Regresores cualitativos

Contents

1 Regresores cualitativos con dos niveles

El archivo *Credit Approval Decisions.csv* contiene información acerca de la aceptación o rechazo de un crédito. Se dispone también de información acerca de otras variables adicionales. El objetivo es analizar la variable *Decision* en función del resto de variables.

```r
d = read.csv("datos/Credit Approval Decisions.csv")
str(d)
```

```
## 'data.frame':    50 obs. of  6 variables:
##  $ Homeowner            : chr  "Y" "Y" "Y" "N" ...
##  $ Credit_Score         : int  725 573 677 625 527 795 733 620 591 660 ...
##  $ Years_Credit_History : int  20 9 11 15 12 22 7 5 17 24 ...
##  $ Revolving_Balance    : num  11320 7200 20000 12800 5700 ...
##  $ Revolving_Utilization: num  0.25 0.7 0.55 0.65 0.75 0.12 0.2 0.62 0.5 0.35 ...
##  $ Decision             : chr  "Approve" "Reject" "Approve" "Reject" ...
```

Convertimos las variables cualitativas a factor:

```r
d$Homeowner = factor(d$Homeowner)
d$Decision = factor(d$Decision)
str(d)
```

```
## 'data.frame':    50 obs. of  6 variables:
##  $ Homeowner            : Factor w/ 2 levels "N","Y": 2 2 2 1 1 2 1 1 2 2 ...
##  $ Credit_Score         : int  725 573 677 625 527 795 733 620 591 660 ...
##  $ Years_Credit_History : int  20 9 11 15 12 22 7 5 17 24 ...
##  $ Revolving_Balance    : num  11320 7200 20000 12800 5700 ...
##  $ Revolving_Utilization: num  0.25 0.7 0.55 0.65 0.75 0.12 0.2 0.62 0.5 0.35 ...
##  $ Decision             : Factor w/ 2 levels "Approve","Reject": 1 2 1 2 2 1 1 2 2 1 ...
```

1.1 Variables auxiliares

Se crean variables auxiliares con valores cero - uno. Por ejemplo

- HomeownerY = 1 si Homeowner = "Y"
- HomeownerY = 0 si Homeowner = "N"

```r
HomeownerY = ifelse(d$Homeowner == "Y", 1, 0)
```

El modelo que se estima es:

$$P(Decision_i = 1) = \frac{exp(\beta_0 + \beta_1 HomeownerY_i)}{1 + exp(\beta_0 + \beta_1 HomeownerY_i)}$$

Al final estamos trabajando con dos modelos:

- HomeownerY = 0:

$$P(Decision_i = 1) = \frac{exp(\beta_0)}{1 + exp(\beta_0)}$$

- HomeownerY = 1:

$$P(Decision_i = 1) = \frac{exp(\beta_0 + \beta_1)}{1 + exp(\beta_0 + \beta_1)}$$

Por tanto, β_1 representa diferencias de probabilidad de que se apruebe el crédito entre personas con casa y personas sin casa en propiedad.

- Si $\beta_1 = 0$, entonces P(Decision = 1 | Homeowner = Y) = P(Decision = 1 | Homeowner = N).
- Si $\beta_1 > 0$, entonces:

$$\beta_0 < \beta_0 + \beta_1 - \beta_0 > -\beta_0 - \beta_1 exp(-\beta_0) > exp(-\beta_0 - \beta_1) 1 + exp(-\beta_0) > 1 + exp(-\beta_0 - \beta_1) \frac{1}{1 + exp(-\beta_0)} < \frac{1}{1 + exp(-\beta_0 - \beta_1)}$$

Multiplicando numerador y denominador por el mismo número la desigualdad no cambia:

$$\frac{exp(\beta_0)}{1 + exp(\beta_0)} < \frac{exp(\beta_0 + \beta_1)}{1 + exp(\beta_0 + \beta_1)} P(Decision = 1 | Homeowner = N) < P(Decision = 1 | Homeowner = Y)$$

Por tanto, si $\beta_1 > 0$, la probabilidad de que se apruebe el crédito si se posee una casa es mayor que si no se posee.

Veamos que se obtiene con R. Recordemos que la variable respuesta toma valores 0 y 1. Por tanto si definimos:

```r
Decision1 = ifelse(d$Decision == "Approve", 1, 0)
```

el modelo que estamos estimando es P(Y = 1) = P(Decision1 = 1) = P(Decision = "Approve").

```r
m1 = glm(Decision1 ~ HomeownerY, family = binomial)
summary(m1)
```

```
##
## Call:
## glm(formula = Decision1 ~ HomeownerY, family = binomial)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.3514     0.7400  -3.177  0.00149 **
## HomeownerY    3.6041     0.8729   4.129 3.64e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 68.994  on 49  degrees of freedom
## Residual deviance: 42.194  on 48  degrees of freedom
## AIC: 46.194
##
## Number of Fisher Scoring iterations: 5
```

Como $\beta_1 > 0$ quiere decir que P(Decision = Approve|Homeowner = N) < P(Decision = Approve|Homeowner = Y).

