Understanding the structure of an Econometrics study

FLORENCE

2025-01-30

Part A: Structure of an Econometrics Study

Microeconomics Research Question

How does an individual's income and credit history affect their probability of credit card approval?

General Structure

- **Type of Research:** Secondary
- **Variables:**
  - **Dependent:** `Acc` (Credit Card Application Accepted)
  - **Independent:** `MDR`, `Age`, `Income`, `Avgexp`, `Ownrent`, `Selfempl`
- **Econometric Model:** Logistic Regression

$$P(Acc = 1) = \frac{e^{\beta_0 + \beta_1 MDR + \beta_2 Age + \beta_3 Income + \beta_4 Avgexp + \beta_5 Ownrent + \beta_6 Selfempl}}{1 + e^{\beta_0 + \beta_1 MDR + \beta_2 Age + \beta_3 Income + \beta_4 Avgexp + \beta_5 Ownrent + \beta_6 Selfempl}}$$

Macroeconomics Research Question

How does economic growth affect overall credit card approval rates in an economy?

General Structure

- **Type of Research:** Secondary
- **Variables:**
  - **Dependent:** `Credit Card Approval Rate`
  - **Independent:** `GDP Growth`, `Unemployment Rate`, `Inflation Rate`, `Interest Rate`
- **Econometric Model:** Multiple Linear Regression

$$CreditCardApprovalRate = \beta_0 + \beta_1 GDPGrowth + \beta_2 Unemployment + \beta_3 Inflation + \beta_4 InterestRate + \epsilon$$

Part B: Modelling Econometric Data in R

Step 1: Load Required Libraries and Dataset

```r
library(tidyverse)  # Data manipulation
library(dplyr)      # Data filtering
library(ggplot2)    # Visualization
library(caret)      # Machine Learning
library(stargazer)  # Regression tables
library(MatchIt)    # Causal inference
library(broom)      # Tidying model output

# The dataset
df <- read.csv("dataset_for_assignment(2)(1).csv")

# First few rows
head(df)
```

```
##   MDR Acc Age Income Avgexp Ownrent Selfempl
## 1   0   1  38   4.52 124.98       1        0
## 2   0   1  33   2.42   9.85       0        0
## 3   0   1  34   4.50  15.00       1        0
## 4   0   1  31   2.54 137.87       0        0
## 5   0   1  32   9.79 546.50       1        0
## 6   0   1  23   2.50  92.00       0        0
```

```r
# Any missing values?
colSums(is.na(df))
```

```
##      MDR      Acc      Age   Income   Avgexp  Ownrent Selfempl
##        0        0        0        0        0        0        0
```

```r
# Summary statistics
summary(df)
```

```
##       MDR              Acc              Age             Income
##  Min.   :0.00   Min.   :0.00   Min.   :20.00   Min.   : 1.500
##  1st Qu.:0.00   1st Qu.:0.00   1st Qu.:26.00   1st Qu.: 2.365
##  Median :0.00   Median :1.00   Median :31.00   Median : 3.000
##  Mean   :0.36   Mean   :0.73   Mean   :32.08   Mean   : 3.369
##  3rd Qu.:0.00   3rd Qu.:1.00   3rd Qu.:37.00   3rd Qu.: 3.970
##  Max.   :7.00   Max.   :1.00   Max.   :55.00   Max.   :10.000
##      Avgexp           Ownrent         Selfempl
##  Min.   :   0.0   Min.   :0.00   Min.   :0.00
##  1st Qu.:   0.0   1st Qu.:0.00   1st Qu.:0.00
##  Median :  81.3   Median :0.00   Median :0.00
##  Mean   : 189.0   Mean   :0.36   Mean   :0.05
##  3rd Qu.: 252.8   3rd Qu.:1.00   3rd Qu.:0.00
##  Max.   :1898.0   Max.   :1.00   Max.   :1.00
```

Normalizing Income and Avgexp for better model performance

```
df$Income <- df$Income / max(df$Income)
df$Avgexp <- df$Avgexp / max(df$Avgexp)
```

 # Fitting a logistic regression model

```
logit_model <- glm(Acc ~ MDR + Age + Income + Avgexp + Ownrent, data = df, family = binomial)
```

```
## Warning: glm.fit: algorithm did not converge
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
# Model summary
summary(logit_model)
```

```
##
## Call:
## glm(formula = Acc ~ MDR + Age + Income + Avgexp + Ownrent, family = binomial,
##     data = df)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    146.024   7073.890   0.021    0.984
## MDR            -20.068 298320.604   0.000    1.000
## Age             -3.749    256.517  -0.015    0.988
## Income        -310.956  24470.842  -0.013    0.990
## Avgexp       18517.566 809676.522   0.023    0.982
## Ownrent          2.078   5552.584   0.000    1.000
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1.1665e+02  on 99  degrees of freedom
## Residual deviance: 1.8476e-06  on 94  degrees of freedom
## AIC: 12
##
## Number of Fisher Scoring iterations: 25
```

