Análisis de Datos Categóricos

Alumno:
Huertas Quispe, Anthony Enrique ***Cod:*** *20173728*

Semestre: *2017-II*

Tema: PC 2

Prof. Victor Giancarlo Sal Y Rosas

Pontificia Universidad Católica del Perú
Escuela de Posgrado
Maestría en Estadística

Ejercicio 1 (Regresión Multinomial)

Deseamos entender la relación entre el nivel de satisfacción en la atención en el centro de salud (C1P40) y tipo de institución (INSTITUCION), si tenía cita programada o no (C1P5), edad (C1P76EDAD), sexo (C1P77), tener seguro o no (C1P47), quintil de riqueza (C1INDICERIQUEZA) y el tiempo que dura la atención (C1P14).

Listing 1: Lectura de datos.

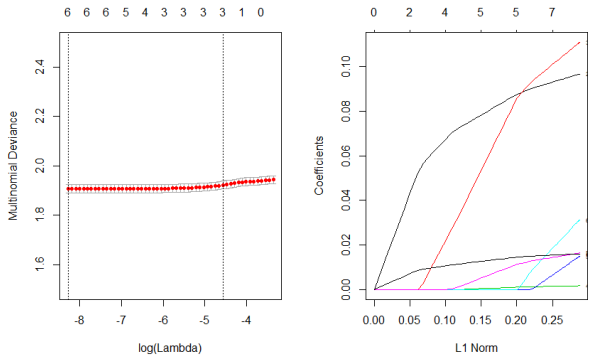
```
#Leyendo la base de datos
salud.spss  = read.spss(file.choose(),to.data.frame=FALSE,use.value.labels = FALSE)

#Seleccionando las variables en estudio
salud = as.data.frame(salud.spss)
salud=salud[,c(143,17,40,252,253,153,312,56)]
salud=na.omit(salud) #Se omiten los datos con NA valores en sus variables
head(salud)
salud.var   = attr(salud.spss, "variable.labels")
salud.label = attr(salud.spss, "label.table")
```

Listing 2: Selección de variables usando LASSO.

```
#Regresion Lasso 1
lasso<- model.matrix(~INSTITUCION+C1P5+C1P76EDAD+C1P77+C1P47+C1INDICERIQUEZA+C1P14,
    salud)
y  = salud$C1P40
cv1  = cv.glmnet(lasso,y,alpha=1,family='multinomial',
type.measure = "auc")
# Salidas
coef(cv1,s="lambda.min")

#Visualizacion
tray1    = glmnet(lasso,y=salud$C1P40,alpha=1)
x1 = par(mfrow=c(1,2))
plot(cv1,ylim=c(1.5,2.5))
plot(tray1,label=TRUE)
par(x1)
```



Selecciono las variables categóricas INSTITUCIÓN, C1P5 y la variable continua C1P14.

a) Construya un modelo multinomial para explicar que variables están asociadas con satisfacción. Intreprete sus resultados mas importantes.

Listing 3: Modelo Multinomial.

```
salud = as.data.frame(salud.spss)
salud=salud[,c(143,17,40,252,253,153,312,56)]
salud.var   = attr(salud.spss, "variable.labels")
salud.label = attr(salud.spss, "label.table")

with(salud,table(C1P40,useNA="always"))
with(salud,round(100*prop.table(table(C1P40,useNA="always")),2))

salud.label$C1P40

mod1 <-multinom(C1P40~as.factor(INSTITUCION)+as.factor(C1P5)+C1P14,data=salud)
summary(mod1)
```

```
> summary(mod1)
Call:
multinom(formula = C1P40 ~ as.factor(INSTITUCION) + as.factor(C1P5) +
    C1P14, data = salud)

Coefficients:
  (Intercept) as.factor(INSTITUCION)2 as.factor(INSTITUCION)3
2    1.813020               0.6022535              -0.2085697
3    2.152066               0.5990996               0.4614664
4    2.789416               0.6356560               1.0917569
5   -1.209338               0.8909508               2.3140468
  as.factor(INSTITUCION)4 as.factor(C1P5)2      C1P14
2              -1.3117022        0.3985915 0.03342639
3              -0.4644431        0.6934507 0.08141182
4               1.0435860        0.7137182 0.10846553
5               2.2436218        0.6830396 0.15281666

Std. Errors:
  (Intercept) as.factor(INSTITUCION)2 as.factor(INSTITUCION)3
2   0.3158380               0.2360495               0.7627805
3   0.3079469               0.2306833               0.7286058
4   0.3048063               0.2283949               0.7217082
5   0.3438326               0.2640757               0.7436759
  as.factor(INSTITUCION)4 as.factor(C1P5)2      C1P14
2               0.8313292        0.2528510 0.02757946
3               0.7393473        0.2464149 0.02693668
4               0.7244783        0.2441156 0.02673926
5               0.7401959        0.2725671 0.02804596

Residual Deviance: 25930.3
AIC: 25978.3
```

