Probabilidad Condicional y Regresion

Luis Ellis

January 6, 2019

Analisis Estadistico

Antes de iniciar con todo el analisis, pasamos a definir todo nuestro environment:

```
#Environment for this task
library(tidyverse)
library(olsrr)
library(vtreat)
library(dummies)
library(stargazer)
#Workspace
diabetes = read_csv("diabetes.csv")
```

Probabilidad Condicional

La primera parte del proyecto consta de realizar 3 Probabilidades condicionales, para esto tomamos los tres casos que pueden haber:

- Por tipos de variables:
  - Dos continuas
  - Dos Categoricas
  - Una continua y una categorica.

Comenzaremos de arriba para abajo, escogiendo la edad y el genero. Para escoger un grupo de categorias, encontramos el promedio de la edad 46.8511166. Con este valor entonces procedemos a encontrar la probabilidad condicional.

```
diabetes %>%
  mutate(age_cat = ifelse(age >= 45, "at least 45", "less than 45")) %>%
  count(age_cat, gender) %>%
  spread(gender, n) %>%
  mutate(prop = female / (female + male))
```

```
## # A tibble: 2 x 4
##   age_cat      female  male  prop
##   <chr>         <int> <int> <dbl>
## 1 at least 45     108    95 0.532
## 2 less than 45    126    74 0.63
```

En este resultado vemos que tenemos a la derecha una columna que se llama "prop" si vemos como fue calculada, es de forma:

$$P(\text{Mujer}|\text{Mayor a 45}) = \frac{P(\text{Mujer} \cap \text{Mayor a 45})}{\text{Mayor a 45}},$$

Entonces los renglones seran B, o el evento dado, y las columnas seran A o el evento de interes para los calculos siguientes. Para este entonces, la probabilidad dado que la persona sea mayor a 45, que sea mujer es de 53.

Los dos casos entonces aqui:

```
#We then do it again with two continuous variables
diabetes%>%
  filter(
    !is.na(chol),
    !is.na(hdl)
  ) %>%
    mutate(chol_cat = ifelse(chol >=240, "High cholesterol", "Borderline or Good cholesterol")) %>%
    mutate(hdl_cat = ifelse(hdl >= 50, "High Density Lipoprotein over 50", "Under 50")) %>%
    count(chol_cat, hdl_cat) %>%
    spread(hdl_cat, n) %>%
    mutate(prop =
             `High Density Lipoprotein over 50` / (`High Density Lipoprotein over 50` + `Under 50`))
```

```
## # A tibble: 2 x 4
##   chol_cat                   `High Density Lipoprotein ov~ `Under 50`  prop
##   <chr>                                              <int>      <int> <dbl>
## 1 Borderline or Good choles~                           134        190 0.414
## 2 High cholesterol                                      37         41 0.474
```

```
#And lastly, for 2 categorical variables:
diabetes %>%
    count(location, gender) %>%
    spread(gender,n) %>%
    mutate(prop = female / (female + male))
```

```
## # A tibble: 2 x 4
##   location   female  male  prop
##   <chr>       <int> <int> <dbl>
## 1 Buckingham    114    86 0.570
## 2 Louisa        120    83 0.591
```

Regresion

El analisis de regresion es un tipo de analisis que usamos para estimar, o poder predecir una variable objetivo a partir de otras, como el futuro es incierto no nos queda mas que intentar predecir y luego verificar nuestras hipotesis de prediccion.

En este analisis de regresion usamos el criterio de informacion de Akaike, que hacemos un intercambio entre bondad de ajuste y complejidad. El metodo fue escogido debido a que el dataset es un tema de salud y me parecio interesante ver como reducir la complejidad de los componentes de salud que simplemente usar el p-value.

