INFORME ENCUESTA CONDICIONES DE VIDA 2016

Abstract

En este informe se parte de una base de datos que recoge información de 477 hogares en los que se estudian 17 variables siendo una de ellas el indicador de si la familia se encuentra en riesgo de pobreza. El objetivo es descubrir qué variables son más relevantes para la predicción del riesgo de pobreza y el establecimiento de un método que permita decidir de manera eficaz qué hogares se encuentran en el umbral de la pobreza y cuáles no.

Se comienza este análisis explicando brevemente las distintas variables presentes en la base de datos así como los posibles valores que estas pueden tomar. La variable a predecir es el riesgo de pobreza y es dicotómico, al igual que la posibilidad de irse de vacaciones una semana al año o la capacidad para afrontar gastos imprevistos. Otras variables categóricas son el género de la persona de mayor edad en la unidad familiar, su ocupación, su régimen de tenencia de la vivienda (propia, hipotecada, en alquiler o cedida), la facilidad para llegar a fin de mes (medida en seis grados), la región donde habitan y la posesión de ordenador o TV a color (según se tenga, no se deseé o no sea posible adquirirla por el hogar). Como variables cuantitativas aparecen la renta de la familia en el año anterior, la cuantía de las ayudas recibidas, la renta neta percibida por los menores de 16 años, el número de miembros de la familia, el número de miembros adultos, la edad del mayor y el número de horas de trabajo semanal del conjunto de miembros del hogar.

Antes de comenzar a trabajar se eliminan dos variables; el código de identificación de hogar (no aporta ninguna información pues no se desean conocer resultados para individuos concretos) y la renta percibida durante el año anterior a la entrevista (el riesgo de pobreza se calcula como el 60% de la mediana de los ingresos anuales por unidad de consumo, así no tiene sentido emplear la renta como variable predictiva pues se incurriría en una definición circular).

Tras la supresión de estas dos variables se transforman en factores las variables categóricas para facilitar el trabajo con ellas en R[1]. Se comprueba a continuación que todos los datos están correctamente cargados y que no existen huecos (no hay valores NA).

A las variables de las que ya se dispone se le añade una nueva pues quizás no resulte tan esencial la cantidad percibida como ayuda si no el hecho en sí de requerir o no una ayuda económica por parte de la familia, así se dicotomiza la variable Aid en Aid_D que será uno cuando la familia haya percibido una ayuda y cero en caso contrario.

A continuación se realiza un pequeño estudio mediante tablas de contingencia y los test de la chi cuadrado para observar qué variables parecen tener una mayor capacidad explicatoria sobre el riesgo de pobreza. El test devuelve una poderosa relación entre el riesgo de pobreza y la ocupación, el régimen de tenencia de la vivienda y la capacidad para afrontar gastos imprevistos. En este sentido se puede observar también por ejemplo que la cuantía de las ayudas no tiene relación con el riesgo pero sí el hecho de percibirlas o no. Por otra parte mediante un test ANOVA parece observarse una fuerte relación entre el riesgo de pobreza y el número de miembros de la familia. A partir de estos datos se construyen y comparan una serie de modelos.

[1] Todo este proceso se encuentra detallado y comentado en el código adjunto en el Anexo

En una primera aproximación se construye una regresión logística basada en todas las variables con el objeto de ver cuáles resultan a priori más significativas. Este análisis confirma la importancia de algunas variables como la dificultad para llegar a fin de mes, el sector laboral y el número de miembros pertenecientes a la familia. Si se construye un nuevo modelo con estas variables se obtiene un índice AIC de 309.57 algo menor que el del modelo general que era además más complejo. La posesión o no de un hogar así como la capacidad para asumir gastos inesperados parecen factores muy reveladores de la situación financiera de las unidades familiares. Un nuevo modelo construido con las variables antes destacadas y estas dos produce el resultado más satisfactorio con un AIC de 301.62 siendo significativas todas las variables. Otros factores como la posesión de un ordenador que parecían relevantes mediante la construcción de tablas de contingencia resultan no ser de tanta importancia como se pensaba y son descartados.

Este último modelo parece el más satisfactorio por lo que a continuación se profundiza sobre él. Para conocer la relevancia de cada variable se realiza un test anova. Este test devuelve p-valores inferiores a 0.05 para todas las variables luego ninguna de ellas es prescindible (si se retirara alguna se perdería una cantidad significativa de información). A la hora de estudiar cómo de relevante es cada variable se estudia la desviación de cada una, así se observa que la variable más relevante es el sector de ocupación (110.588) seguido por la facilidad para llegar a fin de mes (43.484), el régimen de tenencia de la vivienda (11.432), el número de miembros que conforman la unidad familiar (8.093) y por último la capacidad para afrontar gastos inesperados (6.519).

A la hora de medir la calidad predictiva del modelo un test interesante es el pseudo-R^2 de McFadden. Si se realiza este test sobre los diferentes modelos antes planteados se confirma que el mejor modelo es el elegido presentando un coeficiente de 0.4077, un resultado bastante bueno tratándose de este test.

