# Logistic Regression

## Using excel

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT | |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
| Regression Statistics | |  |  |  |  |  |  |  |  |
| Multiple R | 0.1455325 |  |  |  |  |  |  |  |  |
| R Square | 0.0211797 |  |  |  |  |  |  |  |  |
| Adjusted R Square | 0.0211665 |  |  |  |  |  |  |  |  |
| Standard Error | 0.4466794 |  |  |  |  |  |  |  |  |
| Observations | 74343 |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |  |  |  |
|  | df | SS | MS | F | Significance F |  |  |  |  |
| Regression | 1 | 320.9498685 | 320.95 | 1608.590206 | 0 |  |  |  |  |
| Residual | 74341 | 14832.69889 | 0.1995 |  |  |  |  |  |  |
| Total | 74342 | 15153.64876 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  | Coefficients | Standard Error | t Stat | P-value | Lower 95% | Upper 95% | Lower 95.0% | Upper 95.0% |  |
| Intercept | 0.6402046 | 0.002479678 | 258.18 | P-value | 0.635344471 | 0.6450648 | 0.6353445 | 0.6450648 |  |
| male | 0.1324836 | 0.003303234 | 40.107 | 0 | 0.126009259 | 0.1389579 | 0.1260093 | 0.1389579 |  |

## Using R script

| | Coefficient| Std\_Error| Z\_Value| P\_Value |

|:-----------|-----------:|---------:|--------:|-------:|

|(Intercept) | 0.5762524| 0.0115668| 49.81964| 0|

|sex | 0.6472697| 0.0164223| 39.41409| 0|

1. setwd("A:/MA. Program/Semester 4/Thesis/Analyzing Data/r\_language")

2. library(readxl)

3. library(stats)

4. library(knitr)

5.

6. mydata <- read\_excel("education\_1.xlsx")

7.

8. model <- glm(label ~ sex, data = mydata, family = binomial(link = "logit"))

9.

10. coefficients <- coef(model)

11. std\_errors <- sqrt(diag(vcov(model)))

12. z\_values <- coefficients / std\_errors

13. p\_values <- 2 \* (1 - pnorm(abs(z\_values)))

14.

15. # Create a data frame to store the results

16. result <- data.frame(Coefficient = coefficients,

17. Std\_Error = std\_errors,

18. Z\_Value = z\_values,

19. P\_Value = p\_values)

20.

21. # Print the result table using knitr::kable()

22. print(kable(result, format = "markdown"))

## Using Python

Dep. Variable: label No. Observations: 74342

Model: Logit Df Residuals: 74340

Method: MLE Df Model: 1

Date: Mon, 17 Apr 2023 Pseudo R-squ.: 0.01763

Time: 08:24:15 Log-Likelihood: -43656.

converged: True LL-Null: -44439.

Covariance Type: nonrobust LLR p-value: 0.000

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

const 0.5761 0.012 49.809 0.000 0.553 0.599

sex 0.6473 0.016 39.415 0.000 0.615 0.679

1. import pandas as pd

2. import statsmodels.api as sm

3. import numpy as np

4. import matplotlib.pyplot as plt

5. df = pd.read\_excel('../data\_processed/education\_1.xlsx')

6. X = df[['sex']]

7. y = df['label']

8. # Add a constant term to the predictor variables

9. X = sm.add\_constant(X)

10. # Fit the logistic regression model

11. model = sm.Logit(y, X).fit()

12. # Print the summary

13. print(model.summary())

14.

1. const = model.params['const']

2. sex = model.params['sex']

3. x\_min = df['sex'].min()

4. x\_max = df['sex'].max()

5. x = np.linspace(x\_min, x\_max, 100)

6. y = 1 / (1 + np.exp(-const - sex \* x))

7. plt.plot(x, y)

8. plt.xlabel('Sex')

9. plt.ylabel('Probability')

10. plt.title('Logistic Regression Curve')

11. plt.show()

12.

