MBA 652

Statistical Methods in Business Analytics

Estimation of Factors that Determine the Purchase of Eco-labelled Apples

GROUP 16

Ankit Tyagi	231250025
Debdutta Singha Roy	231250039
Shubham Malviya	231250131
Kantule Ritesh Ramdas	210488
Jyoti Tewari	241140605
Praveen Sahu	241290014
Kavita Omar	241290009

Problem Statement

In this study, we aim to understand whether a family will buy eco-labeled apples and what factors influence their decision.

The study aims to:

1. Identify the proportion of families willing to buy eco-labelled apples at given prices.
2. Analyze the impact of various price and non-price variables in the decision.
3. Determine the significance of non-price factors in influencing the decision.
4. Access the predictive accuracy of various models and identify a good-fit model.

Reference

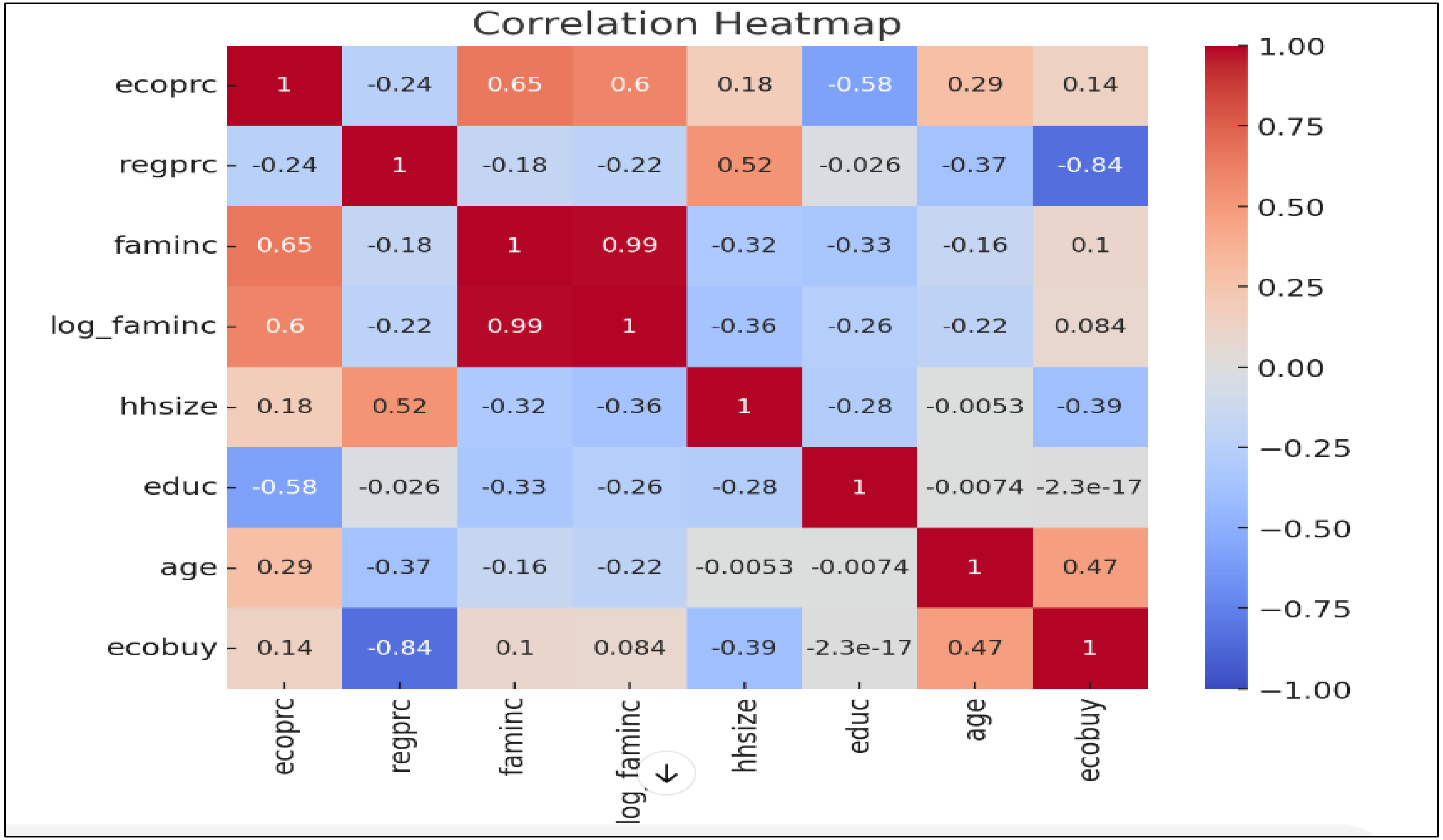
- The dataset used is based on the doctoral dissertation of Jeffrey Blend, Department of Agricultural Economics, Michigan State University, 1998.
- The dataset has 660 observations on 17 variables.