Para testear la calidad predictiva del modelo dentro de nuestra base de datos esta se divide en dos submuestras. Una muestra de entrenamiento (muestra aleatoria simple) conformada por el 70% de los hogares y una muestra de validación conformada por el 30% de hogares restantes. Una vez tomadas las muestras y antes de comenzar a trabajar con ellas se comprueba que estén bien balanceadas en el sentido de que haya una proporción parecida de individuos con riesgo de pobreza en ambas. Tal y como se observa en el código las muestras están bien balanceadas luego se puede trabajar con ellas.

Así pues, se entrena el modelo con la muestra de entrenamiento y una vez entrenado se testea con la muestra de validación obtenido el siguiente resultado:

		PREDICTED	
		YES	*NO*
ACTUAL	*YES*	35	25
	NO	14	70

Figura 1. Matriz de confusión

Lo que implicaría un porcentaje de acierto del 73% siendo el mayor problema que algunos casos de riesgo no son detectados (falsos negativos). En esta línea se comprobó si quizá alguno de los otros modelos invertía esta proporción, pues considero que es más problemático no detectar situaciones de riesgo que generar falsas alarmas, pero los otros modelos devolvieron una proporción muy similar de falsos positivos y negativos y una precisión inferior.

Por último y para concluir el estudio se presenta a continuación (**Figura 2**) la curva ROC y el área debajo de la curva (AUC) que es otra de las medidas más habituales a la hora de baremar la calidad clasificatoria de un determinado modelo siendo la de este 0.891 lo cuál muestra una calidad bastante aceptable.

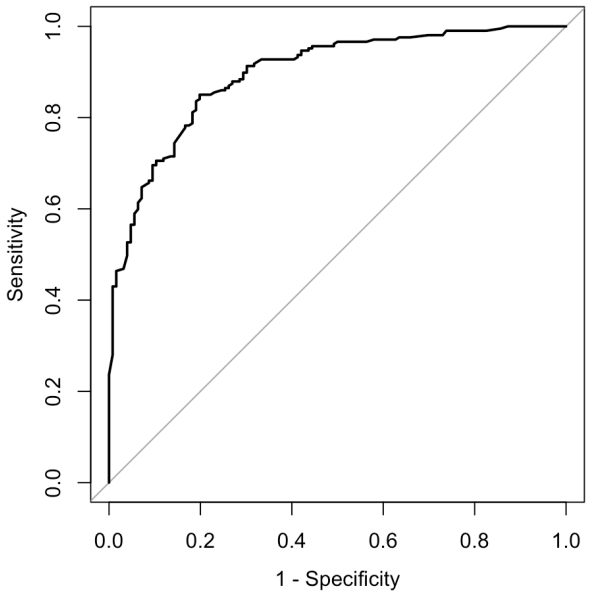


Figura 2. Curva ROC asociada a la regresión

Conclusión

En este informe se presentan como variables esenciales para la detección de riesgo de pobreza el sector empleador, el número de miembros de la familia, el régimen de tenencia de la vivienda, la capacidad para afrontar gastos imprevistos y la dificultad que supone llegar a fin de mes. Mediante el estudio de estas variables se construye un modelo que logra un 72% de precisión sobre la muestra de validación. Queda quizá cómo línea de continuación sobre este estudio lograr un modelo (quizá añadiendo alguna variable de todas las disponibles en la encuesta sobre condiciones de vida) que reduzca los falsos negativos pues cuando estos se producen se pone en peligro la seguridad económica de una familia lo que supone un riesgo nada desdeñable.

Código

El código anexo se encuentra también accesible para su interacción en el siguiente enlace.

