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|
```

```r
setwd("A:/MA. Program/Semester 4/Thesis/Analyzing Data/r_language")
library(readxl)
library(stats)
library(knitr)

mydata <- read_excel("education_1.xlsx")

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

coefficients <- coef(model)
std_errors <- sqrt(diag(vcov(model)))
z_values <- coefficients / std_errors
p_values <- 2 * (1 - pnorm(abs(z_values)))

# Create a data frame to store the results
result <- data.frame(Coefficient = coefficients,
                     Std_Error = std_errors,
                     Z_Value = z_values,
                     P_Value = p_values)

# Print the result table using knitr::kable()
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
==============================================================================================
```

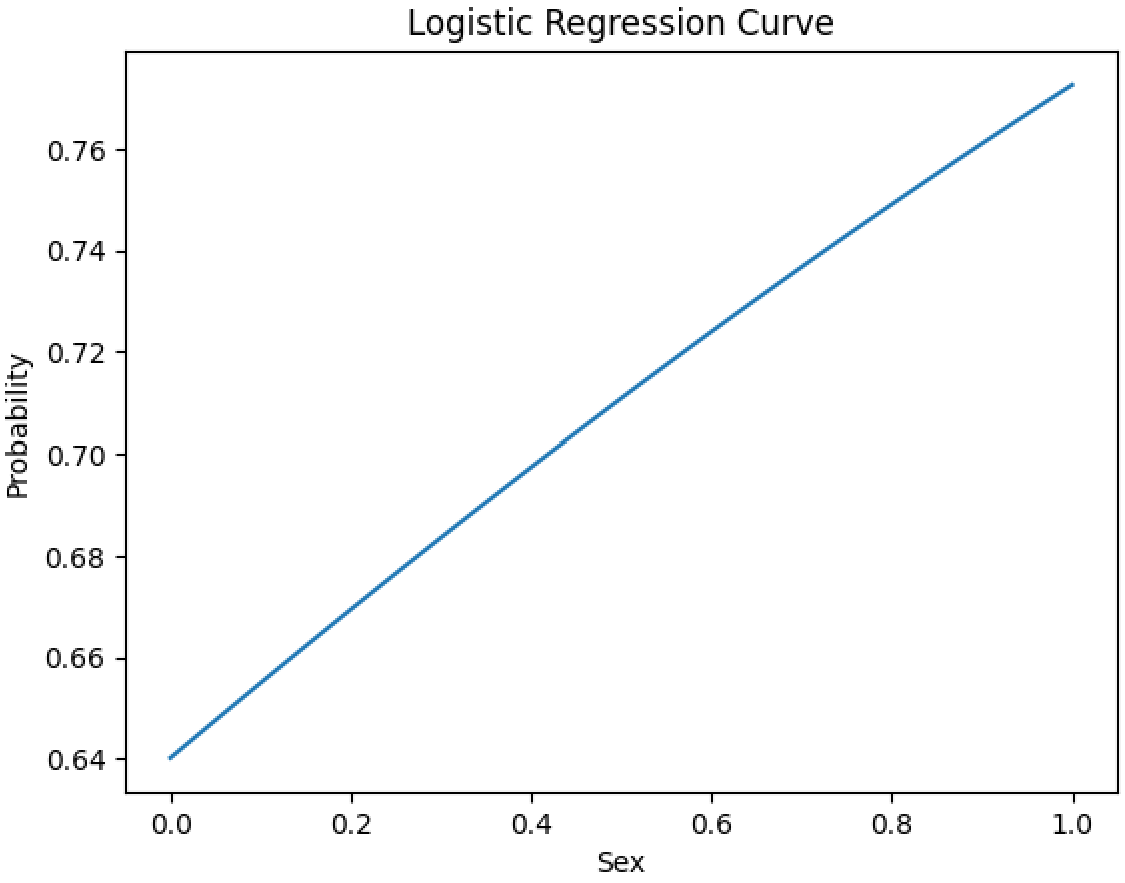
```python
import pandas as pd
import statsmodels.api as sm
import numpy as np
import matplotlib.pyplot as plt
df = pd.read_excel('../data_processed/education_1.xlsx')
X = df[['sex']]
y = df['label']
# Add a constant term to the predictor variables
X = sm.add_constant(X)
# Fit the logistic regression model
model = sm.Logit(y, X).fit()
# Print the summary
print(model.summary())

```

```python
const = model.params['const']
sex = model.params['sex']
x_min = df['sex'].min()
x_max = df['sex'].max()
```

```
x = np.linspace(x_min, x_max, 100)
y = 1 / (1 + np.exp(-const - sex * x))
plt.plot(x, y)
plt.xlabel('Sex')
plt.ylabel('Probability')
plt.title('Logistic Regression Curve')
plt.show()

```


Logistic Regression Curve
Probability
Sex