Key Variables

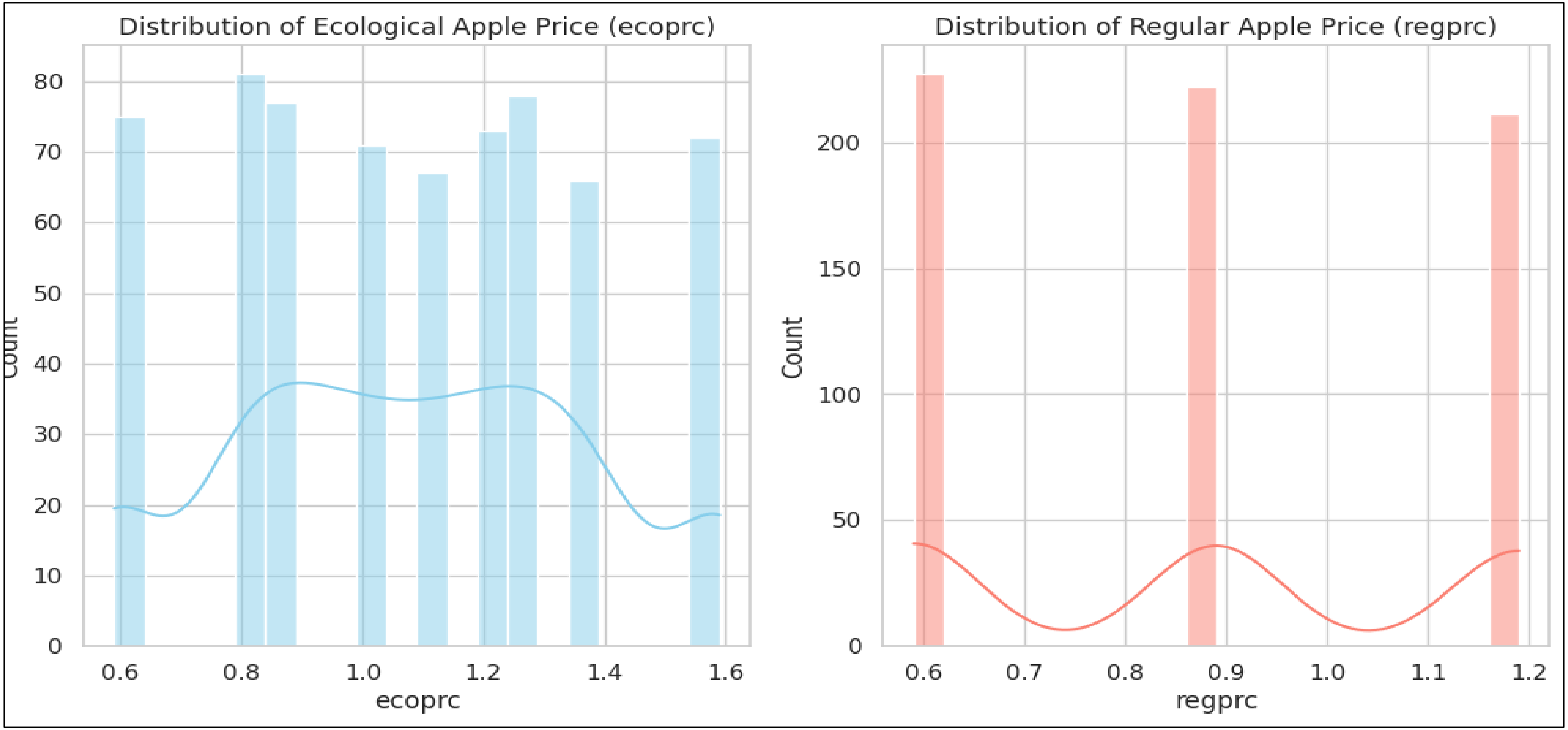
- ecobuy (Dependent Variable)
- regprc: price of regular apples
- ecoprc: price of eco-labeled apples
- educ: years of schooling
- hhsize: household size
- faminc: family income, thousands of dollars
- age: in years