INTERPRETACIONES: Para un paciente que se atiende en la Institucion 3, el odds estimado de que se encuentre satisfecho en lugar de muy insatisfecho está asociado en 3 veces el odds estimado para un paciente que se atiende en la Institución 1. El mayor riesgo relativo se presenta en esta institución con respecto a lo indicado.

Condicionando las variables de institución y el tiempo que dura la atención, para pacientes que no tenían cita el riesgo relativo de estar muy satisfecho vs muy insatisfecho aumenta en 30 % veces en comparación con una persona que tenía cita.

b) Construya un modelo ordinal tipo categoría adjacente para explicar que variables estan asociadas con el nivel de satisfacción. Intreprete sus resultados mas importantes.

Listing 4: Modelo ordinal categoría adjacente.

```
salud$muyinsat = ifelse(salud$C1P40==1,1,0)
salud$insat = ifelse(salud$C1P40==2,1,0)
salud$nisatinsat = ifelse(salud$C1P40==3,1,0)
salud$sat = ifelse(salud$C1P40==4,1,0)
salud$muysat = ifelse(salud$C1P40==5,1,0)

modelord = vglm(cbind(muyinsat,insat,nisatinsat,sat,muysat)~factor(INSTITUCION)
+ factor(C1P5)+C1P14,family = acat(parallel = TRUE),data=salud)
summary(modelord)
```

```
Call:
vglm(formula = cbind(muyinsat, insat, nisatinsat, sat, muysat) ~
    factor(INSTITUCION) + factor(C1P5) + C1P14, family = acat(parallel = TRUE),
    data = salud)


Pearson residuals:
                          Min        1Q   Median       3Q    Max
loge(P[Y=2]/P[Y=1]) -107.9316  0.008102  0.01042  0.03958 1.5840
loge(P[Y=3]/P[Y=2])  -13.5258  0.105874  0.14174  0.17842 1.1033
loge(P[Y=4]/P[Y=3])   -8.2817 -0.927423  0.65890  0.73689 0.9244
loge(P[Y=5]/P[Y=4])   -0.6063 -0.215873 -0.20557 -0.03872 8.3175

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept):1        2.083477   0.108764  19.156  < 2e-16 ***
(Intercept):2        0.496005   0.048343  10.260  < 2e-16 ***
(Intercept):3        0.509668   0.043329  11.763  < 2e-16 ***
(Intercept):4       -3.818100   0.069951 -54.582  < 2e-16 ***
factor(INSTITUCION)2  0.057470   0.028294   2.031   0.0422 *
factor(INSTITUCION)3  0.723193   0.082048   8.814  < 2e-16 ***
factor(INSTITUCION)4  1.202150   0.087758  13.698  < 2e-16 ***
factor(C1P5)2         0.115115   0.029383   3.918 8.94e-05 ***
C1P14                 0.035174   0.002865  12.275  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of linear predictors:  4

Names of linear predictors:
loge(P[Y=2]/P[Y=1]), loge(P[Y=3]/P[Y=2]), loge(P[Y=4]/P[Y=3]), loge(P[Y=5]/P[Y=4])

Residual deviance: 25972.77 on 54559 degrees of freedom

Log-likelihood: -12986.38 on 54559 degrees of freedom

Number of iterations: 6
```

INTERPRETACIÓN: El odds de tener un mejor nivel de satisfacción aumenta en 12 % si el paciente no tenía cita vs si tenía cita, condicionado a las variables de institución y tiempo de demora de atención.

