UNIVERSIDAD DE EL SALVADOR
FACULTAD MULDISCIPLINARIA DE OCCIDENTE

DEPARTAMENTO DE MATEMÁTICA

LICENCIATURA EN ESTADÍSTICA

PROYECTOS DE ESTUDIOS ESTADÍSTICOS

Predicción de la clase social:
Una aplicación de la Regresión Logística Multinomial

DOCENTE:

Lic. Jaime Isaac Peña

PRESENTADO POR:

Hugo Fernando López Cortés

Tabla de contenidos

1. Regresión Logística

La regresión logística (RL) forma parte del conjunto de métodos estadísticos que caen bajo tal denominación y es la variante que corresponde al caso en que se valora la contribución de diferentes factores en la ocurrencia de un evento simple. a RL es una de las técnicas estadístico-inferenciales más empleadas en la producción científica contemporánea. Surge en la década del 60, su generalización dependía de la solución que se diera al problema de la estimación de los coeficientes. El algoritmo de Walker-Duncan para la obtención de los estimadores de máxima verosimilitud vino a solucionar en parte este problema, pero era de naturaleza tal que el uso de computadoras era imprescindible.

La regresión logística es una regresión múltiple cuando la variable dependiente es no métrica. Si dicha variable únicamente posee dos niveles se conoce como regresión logística binomial, y si posee más se denomina regresión logística multinomial. Por ejemplo, esta técnica estadística permite estudiar problemas típicos como la explicación o predicción de la quiebra de empresas, o la decisión del consumidor de recomprar o no un producto o servicio. En este sentido, dada la naturaleza de los objetivos de la aplicación, resulta pertinente aclarar la importancia de no abordar estos problemas mediante una regresión lineal múltiple. La razón es que cuando la variable dependiente es no métrica, por ejemplo, dicotómica, es imposible que cumpla las exigencias de una regresión múltiple: no puede seguir una distribución normal ni tener varianza constante. Tampoco podría cumplirse la hipótesis de linealidad, entonces, la solución pasa por linealizar de alguna forma lo que es una relación no lineal, que en lo que se basa el modelo de regresión logística.

1.1. Modelo de regresión logística binomial

Suponiendo que se desea analizar la relación de una variable dependiente limitada dicotómica Y que toma los valores 0 y 1 en función de una variable métrica X. La relación entre la variable X y Y en un modelo de regresión lineal se plantearía del siguiente modo:

$$Y_i = \beta_0 + \beta_1 X_1 i + \varepsilon_i$$

Donde β_0 sería el intercepto al eje de coordenadas y β_1 la pendiente de esa recta. De forma general, si se tuvieran n variables explicativas, entonces la expresión anterior quedaría:

$$Y_i = \beta_0 + \beta_1 X_1 i + \beta_2 X_2 i + \cdots + \beta_n X_n i + \varepsilon_i$$

A diferencia de la regresión lineal, la regresión logística no predice el valor Y_i dado los valores de X_i, sino que directamente estima la probabilidad de que ocurra $Y_i (Y_i = 1)$ dados los valores de X_i. Por supuesto que incorpora las expresiones anteriores para tener en cuenta la relación entre la variable dependiente y las independientes, pero las envuelve en la siguiente función para calcular la probabilidad de ocurrencia en lugar de predecir Y_i:

$$Pr(Y) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1 i+)}} = \frac{1}{1+e^{-Y}}$$

Y para el caso de n variables:

$$Pr(Y) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1 i+\beta_2 X_2 i+\cdots+\beta_n X_n i)}} = \frac{1}{1+e^{-Y}}$$

Donde $Pr(Y)$ es la probabilidad de ocurrencia de Y, y e es la base del logaritmo natural. Es la forma de linealizar la relación que no era lineal e impedía la estimación de una regresión lineal.

1.1.1. Estimación del modelo

El modelo se estima mediante la minimización de la función de máxima-verosimilitud, que es un planteamiento análogo al evaluar cuanta información queda por explicar después que el modelo se ha estimado:

$$LL = \sum_{i=1}^{N} \left[Y_i \ln\left(\Pr\left(Y_i\right)\right) + (1 - Y_i) \ln\left(1 - \Pr\left(Y_i\right)\right)\right]$$

Suponiendo que, para un valor real de $Y = 0$, la función estimada ha pronosticado una probabilidad de ocurrencia cercana a 0 entonces, la función logarítmica es 0 para $X = 1$. Dicho de otra manera, la función de máxima verisimilitud toma valores cercanos a 0 cuando la probabilidad predicha (cercana a 0 en caso de no ocurrencia, cercana a 1 en caso de ocurrencia) acierta a clasificar al caso como 0 a 1. Sin embargo, en el caso de desacierto, por ejemplo, con un $Y = 0$, se tiene que la función de máxima verisimilitud crece haciéndose grande, puesto que $ln(x)$ tiende a $-\infty$ cuando x tiende a 0. En conclusión, mayor valor implicará menos capacidad de la estimación para replicar los valores reales.