Spanish_Surve_of_Life_Conditions.R

arturosanchezpalacio

Sun Dec 9 23:01:27 2018

```
# ASSIGNMENT 2. LOGISTIC REGRESSION

# Author: Arturo Sánchez Palacio
# Date: 9/XII/18

# The assignment works on a database extracted from 2016 Spanish Survey of Life Conditions.
# The goal of this assignment is to find out which families are in risk of poverty and why. (Which variables
are responsible?)

# Firstly, we load the data in R:

setwd("/Users/arturosanchezpalacio/Documents/CUNEF/Clasificación/Tareas/Tarea 2. Regresión Logística") #We s
et the working directory

library(openxlsx) #This library is required to read from an Excel file.
data <- read.xlsx("data.xlsx") #We charge the rough data and start working on it.

# We give the variables proper names:

names(data) <- c("ID", "Aid", "Minors_rent", "Holidays", "Unexpected_expenses", "TV", "Computer", "Ends_meet
",
                 "Home", "Members", "Rent", "Poverty_risk", "Region", "Age_older", "Working_hours", "Adults
",
                 "Gender_older", "Occupation")

# We have a first glance at the data:

dim(data) #We have 418 homes in this study in which 18 variables are observed.
```

```
## [1] 477  18
```

```
str(data) #All data has been loades as numeric.
```

```
## 'data.frame':    477 obs. of  18 variables:
##  $ ID                  : num  9 66 97 138 183 208 264 294 307 326 ...
##  $ Aid                 : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Minors_rent         : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Holidays            : num  2 2 1 1 2 1 2 2 2 1 ...
##  $ Unexpected_expenses: num  2 2 1 1 2 1 2 2 2 2 ...
##  $ TV                  : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ Computer            : num  1 1 1 1 1 1 1 3 1 1 ...
##  $ Ends_meet           : num  2 2 4 5 2 5 3 2 1 3 ...
##  $ Home                : num  3 2 1 5 3 1 1 1 3 2 ...
##  $ Members             : num  3 3 3 2 2 2 3 2 4 3 ...
##  $ Rent                : num  12672 19853 54832 32373 10893 ...
##  $ Poverty_risk        : num  1 0 0 0 1 0 1 1 1 0 ...
##  $ Region              : chr  "ES21" "ES42" "ES52" "ES52" ...
##  $ Age_older           : num  33 50 51 52 39 50 46 47 37 49 ...
##  $ Working_hours       : num  4 40 45 40 0 40 0 0 59 40 ...
##  $ Adults              : num  1 2 1 1 2 1 2 2 1 2 ...
##  $ Gender_older        : num  0 0 1 0 0 0 0 0 0 0 ...
##  $ Occupation          : num  2 1 1 1 1 3 5 5 1 1 ...
```

```r
# The ID does not give any interesting information:

data <- data[,-1]

# Poverty_risk is the variable we are going to try to classify:

data$Poverty_risk <- factor(data$Poverty_risk, levels = c(1,0), labels = c("Sí", "No"))

# This index is built upon the rent so it makes no sense to use the rent to predict it.  (It would be
# a circular definition.  We erase this variable too)

data <- data[, -10]

# Some variables must be translated to factors:

data$Aid_D <- ifelse(data$Aid > 0, 1, 0)
data$Aid_D <- as.factor(data$Aid_D)
data$Holidays <- as.factor(data$Holidays)
data$Unexpected_expenses <- as.factor(data$Unexpected_expenses)
data$TV <- as.factor(data$TV)
data$Computer <- as.factor(data$Computer)
data$Ends_meet <- as.factor(data$Ends_meet)
data$Home <- as.factor(data$Home)
data$Occupation <- as.factor(data$Occupation)

data$Gender_older <- factor(data$Gender_older, levels = c(0,1), labels = c("Mujer", "Hombre"))
data$Region <- factor(data$Region, levels = c("ES21", "ES42", "ES52", "ES61", "ES41", "ES43", "ES53", "ES51
",
                                              "ES11", "ES23", "ES30", "ES62", "ES22", "ES12", "ES70", "ES1
3", "ES24", "ES64", "ES63"),
                      labels = c("País Vasco", "Castilla la Mancha", "C. Valenciana", "Andalucía", "Castill
a León",
                                 "Extremandura", "Baleares", "Cataluña", "Galicia", "Rioja", "Madrid", "Mur
cia",
                                 "Navarra", "Asturias", "Canarias", "Cantabria", "Aragón", "Melilla", "Ceut
a"))

str(data) #Now the data seems like something we can work with
```

```
## 'data.frame':    477 obs. of  17 variables:
##  $ Aid                : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Minors_rent        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Holidays           : Factor w/ 2 levels "1","2": 2 2 1 1 2 1 2 2 2 1 ...
##  $ Unexpected_expenses: Factor w/ 2 levels "1","2": 2 2 1 1 2 1 2 2 2 2 ...
##  $ TV                 : Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Computer           : Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 3 1 1 ...
##  $ Ends_meet          : Factor w/ 6 levels "1","2","3","4",..: 2 2 4 5 2 5 3 2 1 3 ...
##  $ Home               : Factor w/ 5 levels "1","2","3","4",..: 3 2 1 5 3 1 1 1 3 2 ...
##  $ Members            : num  3 3 3 2 2 2 3 2 4 3 ...
##  $ Poverty_risk       : Factor w/ 2 levels "Sí","No": 1 2 2 2 1 2 1 1 1 2 ...
##  $ Region             : Factor w/ 19 levels "País Vasco","Castilla la Mancha",..: 1 2 3 3 4 5 6 6 7 7 ..
.
##  $ Age_older          : num  33 50 51 52 39 50 46 47 37 49 ...
##  $ Working_hours      : num  4 40 45 40 0 40 0 0 59 40 ...
##  $ Adults             : num  1 2 1 1 2 1 2 2 1 2 ...
##  $ Gender_older       : Factor w/ 2 levels "Mujer","Hombre": 1 1 2 1 1 1 1 1 1 1 ...
##  $ Occupation         : Factor w/ 9 levels "1","2","3","4",..: 2 1 1 1 1 3 5 5 1 1 ...
##  $ Aid_D              : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
```

```r
sapply(data,function(x) sum(is.na(x))) #There are no unknown values in this dataset.
```

```
##                 Aid         Minors_rent            Holidays 
##                   0                   0                   0 
## Unexpected_expenses                  TV            Computer 
##                   0                   0                   0 
##           Ends_meet                Home             Members 
##                   0                   0                   0 
##        Poverty_risk              Region           Age_older 
##                   0                   0                   0 
##       Working_hours              Adults        Gender_older 
##                   0                   0                   0 
##          Occupation              Aid_D                     
##                   0                   0
```

```
# Now we are going to try to find and justify which variables will be used in the construction of the logisti
c regression:
# In order to do that we use Anova and Chi-Squared tests on contingency tables.
corr_pov_members <- lm(data$Members ~ data$Poverty_risk, data = data)
anova(corr_pov_members)
```

```
## Analysis of Variance Table
##
## Response: data$Members
##                     Df  Sum Sq Mean Sq F value   Pr(>F)   
## data$Poverty_risk   1   2.581 2.58092  6.7004 0.009935 **
## Residuals         475 182.966 0.38519                    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
library(vcd)
```

```
## Loading required package: grid
```

```
tab <- xtabs(~ data$Home + data$Poverty_risk, data = data)
summary(assocstats(tab))
```

```
##
## Call: xtabs(formula = ~data$Home + data$Poverty_risk, data = data)
## Number of cases in table: 477
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 22.655, df = 4, p-value = 0.0001484
##                     X^2 df   P(> X^2)
## Likelihood Ratio 22.375  4 0.00016878
## Pearson          22.655  4 0.00014836
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.213
## Cramer's V        : 0.218
```

```
tab1 <- xtabs(~ data$Occupation + data$Poverty_risk, data = data)
summary(assocstats(tab1))
```

```
##
## Call: xtabs(formula = ~data$Occupation + data$Poverty_risk, data = data)
## Number of cases in table: 477
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 133.98, df = 8, p-value = 4.221e-25
##  Chi-squared approximation may be incorrect
##                     X^2 df P(> X^2)
## Likelihood Ratio 141.71  8        0
## Pearson          133.98  8        0
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.468
## Cramer's V        : 0.53
```

```
tab2 <- xtabs(~ data$Ends_meet + data$Poverty_risk, data = data)
summary(assocstats(tab2))
```

```
##
## Call: xtabs(formula = ~data$Ends_meet + data$Poverty_risk, data = data)
## Number of cases in table: 477
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 102.37, df = 5, p-value = 1.67e-20
##  Chi-squared approximation may be incorrect
##                     X^2 df P(> X^2)
## Likelihood Ratio 118.56  5        0
## Pearson          102.37  5        0
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.42
## Cramer's V        : 0.463
```

```
tab3 <- xtabs(~ data$Unexpected_expenses + data$Poverty_risk, data = data)
summary(assocstats(tab3))
```

```
##
## Call: xtabs(formula = ~data$Unexpected_expenses + data$Poverty_risk,
##     data = data)
## Number of cases in table: 477
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 84.53, df = 1, p-value = 3.777e-20
##                     X^2 df P(> X^2)
## Likelihood Ratio 89.666  1        0
## Pearson          84.534  1        0
##
## Phi-Coefficient   : 0.421
## Contingency Coeff.: 0.388
## Cramer's V        : 0.421
```

```
tab4 <- xtabs(~ data$TV + data$Poverty_risk, data = data)
summary(assocstats(tab4))
```

```
##
## Call: xtabs(formula = ~data$TV + data$Poverty_risk, data = data)
## Number of cases in table: 477
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 2.2041, df = 2, p-value = 0.3322
##  Chi-squared approximation may be incorrect
##                     X^2 df P(> X^2)
## Likelihood Ratio 2.8724  2  0.23784
## Pearson          2.2041  2  0.33219
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.068
## Cramer's V        : 0.068
```

```
tab5 <- xtabs(~ data$Computer + data$Poverty_risk, data = data)
summary(assocstats(tab5))
```

```
##
## Call: xtabs(formula = ~data$Computer + data$Poverty_risk, data = data)
## Number of cases in table: 477
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 30.074, df = 2, p-value = 2.948e-07
##                     X^2 df   P(> X^2)
## Likelihood Ratio 29.537  2 3.8563e-07
## Pearson          30.074  2 2.9475e-07
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.244
## Cramer's V        : 0.251
```

```
tab6 <- xtabs(~ data$Aid_D + data$Poverty_risk, data = data)
summary(assocstats(tab6))
```

```
##
## Call: xtabs(formula = ~data$Aid_D + data$Poverty_risk, data = data)
## Number of cases in table: 477
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 7.203, df = 1, p-value = 0.007276
##                     X^2 df  P(> X^2)
## Likelihood Ratio 7.0055  1 0.0081260
## Pearson          7.2034  1 0.0072764
##
## Phi-Coefficient   : 0.123
## Contingency Coeff.: 0.122
## Cramer's V        : 0.123
```

```
# We set a seed in order to be able to replicate the experiment:

set.seed(1234)

# We split the data into two chunks: training and testing set.
# The training set will be used to fit our model which we will be testing over the testing set.

train <- sample(nrow(data), 0.7*nrow(data))

data.train <- data[train,]

data.validate <- data[-train,]

#We check that the samples are balanced:

table(data.train$Poverty_risk)
```

```
##
##  Sí  No
## 126 207
```

```
table(data.validate$Poverty_risk)
```

```
##
## Sí No
## 60 84
```

```
model1 <- glm(Poverty_risk ~.,family = binomial(link = 'logit'),data = data.train)
```

