==============================================================================
============
                               coef    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------
------------
const                        -8.8682      0.986     -8.992      0.000     -10.801
-6.935
Pregnancies                   0.1583      0.040      3.938      0.000       0.079
0.237
Glucose                       0.0367      0.005      7.546      0.000       0.027
0.046
BMI                           0.0808      0.020      4.023      0.000       0.041
0.120
DiabetesPedigreeFunction      0.7261      0.419      1.732      0.083      -0.096
1.548
==============================================================================
============
"""
```

1 Generalized linear models

[12]:
```
ingot = pd.read_csv('ingots.csv')
# NR means not ready
ingot
```

[12]:

	soak	heat	NR	n
0	1.0	7	0	10
1	1.0	14	0	31
2	1.0	27	1	56
3	1.0	51	3	13
4	1.7	7	0	17
5	1.7	14	0	43
6	1.7	27	4	44
7	1.7	51	0	1

```
8    2.2     7   0   7
9    2.2    14   2  33
10   2.2    27   0  21
11   2.2    51   0   1
12   2.8     7   0  12
13   2.8    14   0  31
14   2.8    27   1  22
15   4.0     7   0   9
16   4.0    14   0  19
17   4.0    27   1  16
18   4.0    51   0   1
```

[13]:
```
ingot['failure'] = ingot['NR']
ingot['success'] = ingot['n'] - ingot['NR']
ingot.head()
```

[13]:
```
   soak  heat  NR   n  failure  success
0   1.0     7   0  10        0       10
1   1.0    14   0  31        0       31
2   1.0    27   1  56        1       55
3   1.0    51   3  13        3       10
4   1.7     7   0  17        0       17
```

[14]:
```
y = ingot[['success', 'failure']]
X = sm.add_constant(ingot[['heat', 'soak']])
```

[15]:
```
model = sm.GLM(y, X, family=sm.families.Binomial(), intercept=True).fit()
model.summary()
```

[15]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                 Generalized Linear Model Regression Results
==============================================================================
==
Dep. Variable:     ['success', 'failure']   No. Observations:
19
Model:                                GLM   Df Residuals:
16
Model Family:                    Binomial   Df Model:
2
Link Function:                      logit   Scale:
1.0000
Method:                              IRLS   Log-Likelihood:
-14.040
Date:                   Mon, 05 Apr 2021   Deviance:
13.753
Time:                           10:59:47   Pearson chi2:
```

```
13.5
No. Iterations:                          7
Covariance Type:                 nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          5.5592      1.120      4.965      0.000       3.365       7.754
heat          -0.0820      0.024     -3.456      0.001      -0.129      -0.036
soak          -0.0568      0.331     -0.171      0.864      -0.706       0.592
==============================================================================
"""
```