1.1.2. Contraste de hipótesis para el modelo estimado

Contraste de significatividad global: El primer paso es contrastar la hipótesis nula de que todos los coeficientes de regresión son nulos, dado que, de ser así, no tendría sentido continuar con la interpretación del modelo. Los pasos a seguir son: calcular la máxima verosimilitud LL de un modelo en el que la función solo esta formada por el intercepto β_0, es decir un modelo en el que las variables explicativas no jugarían ningún papel o, dicho de otro modo, en que todos los β son nulos $LL(0)$. Estimamos la función de máxima verosimilitud del modelo $LL(M)$, si este es significativamente más pequeño que el primero, se concluye que es más plausible (verosímil) por lo que alguna variable debe estar ejerciendo una influencia significativa en la predicción de la variable dependiente.

Contraste para los coeficientes individuales: Una vez descartada la hipótesis de que todos los coeficientes son nulos, es necesario saber cuál es la contribución individual de los regresores a la explicación de la variable dependiente. El planteamiento en una regresión logística, al igual que en la regresión lineal, se construye un estadístico denominado test de Wald.

1.1.3. Interpretación de los coeficientes de regresión

El papel de los coeficientes estandarizados en la regresión logística la juegan los denominados odds ratios. Se define odd de un acontecimiento como la razón entre su probabilidad de ocurrencia y la de no ocurrencia:

$$odd = \frac{\Pr(Y=1)}{\Pr(Y=0)}$$

Siguiendo el caso de la probabilidad de ocurrencia de Pr(Y) entonces el odd puede escribirse como:

$$odd = \frac{\Pr(Y=1)}{\Pr(Y=0)} = \frac{\frac{1}{1+e^{-Y}}}{1-\frac{1}{1+e^{-Y}}} = e^Y$$

Pero e^Y puede expresarse como:

$$e^Y = e^{\beta_0+\beta_1 X_{1i}+\ \beta_2 X_{2i}+\ldots+\beta_n X_{ni}} = e^{\beta_0} e^{\beta_1 X_{1i}+\ \beta_2 X_{2i}+\ldots+\beta_n X_{ni}}$$

Al termino e^{β_i} se le conoce como odd ratio y su interpretación es la siguiente: es el factor en que se incrementa el odd cuando la variable independiente i-ésima se incrementa en una unidad (y el resto permanecen constantes). Mayores odd ratios pueden interpretarse como mayor influencia relativa de esa variable en la predicción de la ocurrencia del caso.

1.1.4. Ajuste del modelo

- R^2 de McFadden

$$R^2{}_{MF} = \frac{-2LL(0)-(-2LL(M))}{-2LL(0)}$$

Es la proporción de la reducción que supone el modelo con respecto al modelo base. Varia entre 0 y 1.

- R^2 de Cox y Snell

$$R^2{}_{CS} = 1 - e^{\left[\frac{1}{N}(2LL(M))-2LL(0)\right]}$$

Donde N es el tamaño muestral. Esta medida, por su construcción, nunca puede alcanzar el 1, por esta razón Nagelkerke propone:

- R^2 de Nagelkerke

$$R^2{}_N = \frac{R^2{}_{CS}}{1-e^{\left[\frac{2LL(0)}{N}\right]}}$$

- Matriz de confusión

Consiste en comparar los valores reales de la variable dependiente con los valores predichos. Dado que se conoce su pertenencia real, basta con generar una tabla cruzada de valores reales y valores predichos, cuantos más elementos haya en la diagonal de esa tabla, mejor será la precisión del modelo.

1.1.5. Capacidad discriminante del modelo

La capacidad discriminante del modelo es la capacidad del mismo para distinguir entre los los grupos en función de la probabilidad predicha. En cierta forma, la matriz de confusión sirve como indicador, sobre todo su se completa con los porcentajes de sensibilidad y especifidad. Sin embargo, existen indicadores adicionales que pueden utilizarse, fundamentalmente el estadístico c asociado a las curvas ROC (Receiver Operating Characteristic).

1.2. Regresión logística multinomial

La regresión logística multinomial es utilizada en modelos con variable dependiente de tipo nominal con más de dos categorías (politómica). Los modelos de elección discreta con más de dos alternativas se denominan modelos multinomiales. Considerando que el número de alternativas son $J+1$ $(0, 1, 2, \dots J)$, tomándose a la alternativa 0 como categoría de referencia. Para construir el modelo logit multinomial se requiere J vectores de parámetrso, entonces se define:

$$Z_{ij} = \beta_{1j} + \ \beta_{2j} X_{2i} + \dots + \beta_{kj} X_{ki}$$

Las probabilidades de cada alternativa se expresan de la siguiente forma:

$$P_{ij} = Pr\left(Y_i = j\right) = \frac{e^{-\left(\beta_{1j} + \ \beta_{2j} X_{2i} + \dots + \beta_{kj} X_{ki}\right)}}{1 + \sum_{g=1}^{J} e^{-\left(\beta_{1j} + \ \beta_{2j} X_{2i} + \dots + \beta_{kj} X_{ki}\right)}}$$

$$P_{i0} = Pr\left(Y_i = 0\right) = \frac{1}{1 + \sum_{g=1}^{J} e^{-\left(\beta_{1j} + \ \beta_{2j} X_{2i} + \dots + \beta_{kj} X_{ki}\right)}}$$

