logict regresion model loans

2026-02-02

Logistic Regression Model

The Loan Default dataset contains information on 240 individuals, including their annual income and loan default status (default or no default). The objective of this analysis is to model the relationship between a borrower's annual income and the likelihood of loan default.

The variables included in the dataset are described in detail below. Variable Name Description

Income Annual income of the borrower in thousands of dollars. Loan-default-indicator A value of 1 indicates the borrower defaulted on the loan, while 0 indicates no default.

loan default indicator A value of 1 indicates the borrower defaulted on the loan 0 indicates no default

a. Formulate the logistic regression model describing the relationship between borrower annual income and the probability of loan default.

$\log(p(x)/1-p(x)) = 0 + 1x$

b. Estimate the parameters of the logistic regression model using the observed data and report the corresponding regression output.

```
mydata = read.csv( './loan_default.csv' )
names( mydata )
```

```
## [1] "Income"                "Loan.default.indicator"
```

```
# output:[1] "Income"                 "Loan.default.indicator"
str(mydata)
```

```
## 'data.frame':    240 obs. of  2 variables:
##  $ Income                : num  20.1 20.1 20.2 20.6 21.2 ...
##  $ Loan.default.indicator: int  0 1 1 1 1 0 0 1 1 1 ...
```

```
# 'data.frame': 240 obs. of  2 variables:
#  $ Income                : num  20.1 20.1 20.2 20.6 21.2 ...
#  $ Loan.default.indicator: int  0 1 1 1 1 0 0 1 1 1 ...

# Define the predictor variable (x),
x = mydata$Income
# Define the response variable (y)
y = mydata$Loan.default.indicator
# The 'y' variable currently contains labels (e.g., "F", "P")
# To do that, we can use the factor(...) function
y = factor(mydata$Loan.default.indicator)
# In R, the second level of a factor is considered the "success" category by default.
# In this case, "Pass" is considered the success.
# To confirm the factor levels, use the levels() function
levels(y)
```

```
## [1] "0" "1"
```

```r
fitted.model = glm(y ~ x, family = binomial)

#Use the summary function to get the regression ouput
summary(fitted.model)
```

```
##
## Call:
## glm(formula = y ~ x, family = binomial)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.823610   0.457738   1.799    0.072 .
## x           -0.043751   0.009009  -4.856  1.2e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 228.68  on 239  degrees of freedom
## Residual deviance: 199.07  on 238  degrees of freedom
## AIC: 203.07
##
## Number of Fisher Scoring iterations: 5
```

Estimating an interpreting coeffcients

```r
coeficients=coefficients(fitted.model)
print(fitted.model)
```

```
##
## Call:  glm(formula = y ~ x, family = binomial)
##
## Coefficients:
## (Intercept)            x
##     0.82361     -0.04375
##
## Degrees of Freedom: 239 Total (i.e. Null);  238 Residual
## Null Deviance:       228.7
## Residual Deviance: 199.1     AIC: 203.1
```

```r
b0factor=exp(0.823610)
print(b0factor)
```

```
## [1] 2.278711
```

```r
b1factor=exp(-0.043751)
print(b1factor)
```

```
## [1] 0.9571923
```

B0=0.823610 When income is 0 (theoretical case), Recall, The intercept B0/A is estimated to be 0.823610 So, the odds of get a default is estimated to e^0.823610 =>2.28

E^0.823610 2.28

B1=-0.043751 we predict the odds of a borrower decreases by a factor fo e^B1 when B1>0 For every $1,000

increase in income, the odds of default are multiplied by 0.957. Each additional $1,000 reduces the odds of default by about 4.3%.

E^-0.043751 0.957

c. Evaluate the statistical significance and predictive utility of the fitted model in explaining the probability of loan default.

```
fitted.model = glm(y ~ x, family = binomial)
# The model includes only an intercept (denoted by "1").
fitted.model.no.predictor = glm( y ~ 1, family = binomial)
#Use the summary function to get the regression ouput
summary(fitted.model.no.predictor)
```

```
##
## Call:
## glm(formula = y ~ 1, family = binomial)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.4939     0.1668  -8.955   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 228.68  on 239  degrees of freedom
## Residual deviance: 228.68  on 239  degrees of freedom
## AIC: 230.68
##
## Number of Fisher Scoring iterations: 4
```

Since the p-value (1.2e-06) is far below the significance level of 0.05, we reject the null hypothesis that the income coefficient is equal to zero. Therefore, there is strong statistical evidence that annual income is significantly associated with the probability of loan default.

```
# against the null model (intercept only)
anova(
fitted.model.no.predictor, # the null model (without predictor)
fitted.model,
# the alternative model (with predictor)
test = "Chisq"
)
```

```
## Analysis of Deviance Table
##
## Model 1: y ~ 1
## Model 2: y ~ x
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       239     228.68
## 2       238     199.07  1    29.61 5.283e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Likelihood Ratio Test: significant G^2 statistic=29.61, p-value=5.283e-08 ($p < 0.05$), indicating that including income variable significantly improves the model's predictive ability.

---

d. Compute the estimated odds of loan default for borrowers with annual incomes of 20 thousand dollars and interpret the odds.

```
beta0 <- 0.823610
beta1 <- -0.043751

odds_20 <- exp(beta0 + beta1 * 20)
odds_20
```

```
## [1] 0.9498891
```

```
p= odds_20/(1+odds_20)
print(p)
```

```
## [1] 0.4871503
```

e. Compute the estimated probabilities of loan default for borrowers with annual incomes of 20, 30, 40, 50, and 60 thousand dollars.

```
income_vals <- c(20,30,40,50,60)
p <- plogis(beta0 + beta1 * income_vals)
data.frame(Income_k = income_vals, ProbDefault = p)
```

```
##   Income_k ProbDefault
## 1       20   0.4871503
## 2       30   0.3801480
## 3       40   0.2836496
## 4       50   0.2036007
## 5       60   0.1416746
```

f. Interpret the intercept parameter of the logistic regression model in the context of loan default risk.

The intercept is the log-odds of default when income is $0k; since $0k is unrealistic, it mainly serves as a baseline for computing probabilities.

g. Interpret the slope parameter of the logistic regression model, emphasizing its effect on the odds of loan default as annual income changes.

The slope coefficient is b1 =>E^-0.043751 0.957

This means that for every $1,000 increase in annual income, the log-odds of loan default decrease by 0.0437(1-0.957).

In other words, the odds of default decrease by approximately 4.3% for every $1,000 increase in annual income.