El odds de tener un mejor nivel de satisfacción es 3.32 veces mayor si se atiende en la institución 4 que si se atiende en la institución 1. Siendo este el mayor odds de una institución relativo a la institución 1.

c) Construya un modelo ordinal de logit acumulado para explicar que variables estan asociadas con el nivel de satisfacción. Intreprete sus resultados mas importantes.

Listing 5: Modelo ordinal de logit acumulado.

```
modelordacum = vglm(cbind(muyinsat,insat,nisatinsat,sat,muysat)~factor(
    INSTITUCION)
+ factor(C1P5)+C1P14,family = cumulative(parallel = TRUE),data=salud)
summary(modelordacum)
```

```
logit(P[Y<=1]) -0.5079 -0.07659 -0.06652 -0.05952 31.3656
logit(P[Y<=2]) -1.6890 -0.23701 -0.21323 -0.18831  8.3928
logit(P[Y<=3]) -0.9189 -0.72831 -0.64568  0.58456 10.8530
logit(P[Y<=4]) -8.1953  0.11885  0.16018  0.17747  0.7845

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept):1        -4.248457   0.112960 -37.610  < 2e-16 ***
(Intercept):2        -1.563494   0.059573 -26.245  < 2e-16 ***
(Intercept):3        -0.022201   0.056118  -0.396  0.69239
(Intercept):4         4.430585   0.080378  55.122  < 2e-16 ***
factor(INSTITUCION)2 -0.076164   0.041552  -1.833  0.06680 .
factor(INSTITUCION)3 -0.985874   0.107506  -9.170  < 2e-16 ***
factor(INSTITUCION)4 -1.534003   0.104739 -14.646  < 2e-16 ***
factor(C1P5)2        -0.136595   0.042522  -3.212  0.00132 **
C1P14                -0.048783   0.003983 -12.247  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of linear predictors:  4

Names of linear predictors:
logit(P[Y<=1]), logit(P[Y<=2]), logit(P[Y<=3]), logit(P[Y<=4])

Residual deviance: 25988.33 on 54559 degrees of freedom

Log-likelihood: -12994.16 on 54559 degrees of freedom

Number of iterations: 4

Warning: Hauck-Donner effect detected in the following estimate(s):
'(Intercept):1', '(Intercept):4'

Exponentiated coefficients:
factor(INSTITUCION)2 factor(INSTITUCION)3 factor(INSTITUCION)4
           0.9266642            0.3731131            0.2156707
       factor(C1P5)2                C1P14
           0.8723236            0.9523879
```

INTERPRETACIÓN: El odds de estar muy insatisfecho se reduce en un 13 % si la persona no tenía cita programada, condicionando la variable institución y tiempo de atención.

d) Fundamente (usando estadísticos) cual modelo (a,b ó c) describe mejor los datos.

Listing 6: Modelo ordinal de logit acumulado.

```
COMP=cbind(AIC(mod1),AIC(modelord),AIC(modelordacum))
```

```
> COMP=cbind(AIC(mod1),AIC(modelord),AIC(modelordacum))
> COMP
         [,1]     [,2]     [,3]
[1,] 25978.3 25990.77 26006.33
```

INTERPRETACIÓN: El modelo de mejor ajuste es el modelo multinomial pues presenta menor AIC. La razón puede ser por tomar alternar interpretaciones a diferencia de los otros dos modelos que generalizan una interpretación, evitando comparar uno a uno los niveles de satisfacción.