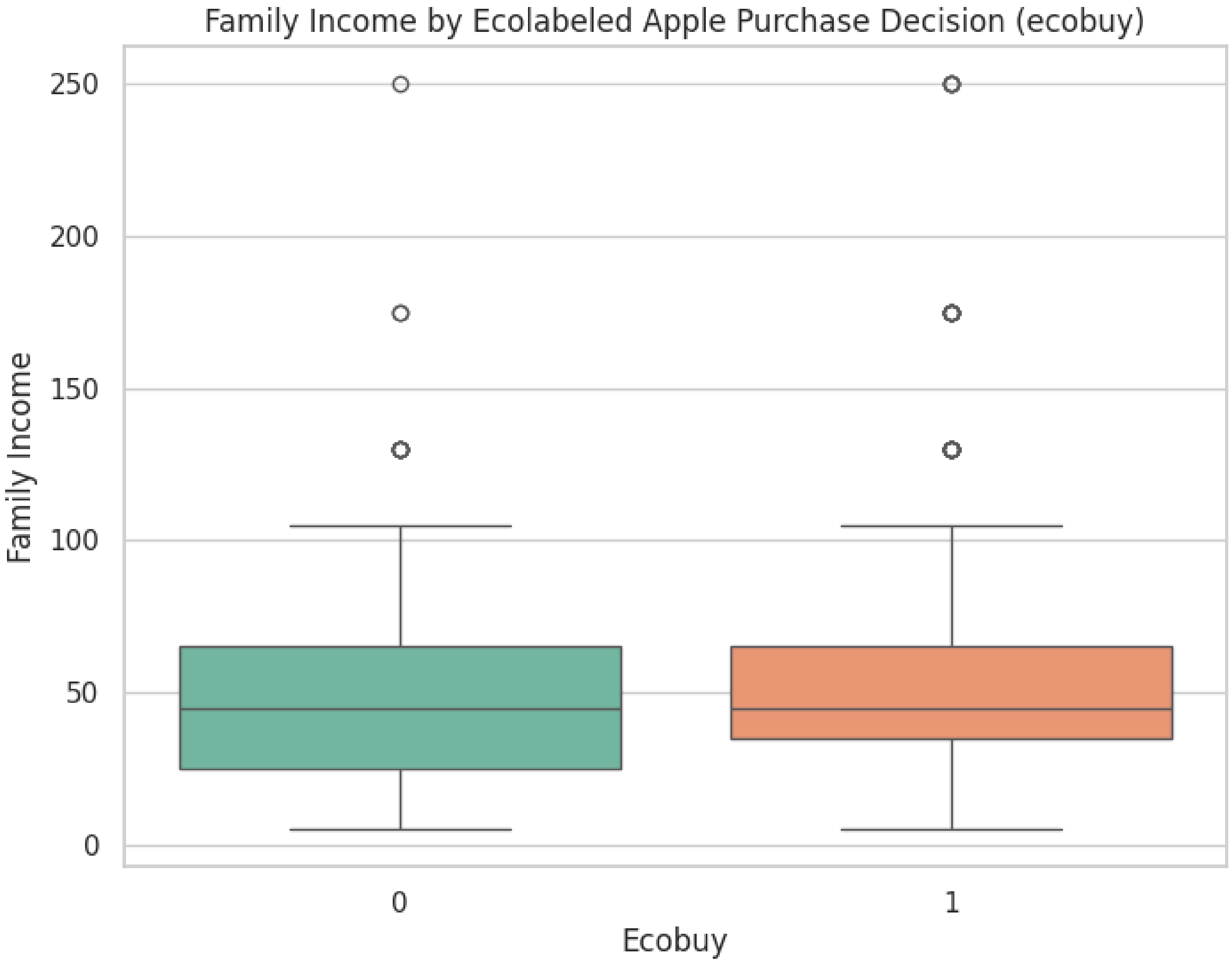
Exploratory Data Analysis

(Correlation matrix)



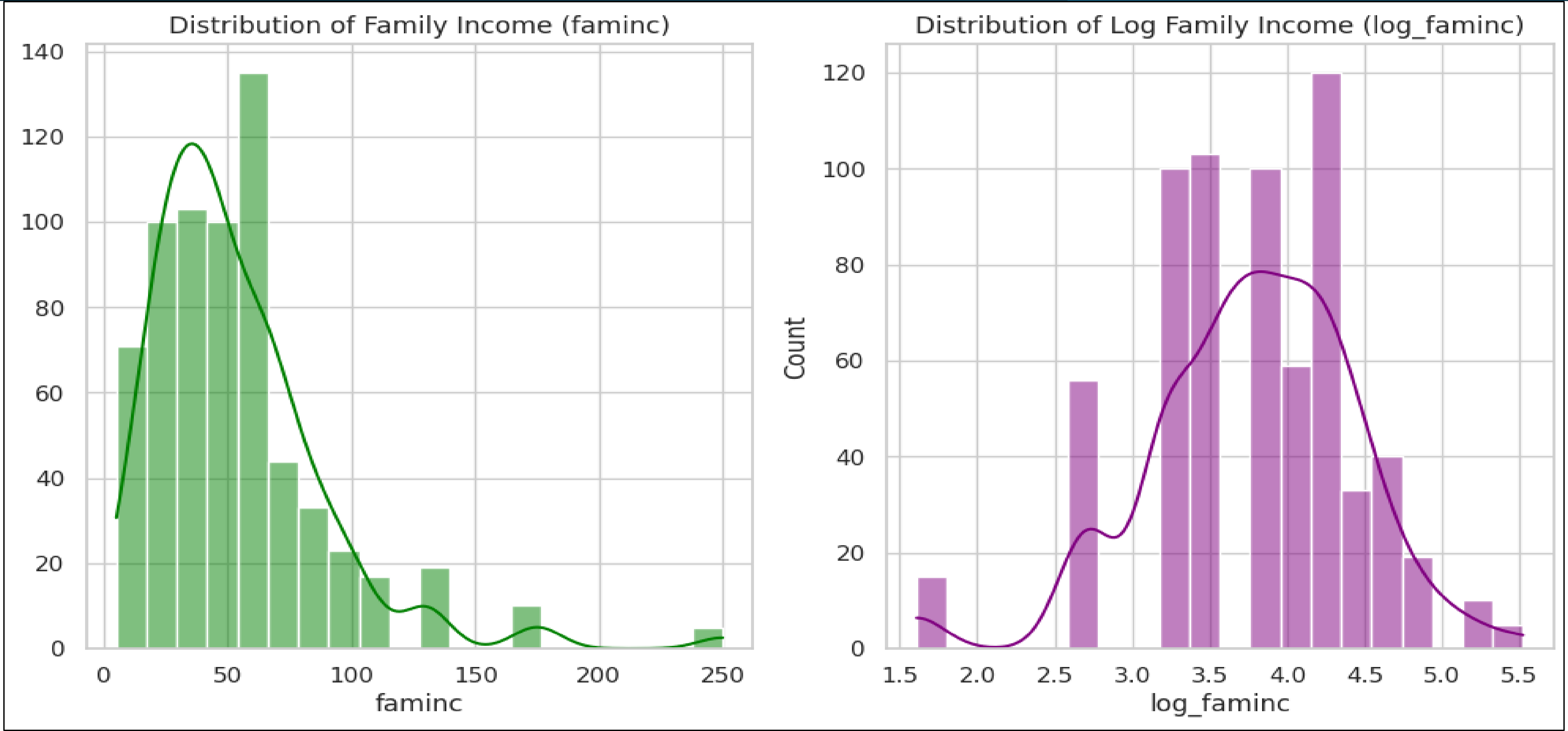
Price Distribution between Eco & Regular Apples