Cuando J es igual a 1, el modelo multinomial es igual al dicotómico. En el modelo anterior el logaritmo neperiano de los odds ratio entre la alternativa j y la alternativa de la categoría de referencia (0) viene dada por:

$$\ln\left[\frac{P_{ij}}{P_{ig}}\right] = \beta_{1j} - \beta_{1g} + \left(\beta_{2j} - \beta_{2g}\right) \ X_{2i} + \dots + \left(\beta_{kj} - \beta_{kg}\right) \ X_{ki}$$

En los modelos multinomiales los parámetros deben interpretarse con mucho cuidado. En principio, se podría pensar que el signo del efecto marginal de una variable sobre la probabilidad de elegir una alternativa solo depende del correspondiente elemento del vector β_j. Sin embargo, al derivar la expresión, se puede establecer que el efecto marginal de la variable X_h es igual a:

$$\frac{\partial P_{ij}}{\partial X_h} = P_{ij}\left[\beta_{hj} - \bar{\beta}_h\right]$$

En donde, $\bar{\beta}_h$ es la media de los parámetros β_{hj} para las J alternativas. Este efecto marginal donde P__j que siempre es positiva y la diferencia entre $\bar{\beta}_h$ y β_{hj}. Por tanto, el signo dependerá del coeficiente y alternativa que se esté analizando.

A continuación, se presenta un caso de aplicación de la regresión logística multinomial a la predicción de la clase social de un individuo tomando como referencia sus calificaciones a lo largo de su paso por el sistema educativo.

2. Predicción de la clase social

Se desea saber en qué medida el desempeño de un individuo a lo largo de su paso por el sistema educativo es capaz de explicar su "éxito" final en la vida medido este como la clase social a la que pertenece tras unos años de acabar sus estudios. También es interesante conocer si el hecho de haber sido mejor estudiante en asignaturas relacionadas con las ciencias sociales o con las ciencias aplicadas influye en este resultado. La base de datos posee 200 registros y para realizar este análisis, se han seleccionado las siguientes variables de estudio:

- *ses*, es la clase social y esta codificada por 1 (Baja), 2 (Media) y 3 (Alta). Es la variable dependiente.
- *female*, corresponde al sexo y esta codificado por 1 (Mujer) y 2 (Hombre).
- *science*, es el puntaje estandarizado que mide el desempeño en la asignatura de ciencias.
- *socst*, es el puntaje estandarizado que mide el desempeño en la asignatura de ciencias sociales.

Para llevar a cabo la aplicación de la regresión logística para predecir la clase social se propone la realización de los cálculos mediante tres softwares: R Statistical Computing, SPSS Statistics y el lenguaje de programación Python. No sin antes realizar un análisis exploratorio para condicionar los datos, como se muestra a continuación.

2.1 Análisis exploratorio

Importando las librerias

```
library(haven)
library(gmodels)
library(nnet)
library(stargazer)
```

```
Please cite as:

 Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Table

 R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

```
library(ggplot2)
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```r
library(cowplot)
library(patchwork)
```

```
Attaching package: 'patchwork'

The following object is masked from 'package:cowplot':

    align_plots
```

```r
library(plotrix)
library(naniar)
library(visdat)
library(outliers)
```

Importando la base de datos, mostrando las dimensiones y los nombres de las variables

```r
datos  <-  read_sav("hsbdemo.sav")
dim(datos)
```

```
[1] 200  13
```

```r
names(datos)
```

```
 [1] "id"      "female"  "ses"     "schtyp"  "prog"    "read"    "write"
 [8] "math"    "science" "socst"   "honors"  "awards"  "cid"
```

Empezaremos análizando las variables numéricas que se utilizaran en la aplicación de RLM, para ello se muestran algunos estadísticos

```r
summary(datos$science)
```

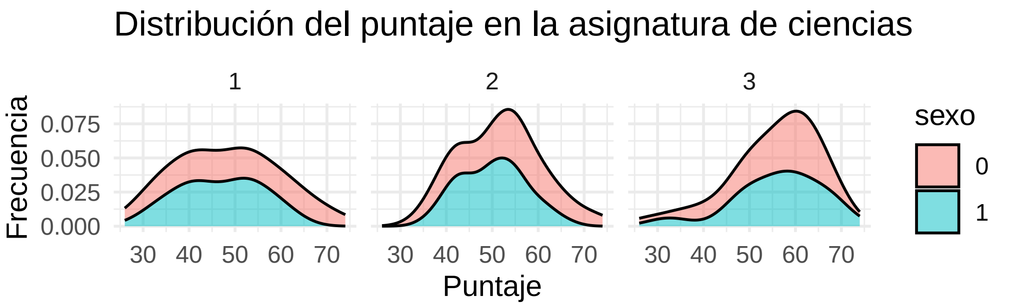
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  26.00   44.00   53.00   51.85   58.00   74.00
```