Ejercicio 2 (Distribución normal y regresión logística)

Suponga que $Y \sim \ Bernoulli(\pi)$ y que la variable de explicación, X, tiene una distribución normal. En particular, $X \sim N(0,1)$ para el grupo con $Y = 0$ y $X \sim N(\mu, 1)$ para el grupo $Y = 1$.

a) Use el **teorema de Bayes** para mostrar que la relación entre X e Y esta dada por una regresión logística. Es decir, muestre que $P(Y = 1 \mid X = x)$ satisface la regresión logística con

$$\begin{aligned} \beta_0 &= \log\left[\frac{\pi}{1-\pi}\right] - \frac{\mu^2}{2} \\ \beta_1 &= \mu \end{aligned}$$

Solución:
Tenemos que demostrar que se satisface

$$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = \beta_0 + \beta_1 x \text{ donde } \hat{\pi} = P(Y = 1 \mid X = x)$$

Tenemos las siguientes distribuciones:

$$\begin{aligned} Y &\sim Bernoulli(\pi) \\ X|Y=0 &\sim N(0,1) \\ X|Y=1 &\sim N(\mu,1) \end{aligned}$$

Por el Teorema de Bayes, denotando $f_{Y|X=x}(1) = \hat{\pi}$:

$$\begin{aligned} f_{Y|X=x}(1) &= \frac{f_{X|Y=1}(x) f_Y(1)}{\sum_{i=0}^{1} f_Y(i) f_{X|Y=i}(x)} \\ \hat{\pi} &= \frac{\left(\frac{1}{2\pi}\exp\left\{-(x-\mu)^2/2\right\}\right)\pi}{\left(\frac{1}{2\pi}\exp\left\{-(x-\mu)^2/2\right\}\right)\pi + \left(\frac{1}{2\pi}\exp\left\{-x^2/2\right\}\right)(1-\pi)} \\ \hat{\pi} &= \frac{\left(\exp\left\{-(x-\mu)^2/2\right\}\right)\pi}{\left(\exp\left\{-(x-\mu)^2/2\right\}\right)\pi + \left(\exp\left\{-x^2/2\right\}\right)(1-\pi)} \end{aligned}$$

Además,

$$1 - \hat{\pi} = \frac{\left(\exp\left\{-x^2/2\right\}\right)(1-\pi)}{\left(\exp\left\{-(x-\mu)^2/2\right\}\right)\pi + \left(\exp\left\{-x^2/2\right\}\right)(1-\pi)}$$

Luego,

$$\begin{aligned} \log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) &= \log\left(\frac{\left(\exp\left\{-(x-\mu)^2/2\right\}\right)\pi}{\left(\exp\left\{-x^2/2\right\}\right)(1-\pi)}\right) \\ &= \log\left(\frac{\pi}{1-\pi}\right) - (x-\mu)^2/2 + x^2/2 \\ &= \underbrace{\log\left(\frac{\pi}{1-\pi}\right) - \frac{\mu^2}{2}}_{\beta_0} + \underbrace{\mu}_{\beta_1} x \end{aligned}$$

b) Considere un punto de corte de 0.5. Es decir, su modelo clasifica de la siguiente forma

$$\hat{Y} = \begin{cases} 1 & \hat{\pi} > 0.5 \\ 0 & \hat{\pi} \leq 0.5 \end{cases}$$

Muestre que la sensibilidad y especificidad dadas por el modelo son

$$\begin{aligned} \phi &= 1 - \Phi\left\{\frac{1}{\beta_1}\log\left[\frac{1-\pi}{\pi}\right] - \frac{\beta_1}{2}\right\} \\ \varphi &= 1 - \Phi\left\{\frac{1}{\beta_1}\log\left[\frac{\pi}{1-\pi}\right] - \frac{\beta_1}{2}\right\} \end{aligned}$$

donde Φ es la función acumulada de distribución para la distribución normal estándar.