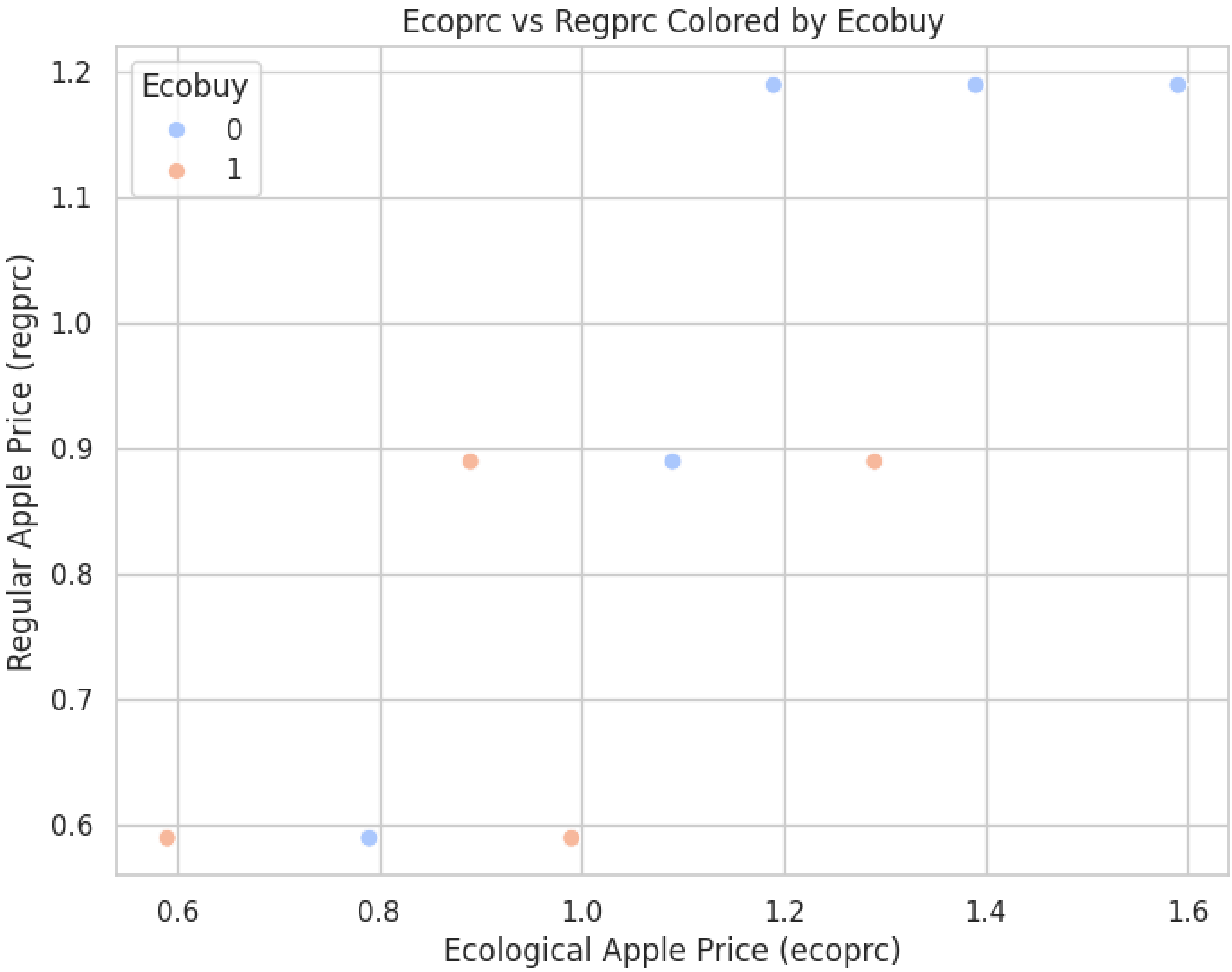




- Boxplot visualizes income differences between those who purchased (ecobuy=1) and didn't purchase (ecobuy=0) eco-labeled apples.
- Family with low income purchase less eco-labeled (ecobuy=1) apple.

Family Income Distribution (non-log vs log)





- Scatter plot indicate that consumers who buy eco-labeled apples may not be as sensitive to price, or that there's minimal impact of regular apple prices on eco-labeled apple purchases.