1.2 Variables auxiliares 1

```
HomeownerN = ifelse(d$Homeowner == "N", 1, 0)
```

```
m2 = glm(Decision1 ~ HomeownerN, family = binomial)
summary(m2)
```

```
##
## Call:
## glm(formula = Decision1 ~ HomeownerN, family = binomial)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.2528     0.4629   2.706   0.0068 **
## HomeownerN   -3.6041     0.8729  -4.129 3.64e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 68.994  on 49  degrees of freedom
## Residual deviance: 42.194  on 48  degrees of freedom
## AIC: 46.194
##
## Number of Fisher Scoring iterations: 5
```

El modelo que se estima ahora es:

$$P(Decision_i = 1) = \frac{exp(\beta_0^* + \beta_1^* HomeownerN_i)}{1 + exp(\beta_0^* + \beta_1^* HomeownerN_i)}$$

Que equivale a dos modelos:

- HomeownerN = 0:

$$P(Decision_i = 1) = \frac{exp(\beta_0^*)}{1 + exp(\beta_0^*)}$$

- HomeownerN = 1:

$$P(Decision_i = 1) = \frac{exp(\beta_0^* + \beta_1^*)}{1 + exp(\beta_0^* + \beta_1^*)}$$

La probabilidad de que la Decision = “Approve” cuando el cliente es propietario de una casa tiene que ser igual en ambos modelos:

$$\frac{exp(\beta_0 + \beta_1)}{1 + exp(\beta_0 + \beta_1)} = \frac{exp(\beta_0^*)}{1 + exp(\beta_0^*)}$$

Por tanto

$$\beta_0 + \beta_1 = \beta_0^* \Rightarrow -2.3514 + 3.6041 = 1.2528$$

Por otro lado, la probabilidad de que la Decision = “Approve” cuando el cliente NO es propietario de una casa tiene que ser igual en ambos modelos:

$$P(Decision_i = 1|HomeownerY = 0) = P(Decision_i = 1|HomeownerN = 1)$$

$$\frac{exp(\beta_0)}{1 + exp(\beta_0)} = \frac{exp(\beta_0^* + \beta_1^*)}{1 + exp(\beta_0^* + \beta_1^*)}$$

Es decir

$$\beta_0 = \beta_0^* + \beta_1^* \Rightarrow -2.3514 = 1.2528 - 3.6041$$

1.3 Factores

Es importante utilizar bien los niveles de las variables:

```
levels(d$Decision)
```

```
## [1] "Approve" "Reject"
```

```
levels(d$Homeowner)
```

```
## [1] "N" "Y"
```

Por tanto, si estimamos el modelo:

```
m3 = glm(Decision ~ Homeowner, data = d, family = binomial)
summary(m3)
```

```
##
## Call:
## glm(formula = Decision ~ Homeowner, family = binomial, data = d)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.3514     0.7400   3.177  0.00149 **
## HomeownerY   -3.6041     0.8729  -4.129 3.64e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 68.994  on 49  degrees of freedom
## Residual deviance: 42.194  on 48  degrees of freedom
## AIC: 46.194
```

```
##
## Number of Fisher Scoring iterations: 5
```

En el fondo estamos estimando $P(\text{Decision} = \text{Reject}|\text{Homeowner} = \text{“Y”})$, ya que

- la variable respuesta tiene que tomar valores 1 y 0. R asigna el cero al nivel de referencia. Por lo tanto, $P(\text{Decision} = 1) = P(\text{Decision} = \text{Reject})$.
- para el regresor de tipo factor, R crea una variable auxiliar que vale 1 y 0, y asigna el cero al nivel de referencia. Por tanto crea HomeownerY, donde HomownerY = 1 cuando Homeowner = “Yes”.

Si queremos estimar el modelo m1 anterior tenemos que hacer:

```
d$Decision = relevel(d$Decision, ref = "Reject")
m4 = glm(Decision ~ Homeowner, data = d, family = binomial)
summary(m4)
```

```
##
## Call:
## glm(formula = Decision ~ Homeowner, family = binomial, data = d)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.3514     0.7400  -3.177  0.00149 **
## HomeownerY    3.6041     0.8729   4.129 3.64e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 68.994  on 49  degrees of freedom
## Residual deviance: 42.194  on 48  degrees of freedom
## AIC: 46.194
##
## Number of Fisher Scoring iterations: 5
```

2 Variables cualitativas y cuantitativas

El funcionamiento es idéntico que el modelo de regresión lineal:

```
d$Decision = relevel(d$Decision, ref = "Approve")
m4 = glm(Decision ~ Homeowner + Credit_Score + Years_Credit_History, data = d, family = binomial)
summary(m4)
```

```
##
## Call:
## glm(formula = Decision ~ Homeowner + Credit_Score + Years_Credit_History,
##     family = binomial, data = d)
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)          38.05066   14.85133   2.562   0.0104 *
## HomeownerY           -3.13441    1.73873  -1.803   0.0714 .
## Credit_Score         -0.04946    0.01933  -2.559   0.0105 *
## Years_Credit_History -0.31716    0.21846  -1.452   0.1466
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 68.994  on 49  degrees of freedom
## Residual deviance: 15.208  on 46  degrees of freedom
## AIC: 23.208
##
## Number of Fisher Scoring iterations: 8
```

3 Variables cualitativas con más de dos niveles

Igual que en regresión lineal, se tienen que crear tantas variables auxiliares como niveles del factor menos una.