Para este dataset en particular se tuvo que eliminar dos variables que tenian muchos valores missing ya que no nos eran particularmente utiles. Luego de esto usamos un step backwaird eliminando variables con Akaike, y luego vimos el modelo final. No necesariamente es el mejor modelo, pero para nuestra data, se ha ajustado de buena forma.

```r
#Regression analysis
#Cleaning
diabetes_cat <- dummy.data.frame(as.data.frame(diabetes),
                                 names = c("location","gender","frame"), sep="_")
rownames(diabetes_cat) = diabetes_cat[,1]
diabetes_cat <- diabetes_cat[,- c(1,20,21)] #Drop IDs, Second Diastolic and Syastolic are dropped due N
#Model building
model <- lm(chol ~ ., data = diabetes_cat)
k <- ols_step_backward_aic(model)
```

```
## Backward Elimination Method
## ---------------------------
##
## Candidate Terms:
##
## 1 . stab.glu
## 2 . hdl
## 3 . ratio
## 4 . glyhb
## 5 . location_Buckingham
## 6 . location_Louisa
## 7 . age
## 8 . gender_female
## 9 . gender_male
## 10 . height
## 11 . weight
## 12 . frame_large
## 13 . frame_medium
## 14 . frame_small
## 15 . frame_NA
## 16 . bp.1s
## 17 . bp.1d
## 18 . waist
## 19 . hip
## 20 . time.ppn
##
##
## Variables Removed:
##
## - location_Buckingham
## - frame_NA
## - bp.1s
## - waist
## - age
## - stab.glu
##
## No more variables to be removed.
```

```r
k
```

```
##
##
##                              Backward Elimination Summary
```

```
## ------------------------------------------------------------------------------
## Variable                  AIC          RSS          Sum Sq       R-Sq      Adj. R-Sq
## ------------------------------------------------------------------------------
## Full Model             3445.558    190969.752    556341.981    0.74446      0.73229
## location_Buckingham    3443.558    190969.752    556341.981    0.74446      0.73229
## gender_female          3441.558    191878.917    555432.817    0.74324      0.73399
## gender_male            3441.558    191883.280    555428.453    0.74324      0.73472
## frame_large            3441.558    191916.396    555395.337    0.74319      0.73541
## frame_medium           3441.558    192721.394    554590.339    0.74211      0.73503
## frame_small            3441.558    193418.547    553893.186    0.74118      0.73480
## frame_NA               3441.558            NA            NA         NA           NA
## bp.1s                  3431.348            NA            NA         NA           NA
## waist                  3429.412            NA            NA         NA           NA
## age                    3428.982            NA            NA         NA           NA
## stab.glu               3428.336            NA            NA         NA           NA
## ------------------------------------------------------------------------------
```

```
formula <- as.formula("chol ~ hdl + ratio + glyhb  + location_Louisa + gender_female +
                      gender_male + height + weight + frame_large + frame_medium +
                      frame_small + bp.1d + hip + time.ppn ")
diabetes_model <- lm(formula, data = diabetes_cat)
summary(diabetes_model)
```

```
##
## Call:
## lm(formula = formula, data = diabetes_cat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -136.003  -10.061   -0.295   11.322   75.946
##
## Coefficients: (1 not defined because of singularities)
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -3.593e+01  3.758e+01  -0.956   0.3396
## hdl              2.453e+00  9.812e-02  25.004   <2e-16 ***
## ratio            2.826e+01  1.009e+00  28.007   <2e-16 ***
## glyhb            5.005e-01  5.865e-01   0.853   0.3940
## location_Louisa  5.894e+00  2.511e+00   2.347   0.0195 *
## gender_female   -1.975e+00  3.910e+00  -0.505   0.6138
## gender_male             NA         NA      NA       NA
## height          -6.928e-01  4.749e-01  -1.459   0.1455
## weight          -5.965e-02  7.319e-02  -0.815   0.4156
## frame_large      3.034e+00  8.293e+00   0.366   0.7147
## frame_medium     2.648e+00  8.078e+00   0.328   0.7433
## frame_small     -1.241e+00  8.300e+00  -0.149   0.8813
## bp.1d            1.327e-01  9.208e-02   1.441   0.1504
## hip              7.120e-01  5.199e-01   1.370   0.1717
## time.ppn         4.078e-04  3.977e-03   0.103   0.9184
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.07 on 361 degrees of freedom
##   (28 observations deleted due to missingness)
## Multiple R-squared:  0.7428, Adjusted R-squared:  0.7336
```

```
## F-statistic: 80.21 on 13 and 361 DF,  p-value: < 2.2e-16
```

```
stargazer(diabetes_model, type = "latex", title = "Modelo")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Tue, Apr 30, 2019 - 3:14:29 PM

Terminamos con el modelo con un R^2adj $= 0.73$ el cual nos arroja un buen resultado para el modelo en general. Para este proyecto no se necesitó cross-validation así que no llevamos a acabo mas experimentación con la muestra y el modelo.

Table 1: Modelo

	Dependent variable:
	chol
hdl	2.453***
	(0.098)
ratio	28.256***
	(1.009)
glyhb	0.500
	(0.586)
location_Louisa	5.894**
	(2.511)
gender_female	−1.975
	(3.910)
gender_male	
height	−0.693
	(0.475)
weight	−0.060
	(0.073)
frame_large	3.034
	(8.293)
frame_medium	2.648
	(8.078)
frame_small	−1.241
	(8.300)
bp.1d	0.133
	(0.092)
hip	0.712
	(0.520)
time.ppn	0.0004
	(0.004)
Constant	−35.933
	(37.576)
Observations	375
R^2	0.743
Adjusted R^2	0.734
Residual Std. Error	23.073 (df = 361)
F Statistic	80.215*** (df = 13; 361)
Note:	*p<0.1; **p<0.05; ***p<0.01