LPM with faminc

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 ecobuy   R-squared:                       0.110
Model:                            OLS   Adj. R-squared:                  0.102
Method:                 Least Squares   F-statistic:                     13.43
Date:                Mon, 11 Nov 2024   Prob (F-statistic):           2.18e-14
Time:                        15:41:34   Log-Likelihood:                -419.60
No. Observations:                 660   AIC:                             853.2
Df Residuals:                     653   BIC:                             884.6
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.4237      0.165      2.568      0.010       0.100       0.748
ecoprc        -0.8026      0.109     -7.336      0.000      -1.017      -0.588
regprc         0.7193      0.132      5.464      0.000       0.461       0.978
faminc         0.0006      0.001      1.042      0.298      -0.000       0.002
hhsize         0.0238      0.013      1.902      0.058      -0.001       0.048
educ           0.0248      0.008      2.960      0.003       0.008       0.041
age           -0.0005      0.001     -0.401      0.689      -0.003       0.002
==============================================================================
Omnibus:                     4015.360   Durbin-Watson:                   2.084
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               69.344
Skew:                          -0.411   Prob(JB):                     8.75e-16
Kurtosis:                       1.641   Cond. No.                         724.
==============================================================================
```

Interpretation

1. R-squared: 0.110 means the model explains 11% of the variance in eco-friendly buying behavior.
2. F-statistic (13.43) with a very low p-value (2.18e-14) indicates the model is statistically significant overall
3. We see strong significant relationships ($p < 0.05$) with three key factors:
   - ecoprc has a negative impact where a unit increase leads to a 0.80 unit decrease in eco-buying, while regprc shows a positive effect with a 0.72 unit increase in eco-buying per unit change, and educ demonstrates a small but significant positive influence with a 0.02 unit increase per education level.
   - lnhsize shows a marginally positive relationship ($p = 0.058$) with eco-buying, though this effect is just above the conventional significance threshold. faminc and age appear to have no statistically significant impact on eco-buying behavior ($p > 0.05$).

LPM with log(faminc)

```
LPM with log(faminc)
                            OLS Regression Results
==============================================================================
Dep. Variable:                 ecobuy   R-squared:                       0.112
Model:                            OLS   Adj. R-squared:                  0.103
Method:                 Least Squares   F-statistic:                     13.67
Date:                Mon, 11 Nov 2024   Prob (F-statistic):           1.16e-14
Time:                        15:41:34   Log-Likelihood:                -418.94
No. Observations:                 660   AIC:                             851.9
Df Residuals:                     653   BIC:                             883.3
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.3038      0.179      1.697      0.090      -0.048       0.655
ecoprc        -0.8007      0.109     -7.326      0.000      -1.015      -0.586
regprc         0.7214      0.132      5.485      0.000       0.463       0.980
log_faminc     0.0445      0.029      1.550      0.122      -0.012       0.101
hhsize         0.0227      0.013      1.810      0.071      -0.002       0.047
educ           0.0231      0.008      2.733      0.006       0.006       0.040
age           -0.0004      0.001     -0.309      0.758      -0.003       0.002
==============================================================================
Omnibus:                     3113.689   Durbin-Watson:                   2.088
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               68.818
Skew:                          -0.412   Prob(JB):                     1.14e-15
Kurtosis:                       1.649   Cond. No.                         499.
==============================================================================
```

```
Prediction Accuracy:
LPM Model (ecobuy=0): 41.13%, LPM Model (ecobuy=1): 82.52%
```

Interpretation

New R-squared: 0.112 (slightly improved from 0.110 in the previous model)
- F-statistic: 13.67 (slightly higher than 13.43 before)
- Both models remain highly significant ($p < 0.001$)

- ecoprc and regprc remain similarly significant:
  - ecoprc: -0.8007 (vs -0.8026 before)
  - regprc: 0.7214 (vs 0.7193 before)

- Family income transformation:
  - Previous model: faminc was non-significant ($p = 0.298$)
  - New model: log_faminc shows improved significance ($p = 0.122$), though still not significant at 0.05 level.

Probit with (faminc)

```
Probit with faminc
                          Probit Regression Results
==============================================================================
Dep. Variable:                 ecobuy   No. Observations:                  660
Model:                         Probit   Df Residuals:                      653
Method:                           MLE   Df Model:                            6
Date:                Mon, 11 Nov 2024   Pseudo R-squ.:                 0.08664
Time:                        15:41:34   Log-Likelihood:                -399.04
converged:                       True   LL-Null:                       -436.89
Covariance Type:            nonrobust   LLR p-value:                 2.751e-14
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.2438      0.474     -0.514      0.607      -1.173       0.685
ecoprc        -2.2669      0.321     -7.052      0.000      -2.897      -1.637
regprc         2.0302      0.382      5.318      0.000       1.282       2.778
faminc         0.0014      0.002      0.932      0.351      -0.002       0.004
hhsize         0.0691      0.037      1.893      0.058      -0.002       0.141
educ           0.0714      0.024      2.939      0.003       0.024       0.119
age           -0.0012      0.004     -0.340      0.734      -0.008       0.006
==============================================================================
```

Interpretation

Pseudo R-squared: 0.08664 (8.66% of variation explained)
Log-likelihood: -399.04
The model is statistically significant overall (LLR p-value = 2.75e-14)

- ecoprc (eco-friendly price):
  - Coefficient = -2.2669, p = 0.000
  - Strong negative effect on probability of eco-buying. A unit increase in eco-friendly price significantly decreases likelihood of eco-buying
- regprc (regular price):
  - Coefficient = 2.0302, p = 0.000
  - Strong positive effect on probability of eco-buying. Higher regular prices increase likelihood of eco-buying
- educ (education):
  - Coefficient = 0.0714, p = 0.003
  - Positive effect on probability of eco-buying
  - More educated individuals are more likely to make eco-friendly purchases

Interpretation

- Marginally Significant:
- hhsize (household size):
  - Coefficient = 0.0691, p = 0.058
  - Marginally positive effect on eco-buying probability
- Non-Significant Variables:

  famninc (family income): p = 0.351

  age: p = 0.734

  constant: p = 0.607

Probit with log(faminc)