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

```
summary(model1)
```

```
##
## Call:
## glm(formula = Poverty_risk ~ ., family = binomial(link = "logit"),
##     data = data.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6423  -0.4837   0.1013   0.5098   2.5541
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                3.489e+00  1.949e+00   1.790 0.073387 .
## Aid                       -7.700e-02  6.196e+00  -0.012 0.990084
## Minors_rent                1.574e-04  3.282e-04   0.480 0.631475
## Holidays2                 -3.714e-02  4.761e-01  -0.078 0.937824
## Unexpected_expenses2      -7.313e-01  4.874e-01  -1.500 0.133511
## TV3                       -8.796e+01  3.622e+04  -0.002 0.998062
## Computer2                  2.735e-01  6.191e-01   0.442 0.658677
## Computer3                  5.792e-01  8.756e-01   0.661 0.508310
## Ends_meet2                 1.311e+00  4.912e-01   2.670 0.007596 **
## Ends_meet3                 1.526e+00  5.296e-01   2.881 0.003960 **
## Ends_meet4                 1.947e+00  8.029e-01   2.424 0.015332 *
## Ends_meet5                 8.810e+01  6.630e+03   0.013 0.989399
## Ends_meet6                -2.270e+01  2.923e+04  -0.001 0.999380
## Home2                      3.632e-01  4.773e-01   0.761 0.446669
## Home3                     -1.313e+00  5.492e-01  -2.391 0.016785 *
## Home4                      1.212e+00  1.298e+00   0.933 0.350572
## Home5                     -5.551e-01  6.830e-01  -0.813 0.416369
## Members                   -8.513e-01  3.147e-01  -2.705 0.006829 **
## RegionCastilla la Mancha -8.785e-01  1.475e+00  -0.596 0.551424
## RegionC. Valenciana      -8.828e-01  1.093e+00  -0.808 0.419357
## RegionAndalucía          -1.364e+00  1.031e+00  -1.323 0.185796
## RegionCastilla León      -5.606e-01  1.196e+00  -0.469 0.639381
## RegionExtremandura       -2.081e+00  1.371e+00  -1.518 0.128982
## RegionBaleares           -1.156e+00  1.275e+00  -0.907 0.364440
## RegionCataluña           -2.032e-01  9.485e-01  -0.214 0.830387
## RegionGalicia            -1.676e+00  1.252e+00  -1.338 0.180943
## RegionRioja               2.126e+01  2.054e+04   0.001 0.999174
## RegionMadrid             -1.774e-02  9.848e-01  -0.018 0.985626
## RegionMurcia             -1.532e+00  1.255e+00  -1.221 0.222091
## RegionNavarra            -2.889e-02  1.738e+00  -0.017 0.986737
## RegionAsturias           -2.477e+00  1.423e+00  -1.741 0.081692 .
## RegionCanarias           -7.013e-01  1.270e+00  -0.552 0.580650
## RegionCantabria          -3.213e+00  1.529e+00  -2.101 0.035633 *
## RegionAragón              9.915e-03  1.367e+00   0.007 0.994215
## RegionMelilla            -1.920e+00  3.129e+00  -0.614 0.539426
## RegionCeuta              -8.234e-01  2.257e+00  -0.365 0.715231
## Age_older                -2.473e-04  2.547e-02  -0.010 0.992251
## Working_hours             1.317e-02  1.845e-02   0.714 0.475403
## Adults                    9.416e-02  3.321e-01   0.284 0.776750
## Gender_olderHombre       -3.356e-01  5.707e-01  -0.588 0.556463
## Occupation2              -2.316e+00  6.481e-01  -3.573 0.000352 ***
## Occupation3              -1.724e+00  7.695e-01  -2.241 0.025047 *
## Occupation4               2.081e+01  2.034e+04   0.001 0.999184
## Occupation5              -2.474e+00  8.734e-01  -2.832 0.004620 **
## Occupation7               1.988e+01  1.091e+04   0.002 0.998546
## Occupation8              -1.247e+00  1.107e+00  -1.127 0.259895
## Occupation10             -5.448e-01  9.901e-01  -0.550 0.582131
## Occupation11             -1.034e+00  1.817e+00  -0.569 0.569563
## Aid_D1                    2.145e+01  1.797e+03   0.012 0.990473
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 441.73  on 332  degrees of freedom
## Residual deviance: 233.09  on 284  degrees of freedom
## AIC: 331.09
##
## Number of Fisher Scoring iterations: 20
```

```
#Regions don't seem important at all. How troubled is to make ends meet, the occupation and the number of me
mbers
#seem like essential facts.

model2 <- glm(Poverty_risk ~ Occupation + Members + Ends_meet,family = binomial(link = 'logit'),data = data.
train)
```