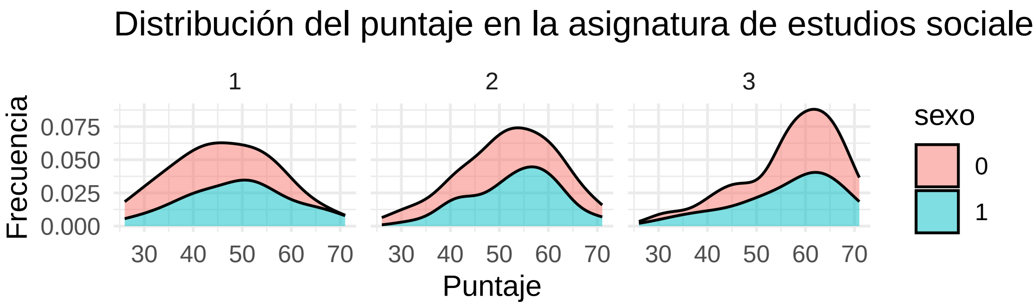
```
summary(datos$socst)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  26.00   46.00   52.00   52.41   61.00   71.00 
```

Distribuciones de las variables del puntaje en las asignaturas de ciencias y sociales por sexo y clase social

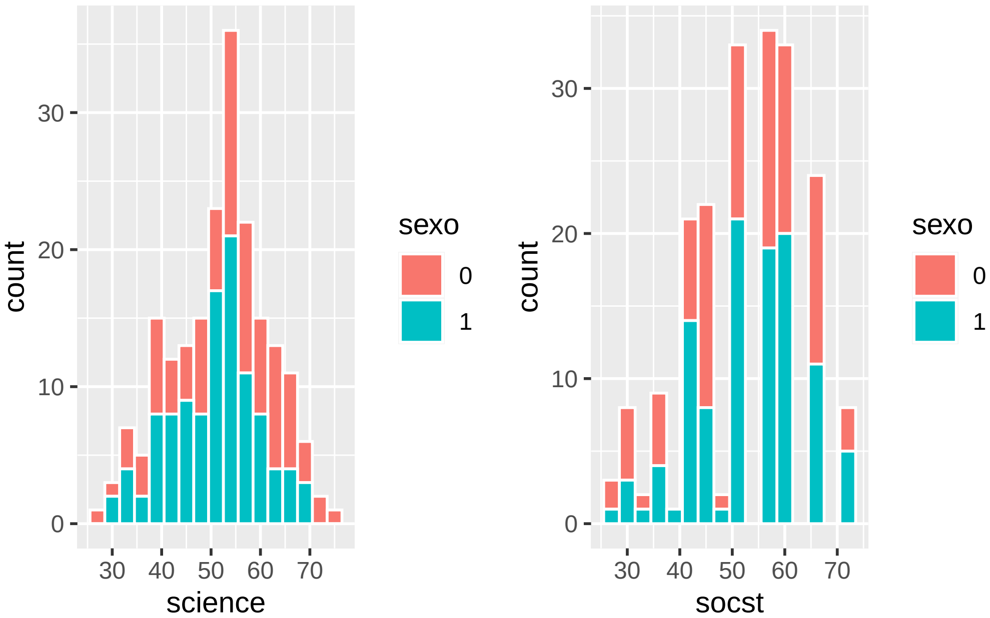
```
#Preparando variables para graficar
sexo = as.factor(datos$female)
cs = as.factor(datos$ses)

pltscience = ggplot(data = datos) +
  geom_density(mapping = aes(x = science, fill =sexo), position = 'stack', alpha = 0.5
  xlab("Puntaje") +
  ylab("Frecuencia") +
  ggtitle("Distribución del puntaje en la asignatura de ciencias") +
  theme_minimal() +
  facet_wrap(~ses)

pltsocst =ggplot(data = datos) +
  geom_density(mapping = aes(x = socst, fill =sexo), position = 'stack', alpha = 0.5)
  xlab("Puntaje") +
  ylab("Frecuencia") +
  ggtitle("Distribución del puntaje en la asignatura de estudios sociales") +
  theme_minimal() +
  facet_wrap(~ses)

pltscience / pltsocst
```





Histogramas de las variables del puntaje en las asignaturas de ciencias y sociales por sexo

```
g1 = ggplot(data = datos, aes(x = science, fill = sexo)) +
  geom_histogram( binwidth = 3, color = "white")
g2 = ggplot(data = datos, aes(x = socst, fill = sexo)) +
  geom_histogram( binwidth = 3, color = "white")
g1+g2
```



Ahora se estudiaran las variables categóricas female (sexo) y clases social (ses).

```
str(datos$ses)
```