```
Probit with log(faminc)
                          Probit Regression Results
==============================================================================
Dep. Variable:                 ecobuy   No. Observations:                  660
Model:                         Probit   Df Residuals:                      653
Method:                           MLE   Df Model:                            6
Date:                Mon, 11 Nov 2024   Pseudo R-squ.:                 0.08801
Time:                        15:41:34   Log-Likelihood:                -398.43
converged:                       True   LL-Null:                       -436.89
Covariance Type:            nonrobust   LLR p-value:                 1.556e-14
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.5620      0.516     -1.090      0.276      -1.572       0.448
ecoprc        -2.2649      0.322     -7.041      0.000      -2.895      -1.634
regprc         2.0393      0.382      5.335      0.000       1.290       2.788
log_faminc     0.1186      0.082      1.440      0.150      -0.043       0.280
hhsize         0.0662      0.037      1.810      0.070      -0.005       0.138
educ           0.0667      0.025      2.719      0.007       0.019       0.115
age           -0.0009      0.004     -0.253      0.800      -0.008       0.006
==============================================================================


Prediction Accuracy:
Probit Model (ecobuy=0): 41.13%, Probit Model (ecobuy=1): 82.04%
```

Interpretation

- Model Fit:
- Pseudo R-squared slightly improved: 0.08801 vs 0.08664
- Log-likelihood slightly better: -398.43 vs -399.04
- Both models remain highly significant (LLR p-value < 0.001)
- Key Changes in Coefficients: a) Consistently Significant Variables:
- ecoprc: -2.2649 (vs -2.2669)
  - Remains strongly negative and significant (p = 0.000)
  - Almost identical effect size
- regprc: 2.0393 (vs 2.0302)
  - Remains strongly positive and significant (p = 0.000)
  - Very similar effect size
- educ: 0.0667 (vs 0.0714)
  - Still significant but slightly less so (p = 0.007 vs 0.003)
  - Slightly smaller coefficient

b) Changes in Family Income:

- Previous model: faminc coefficient = 0.0014 (p = 0.351)
- New model: log_faminc coefficient = 0.1186 (p = 0.150)
  - Log transformation improved the p-value somewhat
  - Still not significant at conventional levels
- Other Variables:
- hhsize: Remains marginally significant (p = 0.070 vs 0.058)
- age: Remains non-significant in both models
- New Information - Prediction Accuracy:
- Non-eco-buying (ecobuy = 0): 41.13%
- Eco-buying (ecobuy = 1): 82.04%
  - Very similar to previous Probit model's accuracy rates

Logit with (faminc)

```
Logit with faminc
                           Logit Regression Results
==============================================================================
Dep. Variable:                 ecobuy   No. Observations:                  660
Model:                          Logit   Df Residuals:                      653
Method:                           MLE   Df Model:                            6
Date:                Mon, 11 Nov 2024   Pseudo R-squ.:                 0.08642
Time:                        15:41:34   Log-Likelihood:                -399.13
converged:                       True   LL-Null:                       -436.89
Covariance Type:            nonrobust   LLR p-value:                 3.017e-14
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.4278      0.786     -0.544      0.586      -1.968       1.112
ecoprc        -3.6773      0.533     -6.898      0.000      -4.722      -2.632
regprc         3.2742      0.630      5.196      0.000       2.039       4.509
faminc         0.0026      0.003      1.012      0.311      -0.002       0.008
hhsize         0.1145      0.061      1.878      0.060      -0.005       0.234
educ           0.1186      0.041      2.925      0.003       0.039       0.198
age           -0.0022      0.006     -0.372      0.710      -0.014       0.009
==============================================================================
```

Interpretation

- **Model Fit**:
  - **Pseudo R-squared**: 0.0864, explaining about 8.64% of variation in eco-buying.
  - **Log-likelihood**: -399.13, statistically significant (LLR p-value < 0.001).
- **Significant Variables**:
  - **Eco-Friendly Price (`ecoprc`)**: Strong negative impact (coefficient: -3.6773, p < 0.001); higher prices decrease likelihood of eco-buying.
  - **Regular Price (`regprc`)**: Strong positive impact (coefficient: 3.2742, p < 0.001); higher prices for regular apples increase eco-buying probability.
  - **Education (`educ`)**: Positive effect (coefficient: 0.1186, p = 0.003); higher education levels correlate with increased eco-buying.
- **Marginally Significant**:
  - **Household Size (`hhsize`)**: Positive effect (coefficient: 0.1145, p = 0.060); larger households may slightly favor eco-buying.
- **Non-Significant Variables**:
  - **Family Income (`faminc`)**, **Age**, and **Constant** show no significant effect on eco-buying probability.

Logit with log (faminc)

