로지스틱회귀 + 더미변수

```
from pandas import read_excel, DataFrame, merge
from matplotlib import pyplot as plt
import seaborn as sb
import numpy as np
from patsy import dmatrix
import sys
import os

sys.path.append(os.path.dirname(os.path.dirname(os.getcwd())))
from helper import my_logit, scalling
```

데이터 가져오기

```
df = read_excel("https://data.hossam.kr/E05/gradeuate.xlsx")
df.head()
```

	합격여부	필기점수	학부성적	병원경력
0	0	380	3.61	3
1	1	660	3.67	3
2	1	800	4.00	1

로지스틱회귀 + 더미변수

데이터 가져오기

표준화 적용

	합격여부	필기점수	학부성적	병원경력
3	1	640	3.19	4
4	0	520	2.93	4

```
dv = dmatrix('C(병원경력)', df)
dv
```

```
DesignMatrix with shape (400, 4)
  Intercept  C(병원경력)[T.2]  C(병원경력)[T.3]  C(병원경력)[T.4]
          1               0               1               0
          1               0               1               0
          1               0               0               0
          1               0               0               1
          1               0               0               1
          1               1               0               0
          1               0               0               0
          1               1               0               0
          1               0               1               0
          1               1               0               0
          1               0               0               1
          1               0               0               0
          1               0               0               0
          1               1               0               0
          1               0               0               0
          1               0               1               0
          1               0               0               1
          1               0               1               0
          1               1               0               0
```

로지스틱회귀 + 더미변수

데이터 가져오기

표준화 적용

```
        1           0           0           0
        1           0           1           0
        1           1           0           0
        1           0           0           1
        1           0           0           1
        1           1           0           0
        1           0           0           0
        1           0           0           0
        1           0           0           1
        1           1           0           0
        1           0           0           0
[370 rows omitted]
Terms:
  'Intercept' (column 0)
  'C(병원경력)' (columns 1:4)
(to view full data, use np.asarray(this_obj))
```

```
dummy_df = DataFrame(np.asarray(dv))
dummy_df.drop(0, axis=1, inplace=True)
dummy_df.rename(columns={1: '고수', 2: '중수', 3: '하수'}, inplace=True)
dummy_df.head()
```

	고수	중수	하수
0	0.0	1.0	0.0
1	0.0	1.0	0.0
2	0.0	0.0	0.0

로지스틱회귀 + 더미변수

데이터 가져오기

표준화 적용

	고수	중수	하수
3	0.0	0.0	1.0
4	0.0	0.0	1.0

```
mdf = merge(df.drop('병원경력', axis=1), dummy_df, left_index=True, right
mdf.head()
```

	합격여부	필기점수	학부성적	고수	중수	하수
0	0	380	3.61	0.0	1.0	0.0
1	1	660	3.67	0.0	1.0	0.0
2	1	800	4.00	0.0	0.0	0.0
3	1	640	3.19	0.0	0.0	1.0
4	0	520	2.93	0.0	0.0	1.0

```
logit_result = my_logit(mdf, y='합격여부', x=['필기점수','학부성적','고수','
print(logit_result.summary)
```

```
Optimization terminated successfully.
         Current function value: 0.573147
         Iterations 6
                           Logit Regression Results
```

로지스틱회귀 + 더미변수

데이터 가져오기

표준화 적용

```
==============================================================================
Dep. Variable:                  합격여부   No. Observations:
Model:                          Logit   Df Residuals:
Method:                           MLE   Df Model:
Date:                Mon, 31 Jul 2023   Pseudo R-squ.:                  0
Time:                        14:31:24   Log-Likelihood:                 -
converged:                       True   LL-Null:                        -
Covariance Type:            nonrobust   LLR p-value:                  7.5
==============================================================================
                 coef    std err          z      P>|z|      [0.025
------------------------------------------------------------------------------
Intercept     -3.9900      1.140     -3.500      0.000      -6.224
필기점수          0.0023      0.001      2.070      0.038       0.000
학부성적          0.8040      0.332      2.423      0.015       0.154
고수           -0.6754      0.316     -2.134      0.033      -1.296
중수           -1.3402      0.345     -3.881      0.000      -2.017
하수           -1.5515      0.418     -3.713      0.000      -2.370
==============================================================================
```

```
logit_result.cmdf
```

	Negative	Positive
True	254	30
False	97	19

```
logit_result.odds_rate_df
```

	odds_rate
Intercept	0.018500
필기점수	1.002267
학부성적	2.234545
고수	0.508931
중수	0.261792
하수	0.211938

```
logit_result.result_df
```

	설명력 (Pseudo-Rsqe)	정확도 (Accuracy)	정밀도 (Precision)	재현율 (Recall, TPR)	위양성율 (Fallout, FPR)	특이성 (Specificity, TNR)	RAS
0	0.082922	0.71	0.612245	0.23622	0.069597	0.930403	0.583312

표준화 적용

```
sdf = scalling(mdf.filter(['필기점수', '학부성적']))
sdf.head()
```

로지스틱회귀 + 더미변수

데이터 가져오기

표준화 적용

	필기점수	학부성적
0	-1.800263	0.579072
1	0.626668	0.736929
2	1.840134	1.605143
3	0.453316	-0.525927
4	-0.586797	-1.209974

```
mdf['필기점수'] = sdf['필기점수']
mdf['학부성적'] = sdf['학부성적']
mdf.head()
```

	합격여부	필기점수	학부성적	고수	중수	하수
0	0	-1.800263	0.579072	0.0	1.0	0.0
1	1	0.626668	0.736929	0.0	1.0	0.0
2	1	1.840134	1.605143	0.0	0.0	0.0
3	1	0.453316	-0.525927	0.0	0.0	1.0
4	0	-0.586797	-1.209974	0.0	0.0	1.0

```
logit_result = my_logit(mdf, y='합격여부', x=['필기점수','학부성적','고수','
print(logit_result.summary)
```

로지스틱회귀 + 더미변수

데이터 가져오기

표준화 적용

```
Optimization terminated successfully.
         Current function value: 0.573147
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:                   합격여부   No. Observations:
Model:                          Logit   Df Residuals:
Method:                           MLE   Df Model:
Date:                Mon, 31 Jul 2023   Pseudo R-squ.:                  0
Time:                        14:35:50   Log-Likelihood:                -
converged:                       True   LL-Null:                       -
Covariance Type:            nonrobust   LLR p-value:                 7.5
==============================================================================
                 coef    std err          z      P>|z|      [0.025
------------------------------------------------------------------------------
Intercept      0.0664      0.266      0.250      0.802      -0.454
필기점수           0.2613      0.126      2.070      0.038       0.014
학부성적           0.3056      0.126      2.423      0.015       0.058
고수            -0.6754      0.316     -2.134      0.033      -1.296
중수            -1.3402      0.345     -3.881      0.000      -2.017
하수            -1.5515      0.418     -3.713      0.000      -2.370
==============================================================================
```

```
logit_result.result_df
```

로지스틱회귀 + 더미변수

데이터 가져오기

표준화 적용

	설명력 (Pseudo-Rsqe)	정확도 (Accuracy)	정밀도 (Precision)	재현율 (Recall, TPR)	위양성율 (Fallout, FPR)	특이성 (Specificity, TNR)	RAS
0	0.082922	0.71	0.612245	0.23622	0.069597	0.930403	0.583312