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

```
summary(model2)
```

```
##
## Call:
## glm(formula = Poverty_risk ~ Occupation + Members + Ends_meet,
##     family = binomial(link = "logit"), data = data.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4909  -0.7059   0.3033   0.6378   2.3502
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)     2.2119     0.6948   3.183  0.00146 **
## Occupation2    -1.9495     0.4350  -4.481 7.42e-06 ***
## Occupation3    -0.8948     0.6297  -1.421  0.15528
## Occupation4    16.8839  3295.0931   0.005  0.99591
## Occupation5    -2.6209     0.3931  -6.667 2.61e-11 ***
## Occupation7    16.6285  2461.2046   0.007  0.99461
## Occupation8    -1.5449     0.7658  -2.017  0.04365 *
## Occupation10   -0.7444     0.6029  -1.235  0.21693
## Occupation11   -0.7525     1.0931  -0.688  0.49116
## Members        -0.7625     0.2498  -3.053  0.00227 **
## Ends_meet2      1.0543     0.4000   2.636  0.00839 **
## Ends_meet3      1.5650     0.3943   3.969 7.21e-05 ***
## Ends_meet4      2.3695     0.5655   4.190 2.79e-05 ***
## Ends_meet5     18.3798  1102.8443   0.017  0.98670
## Ends_meet6    -18.4904  6522.6386  -0.003  0.99774
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 441.73  on 332  degrees of freedom
## Residual deviance: 279.57  on 318  degrees of freedom
## AIC: 309.57
##
## Number of Fisher Scoring iterations: 17
```

```
#The AIC is lower so by simplifying the regression we have also improved it.

model3 <- glm(Poverty_risk ~ Occupation + Members + Ends_meet + Unexpected_expenses,family = binomial(link =
'logit'),data = data.train)
```