```
Logit with log(faminc)
                           Logit Regression Results
==============================================================================
Dep. Variable:                 ecobuy   No. Observations:                  660
Model:                          Logit   Df Residuals:                      653
Method:                           MLE   Df Model:                            6
Date:                Mon, 11 Nov 2024   Pseudo R-squ.:                 0.08774
Time:                        15:41:34   Log-Likelihood:                -398.55
converged:                       True   LL-Null:                       -436.89
Covariance Type:            nonrobust   LLR p-value:                 1.740e-14
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.9724      0.856     -1.137      0.256      -2.649       0.704
ecoprc        -3.6751      0.534     -6.888      0.000      -4.721      -2.629
regprc         3.2913      0.631      5.216      0.000       2.055       4.528
log_faminc     0.2022      0.136      1.483      0.138      -0.065       0.470
hhsize         0.1097      0.061      1.796      0.072      -0.010       0.229
educ           0.1113      0.041      2.722      0.006       0.031       0.192
age           -0.0016      0.006     -0.280      0.779      -0.013       0.010
==============================================================================

Prediction Accuracy:
Logit Model (ecobuy=0): 41.13%, Logit Model (ecobuy=1): 82.28%
```

Interpretation

- **Model Fit Improvement**:
  - Slightly better **Pseudo R-squared** (0.0877 vs 0.0864) and **Log-likelihood** (-398.55 vs -399.13).
- **Key Consistently Significant Variables**:
  - **ecoprc**: Strong, negative effect remains highly significant (coefficient: -3.6751).
  - **regprc**: Strong, positive effect also remains highly significant (coefficient: 3.2913).
  - **Education (educ)**: Positive effect on eco-buying with minor change in p-value (0.006 vs 0.003).
- **Income Effect**:
  - **Log Transformation** (log_faminc) increased coefficient size (0.2022) but remains insignificant (p = 0.138).
- **Prediction Accuracy**:
  - Similar to previous model: **Eco-buying accuracy** is high at 82.3%, while **non-eco-buying accuracy** is lower at 41.1%.

Final Model

- Based on Pseudo-R squared values we will choose the Probit model with log of family income as the final model. We would like to note that the fit values for almost all the models are very similar

Final Model

- Probit Model
- $ecobuy = \beta 0 + \beta 1 \cdot ecoprc + \beta 2 \cdot regprc + \beta 3 \cdot \log(faminc) + \beta 4 \cdot hhsize + \beta 5 \cdot educ + \beta 6 \cdot age + u$

Results

- **Price Sensitivity**:
  - **Higher eco-labeled apple prices** decrease purchase probability.
  - **Higher regular apple prices** push consumers towards eco-labeled options.
- **Influential Non-Price Factor**:
  - **Education** significantly increases purchase probability for eco-labeled apples, suggesting a link between education and eco-conscious choices.
- **Prediction Accuracy**:
  - **Accurate for ecobuy = 1 (82.5%)**: All models better predict eco-friendly purchases.
  - **Less Accurate for ecobuy = 0 (41.1%)**: Difficulty in predicting when consumers opt out.

Conclusions and Key Insights

- **Strong Demand for Eco-labeled Products**:
  - Approximately **62.4% of consumers** willing to buy eco-labeled apples, showing a solid base of eco-conscious consumers.
- **Importance of Strategic Pricing**:
  - Price elasticity suggests that **competitive pricing for eco-labeled apples** could maximize purchases in this segment.
- **Role of Education**:
  - Across all models, **education is the most impactful factor** beyond price, indicating educated consumers prioritize environmental choices.

Thank You!

We'd be happy to answer any questions

Appendix

Google collab: https://colab.research.google.com/drive/1Rjq-8jDbSwotC1biNg0iifJqKHiazRLk

```
# Import necessary libraries
import pandas as pd
import numpy as np
import statsmodels.api as sm
from statsmodels.discrete.discrete_model import Probit, Logit
import matplotlib.pyplot as plt
import seaborn as sns
# Load the data
#data_path = '/path_to_file/Apple.csv'
apple_data = pd.read_csv("Apple.csv")


# Define ecobuy as 1 if ecolbs > 0, else 0
apple_data['ecobuy'] = (apple_data['ecolbs'] > 0).astype(int)



# Define the independent variables, including log(faminc)
apple_data['log_faminc'] = np.log(apple_data['faminc'].replace(0, np.nan))  # Avoid log issues with 0
X = apple_data[['ecoprc', 'regprc', 'faminc', 'hhsize', 'educ', 'age']]
X_log_faminc = apple_data[['ecoprc', 'regprc', 'log_faminc', 'hhsize', 'educ', 'age']]
y = apple_data['ecobuy']
X = sm.add_constant(X)
X_log_faminc = sm.add_constant(X_log_faminc)
```

Appendix

```
# EDA: Exploratory Data Analysis
# Set up Seaborn style
sns.set(style="whitegrid")

