Econ104_project3

Ashley Guerra, Eyasu Olana, Evan Titus, Gavin Valenzuela

2025-01-06

```
library(AER)
```

```
## Loading required package: car

## Loading required package: carData

## Loading required package: lmtest

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Loading required package: sandwich

## Loading required package: survival
```

```
library(MASS)
library(caret)
```

```
## Loading required package: ggplot2

## Loading required package: lattice

##
## Attaching package: 'caret'

## The following object is masked from 'package:survival':
##
##     cluster
```

```
library(ggplot2)

data("SmokeBan")
```

```
# Convert the 'smoker' column to a binary numeric variable
SmokeBan$smoker <- ifelse(SmokeBan$smoker == "yes", 1, 0)

# Convert all factor variables to dummy variables using model.matrix
SmokeBan$ban <- as.factor(SmokeBan$ban)
SmokeBan$education <- as.factor(SmokeBan$education)
SmokeBan$afam <- as.factor(SmokeBan$afam)
SmokeBan$hispanic <- as.factor(SmokeBan$hispanic)
SmokeBan$gender <- as.factor(SmokeBan$gender)
```

```r
# Create dummy variables using model.matrix
dummies <- model.matrix(~ ban + education + afam + hispanic + gender - 1, data = SmokeBan)

# Combine the dummy variables with the rest of the data
SmokeBan <- cbind(SmokeBan, dummies)

# Remove the original factor columns
SmokeBan <- SmokeBan[, !(names(SmokeBan) %in% c("ban", "education", "afam", "hispanic", "gender"))]

# Check the structure of the dataset to ensure all variables are numeric
str(SmokeBan)
```

```
## 'data.frame':    10000 obs. of  11 variables:
##  $ smoker               : num  1 1 0 1 0 0 1 1 0 0 ...
##  $ age                  : int  41 44 19 29 28 40 47 36 49 44 ...
##  $ banno                : num  0 0 1 1 0 1 0 1 0 1 ...
##  $ banyes               : num  1 1 0 0 1 0 1 0 1 0 ...
##  $ educationhs          : num  1 0 0 1 0 0 0 0 0 0 ...
##  $ educationsome college: num  0 1 1 0 1 1 1 1 1 1 ...
##  $ educationcollege     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ educationmaster      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ afamyes              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ hispanicyes          : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ genderfemale         : num  1 1 1 1 1 0 1 0 1 0 ...
```

2C)

Fit the Linear Probability Model

```r
lpm <- lm(smoker ~ ., data = SmokeBan)
summary(lpm)
```

```
##
## Call:
## lm(formula = smoker ~ ., data = SmokeBan)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.48682 -0.28725 -0.17239 -0.03619  0.99792
##
## Coefficients: (1 not defined because of singularities)
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             0.4658515  0.0221432  21.038  < 2e-16 ***
## age                    -0.0013543  0.0003503  -3.867 0.000111 ***
## banno                   0.0453435  0.0087250   5.197 2.07e-07 ***
## banyes                         NA         NA      NA       NA
## educationhs            -0.0858065  0.0162692  -5.274 1.36e-07 ***
## `educationsome college` -0.1537486  0.0165818  -9.272  < 2e-16 ***
## educationcollege       -0.2683776  0.0176077 -15.242  < 2e-16 ***
## educationmaster        -0.3099189  0.0197471 -15.694  < 2e-16 ***
## afamyes                -0.0265034  0.0157518  -1.683 0.092491 .
## hispanicyes            -0.1037449  0.0139463  -7.439 1.10e-13 ***
## genderfemale           -0.0328743  0.0085489  -3.845 0.000121 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.417 on 9990 degrees of freedom
## Multiple R-squared:  0.05397,    Adjusted R-squared:  0.05312
## F-statistic: 63.33 on 9 and 9990 DF,  p-value: < 2.2e-16
```