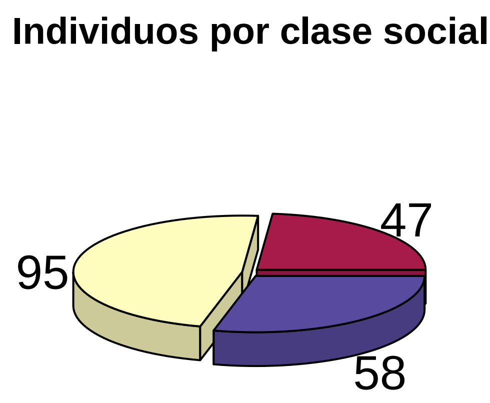
```
dbl+lbl [1:200] 1, 2, 3, 1, 2, 3, 2, 2, 2, 2, 1, 3, 1, 2, 1, 1, 1, 2, 2, 3...
@ format.spss  : chr "F1.0"
@ display_width: int 5
@ labels       : Named num [1:3] 1 2 3
 ..- attr(*, "names")= chr [1:3] "low" "middle" "high"
```

```
table(datos$ses)
```

```
 1  2  3
47 95 58
```

```
tses = table(datos$ses)
pie3D(tses, labels=tses,
      explode=0.05, col = hcl.colors(length(tses), "Spectral"),
```

```
      main="Individuos por clase social")
```




```
str(datos$female)
```

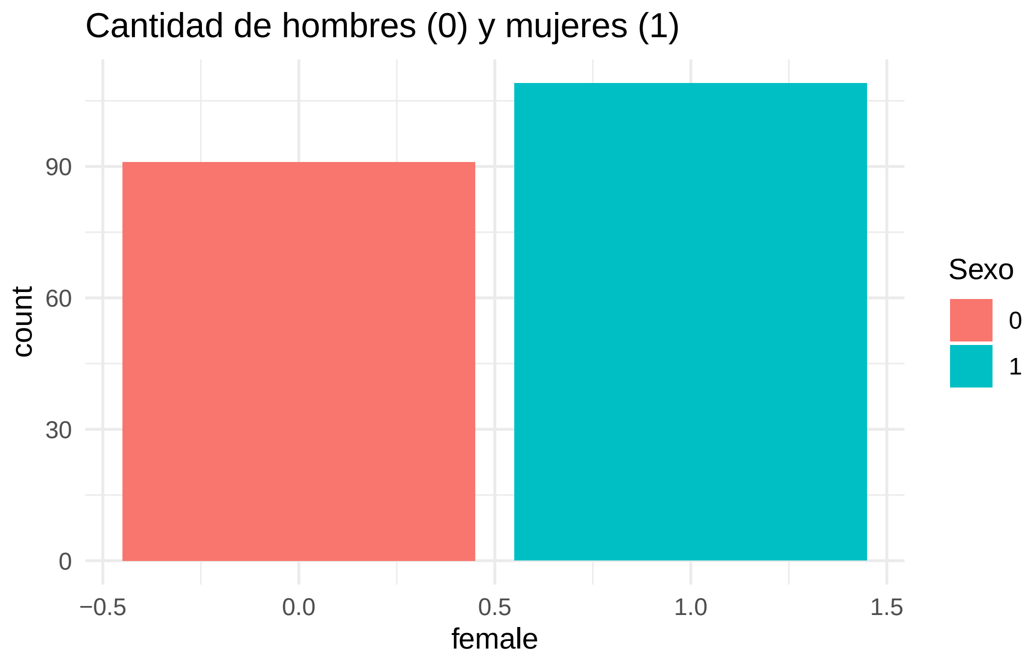
```
dbl+lbl [1:200] 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 1, 1, 0, 1, 1, 0, 0...
@ format.spss: chr "F1.0"
@ labels     : Named num [1:2] 0 1
 ..- attr(*, "names")= chr [1:2] "male" "female"
```

```
table(datos$female)
```

```
 0   1
91 109
```