# 1. Distribution of `ecoprc` and `regprc` (Price of ecolabeled and regular apples)
plt.figure(figsize=(12, 6))
plt.subplot(1, 2, 1)
sns.histplot(apple_data['ecoprc'], bins=20, kde=True, color='skyblue')
plt.title('Distribution of Ecological Apple Price (ecoprc)')
plt.subplot(1, 2, 2)
sns.histplot(apple_data['regprc'], bins=20, kde=True, color='salmon')
plt.title('Distribution of Regular Apple Price (regprc)')
plt.show()



# 2. Distribution of family income and log-transformed family income
plt.figure(figsize=(12, 6))
plt.subplot(1, 2, 1)
sns.histplot(apple_data['faminc'], bins=20, kde=True, color='green')
plt.title('Distribution of Family Income (faminc)')



plt.subplot(1, 2, 2)
sns.histplot(apple_data['log_faminc'], bins=20, kde=True, color='purple')
plt.title('Distribution of Log Family Income (log_faminc)')
plt.show()
```

Appendix

```
# 3. Boxplot of family income by ecobuy (to observe income differences in buyers vs. non-buyers)
plt.figure(figsize=(8, 6))
sns.boxplot(x='ecobuy', y='faminc', data=apple_data, palette='Set2')
plt.title('Family Income by Ecolabeled Apple Purchase Decision (ecobuy)')
plt.xlabel('Ecobuy')
plt.ylabel('Family Income')
plt.show()

# 4. Scatter plot of `ecoprc` and `regprc` with hue `ecobuy` to visualize price sensitivity
plt.figure(figsize=(8, 6))
sns.scatterplot(x='ecoprc', y='regprc', hue='ecobuy', data=apple_data, palette='coolwarm', s=50)
plt.title('Ecoprc vs Regprc Colored by Ecobuy')
plt.xlabel('Ecological Apple Price (ecoprc)')
plt.ylabel('Regular Apple Price (regprc)')
plt.legend(title='Ecobuy')
plt.show()


# 5. Pair plot of numeric variables with hue `ecobuy`
sns.pairplot(apple_data[['ecoprc', 'regprc', 'faminc', 'hhsize', 'educ', 'age', 'ecobuy']], hue='ecobuy', palette='husl')
plt.suptitle('Pairplot of Key Variables by Purchase Decision (Ecobuy)', y=1.02)
plt.show()


# Model 1: Linear Probability Model (LPM)
# LPM with faminc
lpm_model = sm.OLS(y, X).fit()
# LPM with log(faminc)
lpm_model_log_faminc = sm.OLS(y, X_log_faminc, missing='drop').fit()
```

Appendix

```
# Model 2: Probit Model
# Probit with faminc
probit_model = Probit(y, X).fit(disp=0)
# Probit with log(faminc)
probit_model_log_faminc = Probit(y, X_log_faminc, missing='drop').fit(disp=0)
# Model 3: Logit Model
# Logit with faminc
logit_model = Logit(y, X).fit(disp=0)
# Logit with log(faminc)
logit_model_log_faminc = Logit(y, X_log_faminc, missing='drop').fit(disp=0)
# Prediction Accuracy and Threshold Analysis
# Define a threshold of 0.5 for both LPM, Probit, and Logit predictions
# LPM Predictions
lpm_predictions = (lpm_model_log_faminc.predict(X_log_faminc) >= 0.5).astype(int)
# Probit Predictions
probit_predictions = (probit_model_log_faminc.predict(X_log_faminc) >= 0.5).astype(int)
# Logit Predictions
logit_predictions = (logit_model_log_faminc.predict(X_log_faminc) >= 0.5).astype(int)
# Calculate the prediction accuracy for each model
def calculate_accuracy(predictions, actuals):
    correct_0 = ((predictions == 0) & (actuals == 0)).sum() / (actuals == 0).sum() * 100
    correct_1 = ((predictions == 1) & (actuals == 1)).sum() / (actuals == 1).sum() * 100
    return correct_0, correct_1


lpm_correct_0, lpm_correct_1 = calculate_accuracy(lpm_predictions, y)
probit_correct_0, probit_correct_1 = calculate_accuracy(probit_predictions, y)
logit_correct_0, logit_correct_1 = calculate_accuracy(logit_predictions, y)
```

Appendix

```
# Display the summaries and accuracy results
# LPM Report
print("### Linear Probability Model (LPM) Report ###")
print("LPM with faminc")
print(lpm_model.summary())
print("\nLPM with log(faminc)")
print(lpm_model_log_faminc.summary())
print(f"\nPrediction Accuracy:\nLPM Model (ecobuy=0): {lpm_correct_0:.2f}%, LPM Model (ecobuy=1): {lpm_correct_1:.2f}%")


# Probit Report
print("\n### Probit Model Report ###")
print("Probit with faminc")
print(probit_model.summary())
print("\nProbit with log(faminc)")
print(probit_model_log_faminc.summary())
print(f"\nPrediction Accuracy:\nProbit Model (ecobuy=0): {probit_correct_0:.2f}%, Probit Model (ecobuy=1): {probit_correct_1:.2f}%")


# Logit Report
print("\n### Logit Model Report ###")
print("Logit with faminc")
print(logit_model.summary())
print("\nLogit with log(faminc)")
print(logit_model_log_faminc.summary())
print(f"\nPrediction Accuracy:\nLogit Model (ecobuy=0): {logit_correct_0:.2f}%, Logit Model (ecobuy=1): {logit_correct_1:.2f}%")
```