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

```
summary(model3)
```

```
##
## Call:
## glm(formula = Poverty_risk ~ Occupation + Members + Ends_meet +
##     Unexpected_expenses, family = binomial(link = "logit"), data = data.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4805  -0.6351   0.3073   0.6257   2.3995
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)             2.8467     0.7557   3.767 0.000165 ***
## Occupation2            -1.8347     0.4440  -4.132 3.59e-05 ***
## Occupation3            -0.9048     0.6366  -1.421 0.155275
## Occupation4            17.1918  3244.5624   0.005 0.995772
## Occupation5            -2.5105     0.3988  -6.295 3.08e-10 ***
## Occupation7            16.0342  2561.2975   0.006 0.995005
## Occupation8            -1.2091     0.7696  -1.571 0.116151
## Occupation10           -0.6272     0.6162  -1.018 0.308737
## Occupation11           -0.5038     1.0988  -0.459 0.646563
## Members                -0.7273     0.2529  -2.876 0.004028 **
## Ends_meet2              1.1144     0.4024   2.769 0.005618 **
## Ends_meet3              1.3223     0.4112   3.216 0.001300 **
## Ends_meet4              1.6372     0.6271   2.611 0.009037 **
## Ends_meet5             17.6652  1107.6121   0.016 0.987275
## Ends_meet6            -19.2308  6522.6386  -0.003 0.997648
## Unexpected_expenses2   -0.9751     0.3854  -2.530 0.011394 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 441.73  on 332  degrees of freedom
## Residual deviance: 273.05  on 317  degrees of freedom
## AIC: 305.05
##
## Number of Fisher Scoring iterations: 17
```

```
#Unexpected expenses seem like an interesting fact and they lower the AIC so it's good.
#The biggest indicator is when families are able to overcome unexpected expenses.

model4 <- glm(Poverty_risk ~ Occupation + Members + Ends_meet + Unexpected_expenses + Home,family = binomial
(link = 'logit'),
              data = data.train)
```

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

```
summary(model4)
```

```
##
## Call:
## glm(formula = Poverty_risk ~ Occupation + Members + Ends_meet +
##     Unexpected_expenses + Home, family = binomial(link = "logit"),
##     data = data.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5839  -0.5750   0.2412   0.5595   2.2728
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)             3.0651     0.8372   3.661 0.000251 ***
## Occupation2            -2.0646     0.4673  -4.418 9.95e-06 ***
## Occupation3            -1.2202     0.6743  -1.809 0.070376 .
## Occupation4            16.8811  3447.6995   0.005 0.996093
## Occupation5            -2.6115     0.4228  -6.177 6.52e-10 ***
## Occupation7            15.6714  2598.5287   0.006 0.995188
## Occupation8            -1.3747     0.7899  -1.740 0.081785 .
## Occupation10           -0.7331     0.6627  -1.106 0.268653
## Occupation11           -1.3883     1.2683  -1.095 0.273685
## Members                -0.7704     0.2627  -2.932 0.003365 **
## Ends_meet2              1.2760     0.4212   3.029 0.002452 **
## Ends_meet3              1.5091     0.4384   3.443 0.000576 ***
## Ends_meet4              1.7298     0.6498   2.662 0.007765 **
## Ends_meet5             17.6325  1104.4376   0.016 0.987262
## Ends_meet6            -19.3200  6522.6386  -0.003 0.997637
## Unexpected_expenses2   -0.8976     0.3989  -2.250 0.024433 *
## Home2                   0.2686     0.4213   0.637 0.523803
## Home3                  -1.0251     0.4600  -2.228 0.025863 *
## Home4                   1.0352     1.0697   0.968 0.333169
## Home5                  -0.4846     0.5861  -0.827 0.408368
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 441.73  on 332  degrees of freedom
## Residual deviance: 261.62  on 313  degrees of freedom
## AIC: 301.62
##
## Number of Fisher Scoring iterations: 17
```

```
# As a last idea receiving an aid (independently on how big it is) seems like an important factor. However i
t looks like it is not:

model5 <- glm(Poverty_risk ~ Occupation + Members + Ends_meet + Unexpected_expenses + Home + Aid_D,
              family = binomial(link = 'logit'),
              data = data.train)
```