Fit the Probit Model

```r
probit_model <- glm(smoker ~ ., family = binomial(link = "probit"), data = SmokeBan)
summary(probit_model)
```

```
##
## Call:
## glm(formula = smoker ~ ., family = binomial(link = "probit"),
##     data = SmokeBan)
##
## Coefficients: (1 not defined because of singularities)
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -0.041773   0.071629  -0.583 0.559774
## age                    -0.004203   0.001169  -3.596 0.000323 ***
## banno                   0.151762   0.028948   5.243 1.58e-07 ***
## banyes                        NA         NA      NA       NA
## educationhs            -0.242373   0.050733  -4.777 1.78e-06 ***
## `educationsome college` -0.444975   0.052362  -8.498  < 2e-16 ***
## educationcollege       -0.871756   0.058523 -14.896  < 2e-16 ***
## educationmaster        -1.094230   0.071552 -15.293  < 2e-16 ***
## afamyes                -0.079690   0.052721  -1.512 0.130650
## hispanicyes            -0.332704   0.048001  -6.931 4.17e-12 ***
## genderfemale           -0.110625   0.028779  -3.844 0.000121 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11074  on 9999  degrees of freedom
## Residual deviance: 10505  on 9990  degrees of freedom
## AIC: 10525
##
## Number of Fisher Scoring iterations: 4
```

Fit the Logit Model

```r
logit_model <- glm(smoker ~ ., family = binomial(link = "logit"), data = SmokeBan)
summary(logit_model)
```

```
##
## Call:
## glm(formula = smoker ~ ., family = binomial(link = "logit"),
##     data = SmokeBan)
##
## Coefficients: (1 not defined because of singularities)
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -0.016425   0.119620  -0.137 0.890788
## age                    -0.007452   0.001987  -3.751 0.000176 ***
```

```
## banno                      0.250735   0.049164   5.100 3.40e-07 ***
## banyes                           NA         NA      NA       NA
## educationhs               -0.407770   0.083067  -4.909 9.16e-07 ***
## `educationsome college` -0.750995   0.086513  -8.681  < 2e-16 ***
## educationcollege          -1.506322   0.100651 -14.966  < 2e-16 ***
## educationmaster           -1.931075   0.131261 -14.712  < 2e-16 ***
## afamyes                   -0.149472   0.089994  -1.661 0.096732 .
## hispanicyes               -0.584845   0.083085  -7.039 1.93e-12 ***
## genderfemale              -0.188720   0.049105  -3.843 0.000121 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11074  on 9999  degrees of freedom
## Residual deviance: 10502  on 9990  degrees of freedom
## AIC: 10522
##
## Number of Fisher Scoring iterations: 4
```

AIC and BIC for model comparison

```r
# Calculate AIC and BIC for model comparison
aic_values <- c(AIC(lpm), AIC(probit_model), AIC(logit_model))
bic_values <- c(BIC(lpm), BIC(probit_model), BIC(logit_model))

# Predicted probabilities
pred_lpm <- predict(lpm, type = "response")
pred_probit <- predict(probit_model, type = "response")
pred_logit <- predict(logit_model, type = "response")

# Convert probabilities to class labels
threshold <- 0.5
pred_lpm_class <- factor(ifelse(pred_lpm > threshold, 1, 0), levels = c(0, 1))
pred_probit_class <- factor(ifelse(pred_probit > threshold, 1, 0), levels = c(0, 1))
pred_logit_class <- factor(ifelse(pred_logit > threshold, 1, 0), levels = c(0, 1))
actual_smoker <- factor(SmokeBan$smoker, levels = c(0, 1))

# Confusion matrices
confusion_lpm <- caret::confusionMatrix(pred_lpm_class, actual_smoker)
confusion_probit <- caret::confusionMatrix(pred_probit_class, actual_smoker)
confusion_logit <- caret::confusionMatrix(pred_logit_class, actual_smoker)

# Classification reports
accuracy_values <- c(confusion_lpm$overall['Accuracy'],
                     confusion_probit$overall['Accuracy'],
                     confusion_logit$overall['Accuracy'])

# Model comparison table
model_comparison <- data.frame(
  Model = c("LPM", "Probit", "Logit"),
  AIC = aic_values,
  BIC = bic_values,
  Accuracy = accuracy_values
```

```r
)

print(model_comparison)
```

```
##    Model      AIC      BIC Accuracy
## 1    LPM 10895.45 10974.76   0.7577
## 2 Probit 10524.70 10596.80   0.7585
## 3  Logit 10522.19 10594.29   0.7602
```

```r
# Identify the preferred model
preferred_model <- model_comparison[which.min(model_comparison$AIC), ]
print(preferred_model)
```

```
##   Model      AIC      BIC Accuracy
## 3 Logit 10522.19 10594.29   0.7602
```

Answer :

The Logit model has the lowest AIC (10522.19) and BIC (10594.29) values, indicating a better fit compared to the Linear Probability Model (LPM) and the Probit model. Additionally, the Logit model shows the highest accuracy (0.7602) in predicting the binary dependent variable.