# 연습문제 1번 풀이

## 패키지 참조

```
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
```

## 문제 (1)

### 데이터 가져오기

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

| | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | DiabetesI |
|---|---|---|---|---|---|---|---|
| **0** | 6 | 148 | 72 | 35 | 0 | 33.6 | 0.627 |
| **1** | 1 | 85 | 66 | 29 | 0 | 26.6 | 0.351 |
| **2** | 8 | 183 | 64 | 0 | 0 | 23.3 | 0.672 |
| **3** | 1 | 89 | 66 | 23 | 94 | 28.1 | 0.167 |
| **4** | 0 | 137 | 40 | 35 | 168 | 43.1 | 2.288 |

## 데이터 전처리 없이 분석 수행

```
x = list(df.columns)
x.remove("Outcome")
x
```

```
['Pregnancies',
 'Glucose',
 'BloodPressure',
 'SkinThickness',
 'Insulin',
 'BMI',
 'DiabetesPedigreeFunction',
 'Age']
```

```
logit_result = my_logit(df, y="Outcome", x=x)
print(logit_result.summary)
```

```
Optimization terminated successfully.
         Current function value: 0.470993
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:                Outcome   No. Observations:
Model:                          Logit   Df Residuals:
Method:                           MLE   Df Model:
Date:                Tue, 01 Aug 2023   Pseudo R-squ.:
Time:                        10:14:04   Log-Likelihood:                -
converged:                       True   LL-Null:                       -
Covariance Type:            nonrobust   LLR p-value:                 9.6
==============================================================================
                                 coef    std err          z      P>|z|
------------------------------------------------------------------------------
Intercept                    -8.4047      0.717    -11.728      0.000
Pregnancies                   0.1232      0.032      3.840      0.000
Glucose                       0.0352      0.004      9.481      0.000
BloodPressure                -0.0133      0.005     -2.540      0.011
SkinThickness                 0.0006      0.007      0.090      0.929
Insulin                      -0.0012      0.001     -1.322      0.186
BMI                           0.0897      0.015      5.945      0.000
DiabetesPedigreeFunction      0.9452      0.299      3.160      0.002
```

```
Age                    0.0149      0.009      1.593      0.111
```

```
logit_result.cmdf
```

| | Negative | Positive |
|---|---|---|
| **True** | 445 | 156 |
| **False** | 112 | 55 |

```
logit_result.result_df
```

| | 설명력 (Pseudo-Rsqe) | 정확도 (Accuracy) | 정밀도 (Precision) | 재현율 (Recall, TPR) | 위양성율 (Fallout, FPR) | 특이성 (Specificity, TNR) | RAS |
|---|---|---|---|---|---|---|---|
| **0** | 0.27181 | 0.782552 | 0.739336 | 0.58209 | 0.11 | 0.89 | 0.736045 |

```
logit_result.odds_rate_df
```

| | odds_rate |
|---|---|
| **Intercept** | 0.000224 |

| | odds_rate |
|---|---|
| **Pregnancies** | 1.131091 |
| **Glucose** | 1.035789 |
| **BloodPressure** | 0.986792 |
| **SkinThickness** | 1.000619 |
| **Insulin** | 0.998809 |
| **BMI** | 1.093847 |
| **DiabetesPedigreeFunction** | 2.573276 |
| **Age** | 1.014980 |