Solución:
El punto de corte es $\pi_0 = 0.5$.
En el item a) deducimos la siguiente relación:

$$\beta_0 = \log\left[\frac{\pi}{1-\pi}\right] - \frac{\beta_1^2}{2}.$$

Determinemos x_0 que se establece en el punto de corte, usando la relación del item a)

$$\begin{aligned} \log\left(\frac{\pi_0}{1-\pi_0}\right) &= \beta_0 + \beta_1 x_0 \\ \log\left(\frac{0.5}{1-0.5}\right) &= \log\left[\frac{\pi}{1-\pi}\right] - \frac{\beta_1^2}{2} + \beta_1 x_0 \\ 0 &= \log\left[\frac{\pi}{1-\pi}\right] - \frac{\beta_1^2}{2} + \beta_1 x_0 \quad \Rightarrow x_0 = \frac{1}{\beta_1}\log\left[\frac{1-\pi}{\pi}\right] + \frac{\beta_1}{2} \end{aligned}$$

Calculemos primero la sensibilidad (ϕ), la cual es definida como:

$$\begin{aligned} \phi(0.5) &= P_{\pi_0}(\hat{Y} = 1 \mid Y = 1) : \text{Área de Positivos Verdaderos} \\ &= P_{X|Y=1}(x_0 < x) \\ &= P_{Z|Y=1}(x_0 - \mu < z) : (\text{Estandarizando } Z \mid Y = 1 \sim N(0,1)) \\ &\overset{\mu=\beta_1}{=} P_{Z|Y=1}(x_0 - \beta_1 < z) \\ &= 1 - P_{Z|Y=1}(z \leq x_0 - \beta_1) \\ &= 1 - \Phi\{x_0 - \beta_1\} \\ &= 1 - \Phi\left\{\frac{1}{\beta_1}\log\left[\frac{1-\pi}{\pi}\right] + \frac{\beta_1}{2} - \beta_1\right\} \\ &= 1 - \Phi\left\{\frac{1}{\beta_1}\log\left[\frac{1-\pi}{\pi}\right] - \frac{\beta_1}{2}\right\} \end{aligned}$$

Ahora, calculemos la especificidad (ϕ), la cual es definida como:

$$\begin{aligned} \varphi(0.5) &= P_{\pi_0}(\hat{Y} = 0 \mid Y = 0) : \text{Área de Negativos Verdaderos} \\ &= P_{X|Y=0}(x \leq x_0) : \text{Sabemos } X|Y = 0 \sim N(0,1) \\ &\overset{\text{Simetría}}{=} P_{X|Y=0}(-x_0 \leq x) \\ &= 1 - P_{X|Y=0}(x \leq -x_0) \\ &= 1 - \Phi\{-x_0\} \\ &= 1 - \Phi\left\{-\left(\frac{1}{\beta_1}\log\left[\frac{1-\pi}{\pi}\right] + \frac{\beta_1}{2}\right)\right\} \\ &= 1 - \Phi\left\{\frac{1}{\beta_1}\log\left[\frac{\pi}{1-\pi}\right] - \frac{\beta_1}{2}\right\} \end{aligned}$$

Ejercicio 3 (Simulaciones)

El siguiente ejercicio pretende evaluar las diferentes técnicas de construcción de modelos. Considere tres variables aleatorias $X_1 \sim N(10,1)$, $X_2 \sim B(0.5)$, $X_3 \sim Beta(\alpha = 2, \beta = 3)$, $X_j \sim N(10,4)$ para $j = 4, \ldots, 20$. Asuma que $Y \sim B(\pi)$ es una variable binaria y que relación entre Y y $(X_1, \ldots, X_20)$ está dada por un modelo de regresión logística de la forma

$$logit = -6 + 0.5X_1 - 0.5X_2 + 0.25X_3 + 0X_4 + \ldots + 0X_{20}.$$

a) Genere 1000 bases de datos (de tamaño $n = 2000$) para las variables $(X_1, \ldots, X_{20}; Y)$.