Multicollinearity Check

```
correlation_matrix <- cor(df %>% select(-Acc))
print(correlation_matrix)
```

```
##                    MDR         Age     Income      Avgexp       Ownrent
## MDR       1.0000000000  0.08061585 0.03231663 -0.17266824  0.0008290335
## Age       0.0806158504  1.00000000 0.26852063  0.05275400  0.2865031786
## Income    0.0323166294  0.26852063 1.00000000  0.38536546  0.3482643154
```

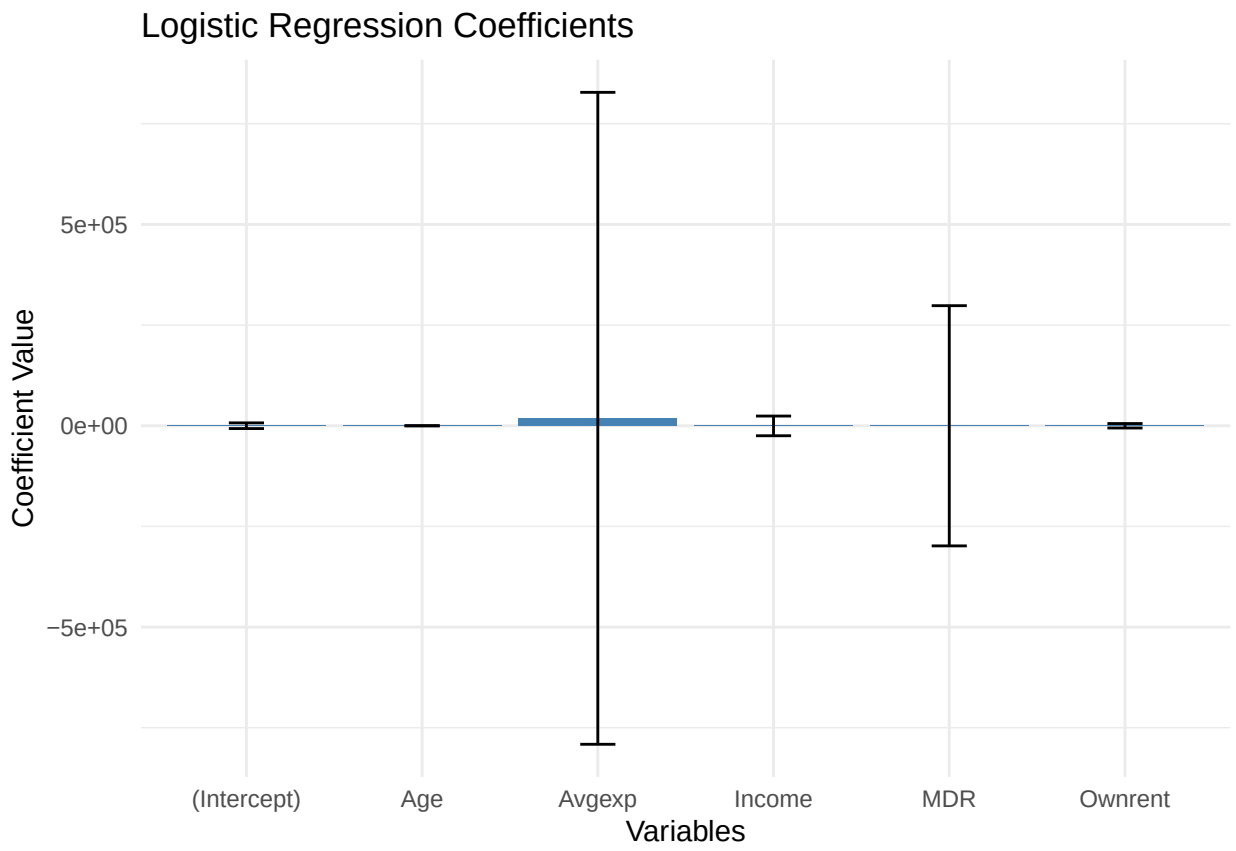
```
## Avgexp   -0.1726682410  0.05275400 0.38536546  1.00000000  0.2108062397
## Ownrent   0.0008290335  0.28650318 0.34826432  0.21080624  1.0000000000
## Selfempl -0.0821635190 -0.02002775 0.20646528 -0.01427031 -0.0764719113
##             Selfempl
## MDR      -0.08216352
## Age      -0.02002775
## Income    0.20646528
## Avgexp   -0.01427031
## Ownrent  -0.07647191
## Selfempl  1.00000000
```