## 표준화 적용하기

```
y_train = df.filter(['Outcome'])
y_train.head()
```

| | Outcome |
|---|---|
| **0** | 1 |
| **1** | 0 |
| **2** | 1 |
| **3** | 0 |

| | Outcome |
|---|---|
| 4 | 1 |

```
x_train = df.drop('Outcome', axis=1)
x_train.head()
```

| | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | DiabetesI |
|---|---|---|---|---|---|---|---|
| 0 | 6 | 148 | 72 | 35 | 0 | 33.6 | 0.627 |
| 1 | 1 | 85 | 66 | 29 | 0 | 26.6 | 0.351 |
| 2 | 8 | 183 | 64 | 0 | 0 | 23.3 | 0.672 |
| 3 | 1 | 89 | 66 | 23 | 94 | 28.1 | 0.167 |
| 4 | 0 | 137 | 40 | 35 | 168 | 43.1 | 2.288 |

```
x_train_std_df = scalling(x_train)
x_train_std_df.head()
```

| | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI |
|---|---|---|---|---|---|---|
| 0 | 0.639947 | 0.848324 | 0.149641 | 0.907270 | -0.692891 | 0.204013 |
| 1 | -0.844885 | -1.123396 | -0.160546 | 0.530902 | -0.692891 | -0.684422 |
| 2 | 1.233880 | 1.943724 | -0.263941 | -1.288212 | -0.692891 | -1.103255 |

| | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI |
|---|---|---|---|---|---|---|
| 3 | -0.844885 | -0.998208 | -0.160546 | 0.154533 | 0.123302 | -0.494043 |
| 4 | -1.141852 | 0.504055 | -1.504687 | 0.907270 | 0.765836 | 1.409746 |

```
result_df = merge(x_train_std_df, y_train, left_index=True, right_index=
result_df.head()
```

| | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI |
|---|---|---|---|---|---|---|
| 0 | 0.639947 | 0.848324 | 0.149641 | 0.907270 | -0.692891 | 0.204013 |
| 1 | -0.844885 | -1.123396 | -0.160546 | 0.530902 | -0.692891 | -0.684422 |
| 2 | 1.233880 | 1.943724 | -0.263941 | -1.288212 | -0.692891 | -1.103255 |
| 3 | -0.844885 | -0.998208 | -0.160546 | 0.154533 | 0.123302 | -0.494043 |
| 4 | -1.141852 | 0.504055 | -1.504687 | 0.907270 | 0.765836 | 1.409746 |

```
logit_result = my_logit(result_df, y="Outcome", x=x)
print(logit_result.summary)
```

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

Logit Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Outcome | No. Observations: | |
| Model: | Logit | Df Residuals: | |
| Method: | MLE | Df Model: | |
| Date: | Tue, 01 Aug 2023 | Pseudo R-squ.: | |
| Time: | 10:14:04 | Log-Likelihood: | - |
| converged: | True | LL-Null: | - |
| Covariance Type: | nonrobust | LLR p-value: | 9.6 |

| | coef | std err | z | P>\|z\| |
|---|---|---|---|---|
| Intercept | -0.8711 | 0.097 | -8.986 | 0.000 |
| Pregnancies | 0.4148 | 0.108 | 3.840 | 0.000 |
| Glucose | 1.1235 | 0.118 | 9.481 | 0.000 |
| BloodPressure | -0.2572 | 0.101 | -2.540 | 0.011 |
| SkinThickness | 0.0099 | 0.110 | 0.090 | 0.929 |
| Insulin | -0.1372 | 0.104 | -1.322 | 0.186 |
| BMI | 0.7068 | 0.119 | 5.945 | 0.000 |
| DiabetesPedigreeFunction | 0.3130 | 0.099 | 3.160 | 0.002 |
| Age | 0.1747 | 0.110 | 1.593 | 0.111 |

```
logit_result.result_df
```

연습문제 1번 풀이

패키지 참조

문제 (1)

데이터 가져오기

데이터 전처리 없이 분석 수행

표준화 적용하기

| | 설명력 (Pseudo-Rsqe) | 정확도 (Accuracy) | 정밀도 (Precision) | 재현율 (Recall, TPR) | 위양성율 (Fallout, FPR) | 특이성 (Specificity, TNR) | RAS |
|---|---|---|---|---|---|---|---|
| **0** | 0.27181 | 0.782552 | 0.739336 | 0.58209 | 0.11 | 0.89 | 0.736045 |