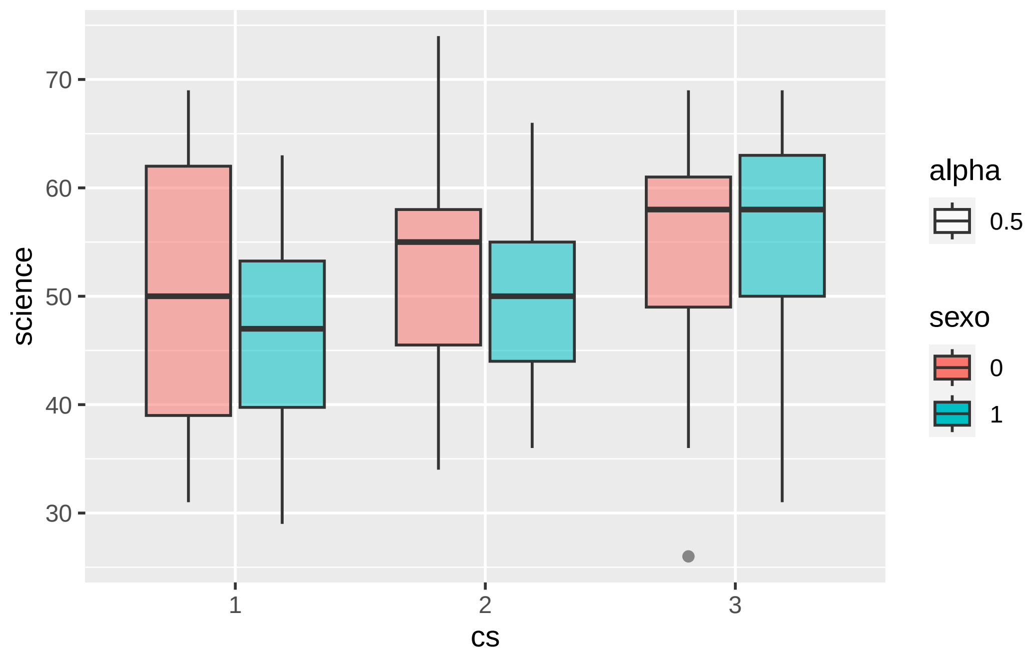
```
ggplot(data = datos, aes(x = female, fill = sexo)) +
  geom_bar() +
  ggtitle("Cantidad de hombres (0) y mujeres (1)") +
  labs(fill = "Sexo") +
```

```
  theme_minimal()
```

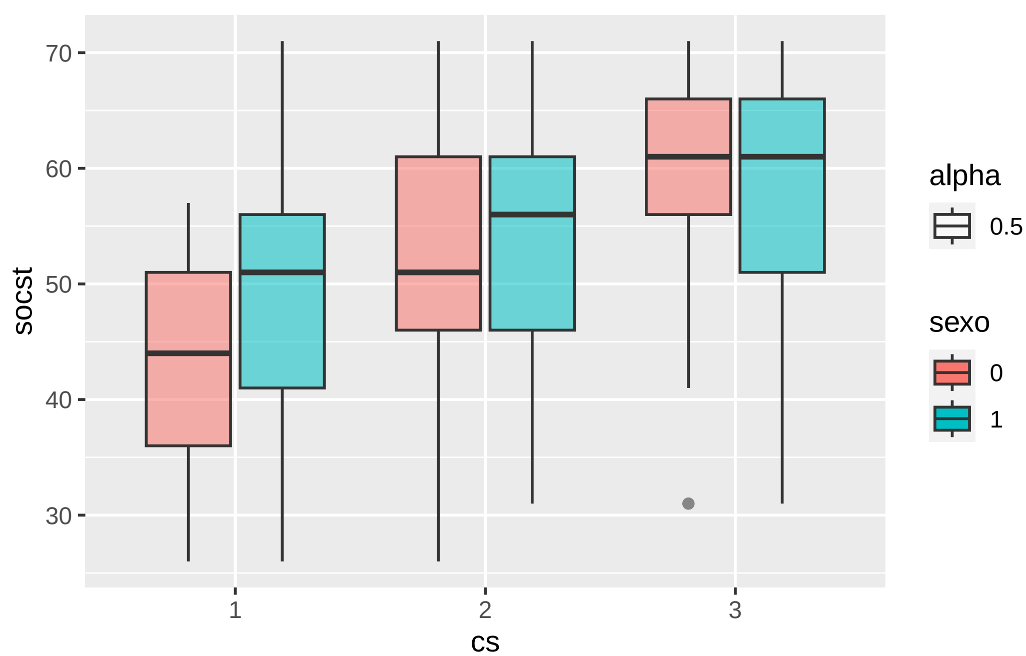


Es pertinente realizar un análisis de datos atípicos, para ello se estudiaran mediante las distancias de Mahalanobis y mediante los graficos de cajas y bigotes se podran visualizar.

```
# Gráfico de cajas y bigotes
ggplot(datos, aes(x = cs, y = science, fill=sexo, alpha=0.5)) +
  geom_boxplot()
```




```
ggplot(datos, aes(x = cs, y = socst, fill=sexo, alpha=0.5)) +
  geom_boxplot()
```



```
#Mahalanobis
#Seleccionando las variables de interés
studyData = datos%>%select(science,socst)

#Cualculando las distancias
mahalanobis_distance = mahalanobis(studyData, colMeans(studyData), cov(studyData))

#Detectamos valores anómalos
outliers = scores(mahalanobis_distance, type = "z", prob =0.99)
cat(sum(outliers), "valores pertenecen al 1% más extremo")
```

```
7 valores pertenecen al 1% más extremo
```

```
#Mostrando 10 casos
data.frame(Index=which(outliers),
           SCIENCE = datos$science[outliers],
           SOCST = datos$socst[outliers])%>%arrange(SCIENCE, SOCST)%>%head(10)
```