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

```
summary(model5)
```

```
##
## Call:
## glm(formula = Poverty_risk ~ Occupation + Members + Ends_meet +
##     Unexpected_expenses + Home + Aid_D, family = binomial(link = "logit"),
##     data = data.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6076  -0.5999   0.2353   0.5713   2.2744
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)             3.0855     0.8424   3.663 0.000250 ***
## Occupation2            -1.9889     0.4726  -4.208 2.57e-05 ***
## Occupation3            -1.2674     0.6766  -1.873 0.061053 .
## Occupation4            16.8166  3453.6003   0.005 0.996115
## Occupation5            -2.6028     0.4241  -6.137 8.42e-10 ***
## Occupation7            15.6907  2593.2406   0.006 0.995172
## Occupation8            -1.4010     0.7888  -1.776 0.075740 .
## Occupation10           -0.7447     0.6633  -1.123 0.261563
## Occupation11           -1.1556     1.2915  -0.895 0.370875
## Members                -0.7924     0.2654  -2.986 0.002828 **
## Ends_meet2              1.2603     0.4222   2.985 0.002834 **
## Ends_meet3              1.5434     0.4421   3.491 0.000481 ***
## Ends_meet4              1.7506     0.6517   2.686 0.007226 **
## Ends_meet5             17.6895  1100.8730   0.016 0.987180
## Ends_meet6            -19.2743  6522.6386  -0.003 0.997642
## Unexpected_expenses2   -0.8432     0.4014  -2.101 0.035645 *
## Home2                   0.3217     0.4251   0.757 0.449253
## Home3                  -1.0307     0.4613  -2.234 0.025481 *
## Home4                   1.1107     1.0633   1.045 0.296197
## Home5                  -0.5128     0.5849  -0.877 0.380663
## Aid_D1                 -1.0659     0.8779  -1.214 0.224684
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 441.73  on 332  degrees of freedom
## Residual deviance: 260.01  on 312  degrees of freedom
## AIC: 302.01
##
## Number of Fisher Scoring iterations: 17
```

```
#Home also appears to be an interesting variable. (Appears with significance in the case of renting).
#This will be our definite model.

anova(model4, test = "Chisq")
```

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

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

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Poverty_risk
##
## Terms added sequentially (first to last)
##
##
##                     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                  332     441.73
## Occupation           8  110.588       324     331.15 < 2.2e-16 ***
## Members              1    8.093       323     323.05  0.004443 **
## Ends_meet            5   43.484       318     279.57 2.947e-08 ***
## Unexpected_expenses  1    6.519       317     273.05  0.010672 *
## Home                 4   11.432       313     261.62  0.022112 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# The bigger the deviance, the more important the variable is. So in order to classify, first the occupation
, then the
# ability to make ends meet, if the home is owned or rented, the number of members of the family and last th
e ability to
# face unexpected expenses.

# As we can see, all the p-values are significant so all the variables are useful in order to explain the va
riable.

library(pscl)
```

```
## Classes and Methods for R developed in the
## Political Science Computational Laboratory
## Department of Political Science
## Stanford University
## Simon Jackman
## hurdle and zeroinfl functions by Achim Zeileis
```

```
pR2(model1)
```

```
##          llh      llhNull           G2     McFadden         r2ML
## -116.5457202 -220.8671387  208.6428369    0.4723266    0.4655703
##         r2CU
##    0.6337711
```

```
pR2(model2)
```

```
##          llh      llhNull           G2     McFadden         r2ML
## -139.7843382 -220.8671387  162.1656010    0.3671112    0.3855230
##         r2CU
##    0.5248044
```

```
pR2(model3)
```

```
##          llh      llhNull           G2     McFadden         r2ML
## -136.5247901 -220.8671387  168.6846972    0.3818692    0.3974356
##         r2CU
##    0.5410207
```

```
pR2(model4)
```

```
##          llh      llhNull           G2     McFadden         r2ML
## -130.8086306 -220.8671387  180.1170161    0.4077497    0.4177713
##         r2CU
##    0.5687033
```

```
fitted.results <- predict(model4, newdata = data.validate, type = 'response')
fitted.results <- ifelse(fitted.results > 0.5,1,0)
fitted.results <- factor(fitted.results, levels = c(0,1), labels = c("Sí", "No"))
(logit.perf <- table(data.validate$Poverty_risk, fitted.results, dnn = c("Actual", "Predicted")))
```