Listing 7: Generando 1000 bases de datos de tamaño $n = 2000$.

```
set.seed(12414)

coef = c(-6,0.5,-0.5,0.25,rep(0,20-3))
pi = function(x)exp(x)/(1+exp(x))

Generador.Base= function(n){
# Variables explicativas
X1 = rnorm(n,10,1)
X2 = rbinom(n,1,0.5)
X3 = rbeta(n,2,3)
Xj=cbind(X1,X2,X3)
for(j in 4:20){
        Xj = cbind(Xj,rnorm(n,10,2))
}

X=matrix(cbind(1,Xj),nrow=n)

#Variable respuesta
Y = rbinom(n,1,prob = pi(X%*%coef))

#Base de datos
Base.de.datos = data.frame(X[,-1])
Base.de.datos=cbind(Y,Base.de.datos)
}

#Generando 1000 bases de datos de tamano 2000
Base.conjunta = lapply(rep(2000,1000),Generador.Base)
```

```
> head(Base.conjunta[[1]],3)
  Y       X1 X2        X3        X4        X5        X6        X7        X8
1 1 9.627269  1 0.5679022  9.425309 10.579598 10.915309  9.701528 10.153843
2 0 9.882158  1 0.7162675  9.548787  7.709658  8.880411 12.523346  7.652174
3 0 9.024324  0 0.3056396 13.496539  8.820128  8.429861  8.375715 10.550201
         X9       X10      X11       X12      X13       X14       X15       X16
1 13.419227 10.169630 15.74950 11.466127 5.647242 10.803244  9.843539  8.792767
2 10.369788  8.446543 10.78719  8.090627 7.244886  4.160571 10.625332 11.246899
3  9.909092  8.025900  9.06679 10.535992 9.781346 10.779682 10.486401 10.711255
        X17       X18      X19       X20
1  7.143451  8.706324  9.53062  9.498608
2 13.472588  8.447055 12.02767 11.237917
3 11.148585 12.161867  8.30441  6.469498
```

b) Aplique el método de regresión de lasso y calcule la proporción de veces cada una de las variables X_i es seleccionada.

Listing 8: Regresión de Lasso - Proporción de selección

```
M = matrix(nrow=1000,ncol=21)

for(j in 1:1000){
        cv1 <- cv.glmnet(as.matrix(Base.conjunta[[j]][,-1]),Base.conjunta[[j
            ]][,1],alpha = 1,family = "binomial",
        type.measure = "auc")
        M[j,] <- as.numeric(coef(cv1,s="lambda.min"))
}

# Calculando la proporcion de seleccion de cada variable Xi
Pr = apply(M != 0,2,sum)/1000
```

```
>     # Proporcion
> Pr = apply(M != 0,2,sum)/1000
> Pr
 [1] 1.000 1.000 0.990 0.246 0.145 0.132 0.142 0.132 0.144 0.136 0.130 0.129
[13] 0.146 0.119 0.129 0.130 0.150 0.136 0.123 0.140 0.136
```

c) Aplique el método de regresión de ridge y calcule la proporción de veces cada una de las variables X_i es seleccionada.

Listing 9: Regresión de Ridge - Proporción de selección

```
M2 <- matrix(nrow=1000,ncol=21)

for(j in 1:1000){
        cv2 <- cv.glmnet(x = as.matrix(Base.conjunta[[j]][,-1]), Base.conjunta[[j
            ]][,1]
                ,alpha = 0, family = "binomial", type.measure = "auc")

        M2[j,] <- as.numeric(coef(cv2,s="lambda.min"))
}

PR2 = apply(M2 != 0,2,sum)/1000
```

```
[1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

d) Estime el sesgo para cada coeficiente del modelo de regresión.

$$\frac{1}{1000}\sum_{i=1}^{1000}\left(\hat{\beta}_k^{(i)} - \beta_k\right), \quad k = 0, \ldots, 20$$

Listing 10: Sesgo

```
round(apply(M2 - matrix(rep(coef,1000),ncol=21,byrow = TRUE),2,sum)/1000,4)
round(apply(M - matrix(rep(coef,1000),ncol=21,byrow = TRUE),2,sum)/1000,4)
```

e) Escriba un párrafo sobre las implicaciones prácticas de este ejercicio.

Como se puede observar, la regresión Ridge mantiene proporciones de 1 en la selección de

todas las variables, es decir que siempre las selecciona a diferencia de la regresión de Lasso, en la que algunas proporciones son casi todas menores que 0.2. La importancia de este ejercicio es que puede observarse la exigencia que Ridge toma en las variables, y como Lasso logra distinguir mejores modelos haciendo uso de menos variable