**Assignment #1.2 Logistic Regression**

|  |
| --- |
| **Submission Instructions**   * Submit the followingfiles through Blackboard:  1. The completed **answer sheet** provided on the last page.  * If you do not follow the instructions, your assignment will be counted late. |

**Before you start**

For this assignment, you’ll be working with the **advertising.csv** file and the **Lab1.2.py** script. This is a dataset from an online dating website. Build a logistic regression model to help the manager understand how gender of customers may influence their probability of clicking on Ad for dating restaurants.

**Part 1. Simple Logistic regression**

In this part, you will build a simple logistic regression model with only one independent variable, and estimate the model with Python.

* 1. Model Identification

1. What is the dependent variable and what is the explanatory variable in this business problem?

The dependent variable is ClickedonAd and the explanatory variable in this business problem is Male

1. What is the logistic regression model you can use to answer this business problem?

log(odds of clicking on ad) = β0 + β1\*Male

where Male is a binary variable representing gender (1 for male and 0 for female). The coefficient β1 represents the effect of gender on the odds of clicking on the ad.

1. Estimate your model with Python

DailyTimeSpentonSite Age AreaIncome DailyInternetUsage \

count 1000.000000 1000.000000 1000.000000 1000.000000

mean 65.000200 36.009000 55000.000080 180.000100

std 15.853615 8.785562 13414.634022 43.902339

min 32.600000 19.000000 13996.500000 104.780000

25% 51.360000 29.000000 47031.802500 138.830000

50% 68.215000 35.000000 57012.300000 183.130000

75% 78.547500 42.000000 65470.635000 218.792500

max 91.430000 61.000000 79484.800000 269.960000

Male ClickedonAd Educationlevel Workinghrsperweek

count 1000.000000 1000.00000 1000.000000 999.000000

mean 0.481000 0.50000 2.142000 40.018018

std 0.499889 0.50025 0.774878 11.958917

min 0.000000 0.00000 1.000000 20.000000

25% 0.000000 0.00000 2.000000 30.000000

50% 0.000000 0.50000 2.000000 45.000000

75% 1.000000 1.00000 3.000000 50.000000

max 1.000000 1.00000 3.000000 60.000000

Optimization terminated successfully.

Current function value: 0.692424

Iterations 3

Logit Regression Results

==============================================================================

Dep. Variable: ClickedonAd No. Observations: 1000

Model: Logit Df Residuals: 998

Method: MLE Df Model: 1

Date: Mon, 06 Mar 2023 Pseudo R-squ.: 0.001043

Time: 22:09:59 Log-Likelihood: -692.42

converged: True LL-Null: -693.15

Covariance Type: nonrobust LLR p-value: 0.2291

==============================================================================

coef std err z P>|z| [0.025 0.975]

------------------------------------------------------------------------------

const 0.0733 0.088 0.834 0.404 -0.099 0.245

Male -0.1523 0.127 -1.202 0.229 -0.401 0.096

==============================================================================

Optimization terminated successfully.

Current function value: 0.090904

Iterations 10

Logit Regression Results

==============================================================================

Dep. Variable: ClickedonAd No. Observations: 1000

Model: Logit Df Residuals: 994

Method: MLE Df Model: 5

Date: Mon, 06 Mar 2023 Pseudo R-squ.: 0.8689

Time: 22:09:59 Log-Likelihood: -90.904

converged: True LL-Null: -693.15

Covariance Type: nonrobust LLR p-value: 3.136e-258

========================================================================================

coef std err z P>|z| [0.025 0.975]

----------------------------------------------------------------------------------------

const 27.3606 2.736 9.999 0.000 21.997 32.724

DailyTimeSpentonSite -0.1927 0.021 -9.286 0.000 -0.233 -0.152

Age 0.1709 0.026 6.607 0.000 0.120 0.222

AreaIncome -0.0001 1.88e-05 -7.245 0.000 -0.000 -9.93e-05

DailyInternetUsage -0.0635 0.007 -9.390 0.000 -0.077 -0.050

Male -0.4217 0.404 -1.043 0.297 -1.214 0.371

========================================================================================

Possibly complete quasi-separation: A fraction 0.23 of observations can be

perfectly predicted. This might indicate that there is complete

quasi-separation. In this case some parameters will not be identified.

**Part 2. Multiple Logistic Regression**

1. Look at the data file. What other factors in the table you also want to include in the model to avoid omitted variable bias? Why

* Working hours per week: Occupations with people who work more hours in a day compared to others are less likely to click on the ad due to time constraints.
* Education level: Customers with different educational levels may have different preferences.

1. Modify your regression model by adding more independent variables. Write down your new models.

ClickedonAd = β0 + β1(DailyTimeSpentonSite) + β2(Age) + β3(AreaIncome) + β4(DailyInternetUsage) + β5(Male) + β6(EducationLevel) + β7(Workinghoursperweek)

1. Estimate your new model in Python. Which variables are statistically significant and interpret the coefficients of these variable.

Education level categories: High School (1), Bachelors (2), Master and above(3) Code did not produce data

1. Can you confidently say that this model tested the “effect” of gender on customers’ probability of clicking on Ad? Why?

Since the coefficient for the male variable was not statistically significant, we cannot reject the null hypothesis.