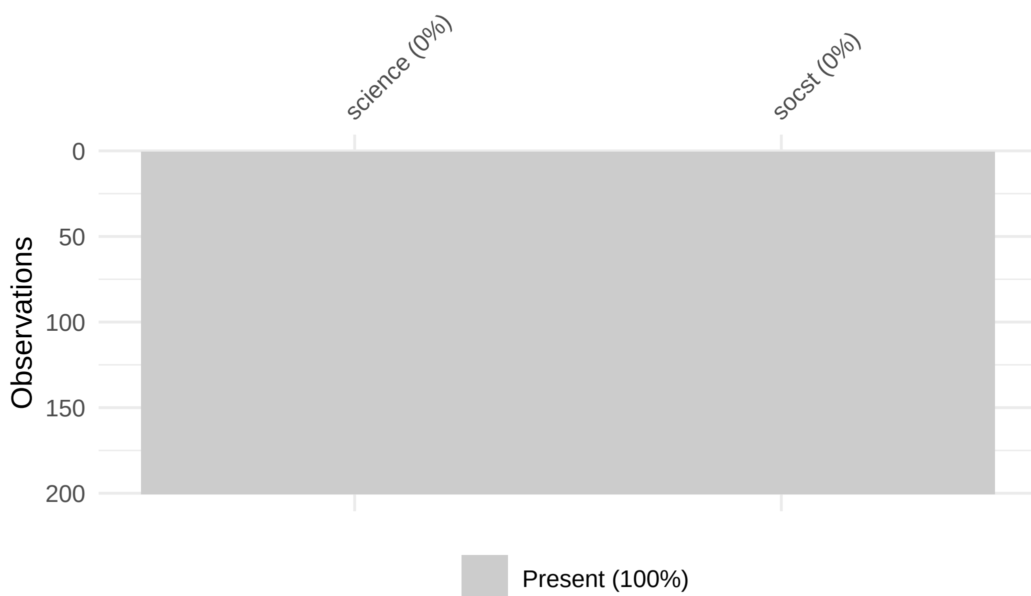
```
  Index SCIENCE SOCST
1     3      26    42
2     1      29    26
3    54      31    56
4    35      50    26
5    60      58    31
6   124      63    31
7   104      72    31
```

Contando y visualizando valores perdidos.

```
miss_var_summary(studyData)
```

```
# A tibble: 2 x 3
  variable n_miss pct_miss
  <chr>     <int>    <dbl>
1 science       0        0
2 socst         0        0
```

```
vis_miss( studyData,sort_miss = TRUE)
```



A continuación se presenta la aplicación de la regresión logística multinomial en el software R Statistical Computing. Primero se obtendrá toda la información y por último en el tercer apartado se presentaran los respectivos análisis e interpretaciones.

2.2 R Statistical Computing

Análisis bivariado de las variables dependiente e independientes

```
CrossTable(datos$female, datos$ses, chisq=T, prop.chisq = F, prop.t = F, prop.c = F)
```

```
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|-------------------------|


Total Observations in Table:  200


             | datos$ses
datos$female |         1 |         2 |         3 | Row Total |
-------------|-----------|-----------|-----------|-----------|
           0 |        15 |        47 |        29 |        91 |
             |     0.165 |     0.516 |     0.319 |     0.455 |
-------------|-----------|-----------|-----------|-----------|
           1 |        32 |        48 |        29 |       109 |
             |     0.294 |     0.440 |     0.266 |     0.545 |
-------------|-----------|-----------|-----------|-----------|
Column Total |        47 |        95 |        58 |       200 |
-------------|-----------|-----------|-----------|-----------|


Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  4.576532     d.f. =  2     p =  0.1014422
```

```
aggregate(cbind(science, socst)~ses, data = datos, mean, na.rm=T)
```

```
  ses  science    socst
1   1 47.70213 47.31915
2   2 51.70526 52.03158
3   3 55.44828 57.13793
```

Estimación del modelo con la categoría bajo en como nivel de referencia en la variable dependiente

```
multi1 =multinom(datos$ses ~ science + socst + female, data = datos)
```

```
# weights:  15 (8 variable)
initial  value 219.722458
iter  10 value 194.078046
final  value 194.034851
converged
```

```
summary(multi1)
```

```
Call:
multinom(formula = datos$ses ~ science + socst + female, data = datos)

Coefficients:
  (Intercept)    science      socst     female
2   -1.912305 0.02356541 0.03892473 -0.8166207
3   -5.969577 0.04648473 0.08192982 -0.8494647
```

```
Std. Errors:
  (Intercept)    science      socst    female
2    1.127259 0.02097473 0.01951656 0.3909821
3    1.437546 0.02510004 0.02383389 0.4482127

Residual Deviance: 388.0697
AIC: 404.0697
```

Estimacion del modelo solo con la constante (ratio de verosimilitud)

```
multi0 = multinom(ses ~ 1, data=datos)
```

```
# weights:  6 (2 variable)
initial  value 219.722458
final  value 210.582537
converged
```