```
##       Predicted
## Actual Sí No
##     Sí 35 25
##     No 14 70
```

```
print(paste('Accuracy',sum(diag(logit.perf))/sum(logit.perf)))
```

```
## [1] "Accuracy 0.729166666666667"
```

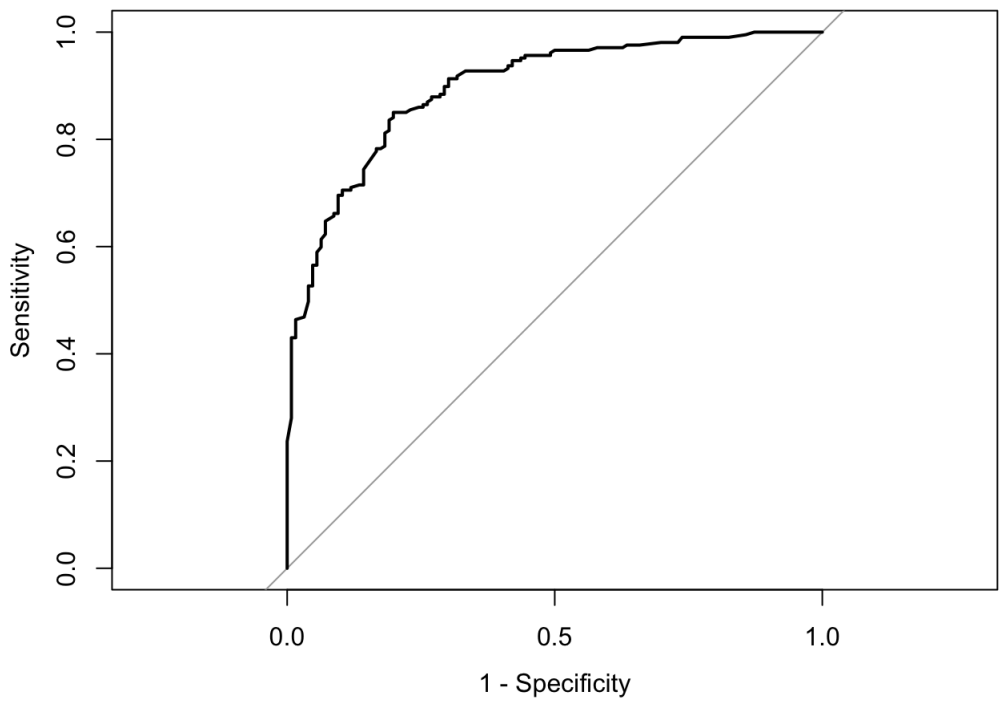
```
# ROC curve
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
resRoc <- roc(data.train$Poverty_risk ~ model4$fitted.values)
plot(resRoc, legacy.axes = TRUE)
```




```
# Area Under the Curve (AUC)

library(rms)
```

```
## Loading required package: Hmisc
```

```
## Loading required package: lattice
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
## Loading required package: SparseM
```

```
##
## Attaching package: 'SparseM'
```

```
## The following object is masked from 'package:base':
##
##     backsolve
```

```
resLrm <- lrm(formula = Poverty_risk ~ Occupation + Members + Ends_meet + Unexpected_expenses + Home,
              data    = data.validate,
              x = TRUE, y = TRUE)
resLrm
```

```
## Logistic Regression Model
##
##  lrm(formula = Poverty_risk ~ Occupation + Members + Ends_meet +
##      Unexpected_expenses + Home, data = data.validate, x = TRUE,
##      y = TRUE)
##
##                        Model Likelihood     Discrimination    Rank Discrim.
##                           Ratio Test           Indexes           Indexes
##  Obs            144    LR chi2      80.32    R2       0.575    C       0.891
##   Sí             60    d.f.            18    g        4.247    Dxy     0.782
##   No             84    Pr(> chi2) <0.0001    gr      69.870    gamma   0.786
##  max |deriv| 0.001                          gp       0.382    tau-a   0.383
##                                             Brier    0.132
##
##                         Coef     S.E.     Wald Z Pr(>|Z|)
##  Intercept               1.4213   1.3394  1.06  0.2886
##  Occupation=2           -1.9355   0.7920 -2.44  0.0145
##  Occupation=3           -1.4453   0.7423 -1.95  0.0515
##  Occupation=5           -1.9821   0.5933 -3.34  0.0008
##  Occupation=7            8.3607 101.7065  0.08  0.9345
##  Occupation=8          -10.9316  67.7318 -0.16  0.8718
##  Occupation=10         -11.6134  55.2956 -0.21  0.8336
##  Occupation=11           6.5357 101.7075  0.06  0.9488
##  Members                -0.1267   0.4126 -0.31  0.7588
##  Ends_meet=2            -0.4771   0.6185 -0.77  0.4405
##  Ends_meet=3             1.5403   0.6816  2.26  0.0238
##  Ends_meet=4             1.2079   0.9158  1.32  0.1872
##  Ends_meet=5             8.3306  31.7395  0.26  0.7930
##  Ends_meet=6             7.8119  70.8527  0.11  0.9122
##  Unexpected_expenses=2  -1.0081   0.6570 -1.53  0.1250
##  Home=2                  0.7166   0.6209  1.15  0.2485
##  Home=3                  0.7234   0.6627  1.09  0.2751
##  Home=4                 -8.7644  39.8801 -0.22  0.8261
##  Home=5                  0.8044   0.8962  0.90  0.3694
##
```