Average Treatment Effect (ATE) for Income

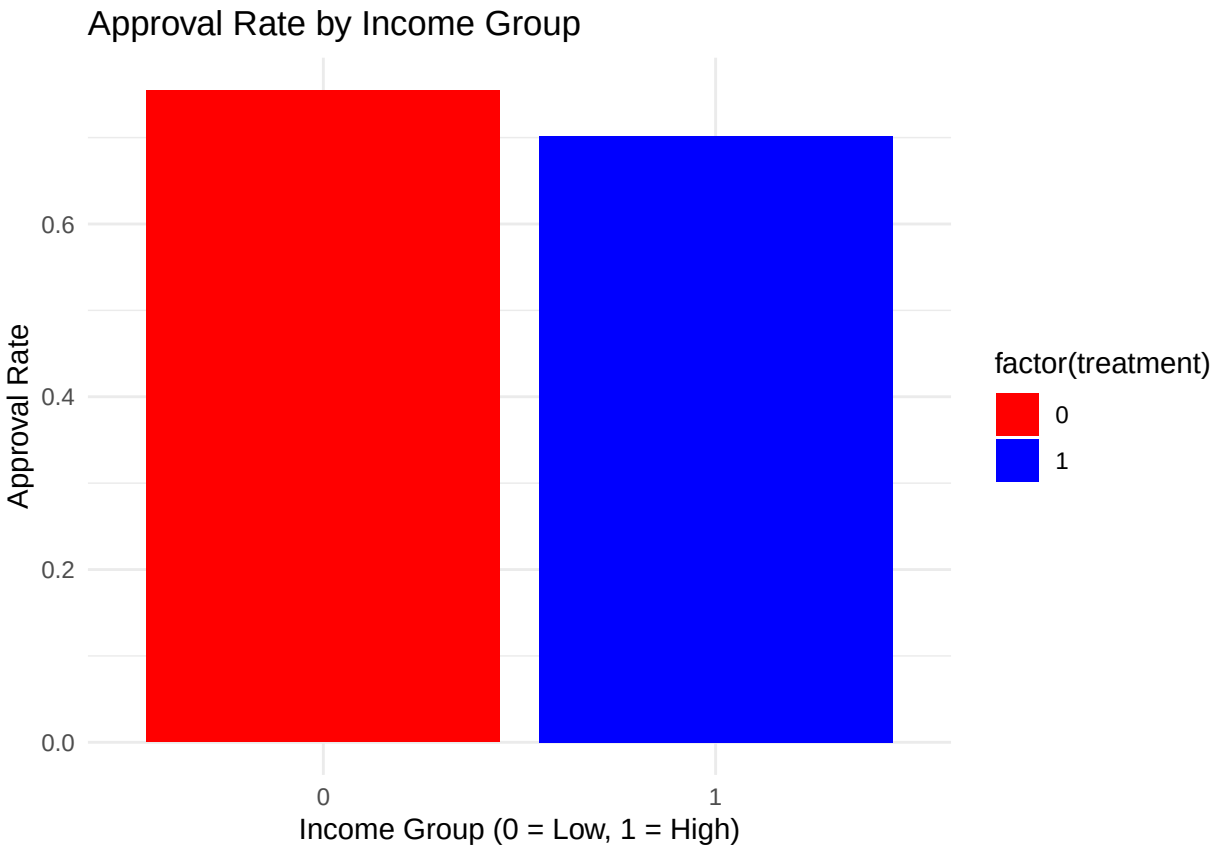
```r
# Median income
median_income <- median(df$Income)

# Treatment and control groups
df$treatment <- ifelse(df$Income > median_income, 1, 0)

# Average approval rate
treatment_mean <- mean(df$Acc[df$treatment == 1])
control_mean <- mean(df$Acc[df$treatment == 0])

# Compute ATE
ATE <- treatment_mean - control_mean
ATE
```

```
## [1] -0.05258932
```

```r
# Coefficients and plot
logit_coeffs <- tidy(logit_model)
ggplot(logit_coeffs, aes(x = term, y = estimate)) +
  geom_col(fill = "steelblue") +
  geom_errorbar(aes(ymin = estimate - std.error, ymax = estimate + std.error), width = 0.2) +
  labs(title = "Logistic Regression Coefficients", x = "Variables", y = "Coefficient Value") +
  theme_minimal()
```




```
# Approval rate for treatment and control groups plot
df |>
  group_by(treatment) |>
  summarise(approval_rate = mean(Acc)) |>
  ggplot(aes(x = factor(treatment), y = approval_rate, fill = factor(treatment))) +
  geom_bar(stat = "identity") +
  labs(title = "Approval Rate by Income Group", x = "Income Group (0 = Low, 1 = High)", y = "Approval Ra
  scale_fill_manual(values = c("red", "blue")) +
  theme_minimal()
```


Approval Rate by Income Group
Approval Rate
0.0
0.2
0.4
0.6
0
1
Income Group (0 = Low, 1 = High)
factor(treatment)
0
1