Calculando el estadistico ji cuadrado como diferencia de sus -2LL (deviance)

```
chi2 = multi0$deviance - multi1$deviance
df.chi2 = multi1$edf - multi0$edf
sig.chi2 = 1 - pchisq(chi2, df.chi2)
print(cbind(chi2, df.chi2, sig.chi2))
```

```
         chi2 df.chi2     sig.chi2
[1,] 33.09537       6 1.005192e-05
```

Calculando el pseudo R^2 de McFadden (Significatividad global del modelo)

```
R2MF = chi2/multi0$deviance
print(R2MF)
```

```
[1] 0.07858052
```

Coeficientes de regresión del modelo y su significatividad

```
z = summary(multi1)$coefficients/summary(multi1)$standard.errors
z
```

```
  (Intercept)  science    socst    female
2   -1.696420 1.123514 1.994446 -2.088640
3   -4.152617 1.851978 3.437535 -1.895227
```

```
p = (1  -  pnorm ( abs (z),  0 ,  1 ))  *  2
p
```

```
   (Intercept)    science        socst     female
2 8.980643e-02 0.26121914 0.0461032998 0.03674018
3 3.286941e-05 0.06402897 0.0005870359 0.05806237
```

```
stargazer(multi1, type = "text")
```

```
==============================================
                      Dependent variable:
                  ----------------------------
                        2              3
                       (1)            (2)
----------------------------------------------
science               0.024          0.046*
                     (0.021)        (0.025)

socst                 0.039**        0.082***
                     (0.020)        (0.024)
```

```
female                 -0.817**       -0.849*
                       (0.391)        (0.448)


Constant               -1.912*       -5.970***
                       (1.127)        (1.438)


-----------------------------------------------
Akaike Inf. Crit.      404.070        404.070
===============================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```

Risk ratios del modelo estimado

```
multi1.rrr = exp(coef(multi1))
stargazer(multi1, type = "text", coef = list(multi1.rrr), p.auto = F)
```

```
===============================================
                      Dependent variable:
                  ----------------------------
                        2              3
                       (1)            (2)
-----------------------------------------------
science               1.024          1.048*
                      (0.021)        (0.025)


socst                 1.040**       1.085***
                      (0.020)        (0.024)


female                0.442**        0.428*
                      (0.391)        (0.448)
```

```
Constant               0.148*        0.003***
                      (1.127)        (1.438)


----------------------------------------------
Akaike Inf. Crit.    404.070        404.070
==============================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```

A continuación se presenta la aplicación de la regresión logística multinomial en el software SPSS Statistics. Primero se obtendrá toda la información y por último en el tercer apartado se presentaran los respectivos análisis e interpretaciones.

2.3 SPSS Statistics

Frecuencias para las variables categóricas de sexo y clase social con sus respectivos porcentajes margiales

Resumen de procesamiento de casos

		N	Porcentaje marginal
ses	low	47	23.5%
	middle	95	47.5%
	high	58	29.0%
female	male	91	45.5%
	female	109	54.5%
Válidos		200	100.0%
Perdidos		0	
Total		200	
Subpoblación		143[a]	

a. La variable dependiente sólo tiene un valor observado en 117 (81.8%) subpoblaciones.

Criterio de información de Akaike (AIC) del modelo estimado

Resumen de los pasos

			Criterios de ajuste de modelo			Pruebas de selección de efecto
Modelo	Acción	Efecto(s)	AIC	normalizado	Logaritmo de la verosimilitud -2	Chi-cuadrado[b]
0	Especificado	Intersección, female, science score, social studies score[a]	348.641	375.027	332.641	.

Método por pasos: Entrada hacia adelante

a. No se pueden añadir efectos en el modelo inicial.

b. El chi-cuadrado para la entrada se basa en la prueba de razón de verosimilitud.

Ajuste de modelo: Criterios AIC. BIC y chi cuadrado

Información de ajuste de los modelos

	Criterios de ajuste de modelo			Pruebas de la razón de verosimilitud		
Modelo	AIC	normalizado	Logaritmo de la verosimilitud -2	Chi-cuadrado	gl	Sig.
Sólo intersección	369.736	376.333	365.736			
Final	348.641	375.027	332.641	33.095	6	.000

Tabla de bondad de ajuste de Pearson y desviación

Bondad de ajuste

	Chi-cuadrado	gl	Sig.
Pearson	294.296	278	.240
Desvianza	287.613	278	.333

Significatividad global del modelo: R^2 de McFadden, R^2 de Cox y Snell y R^2 de Negelkerke

Pseudo R cuadrado

Cox y Snell	.153
Nagelkerke	.174
McFadden	.079

Significatividad del modelo por variable

Pruebas de la razón de verosimilitud

	Criterios de ajuste de modelo			Pruebas de la razón de verosimilitud		
Efecto	AIC de modelo reducido	BIC de modelo reducido	Logaritmo de la verosimilitud -2 de modelo reducido	Chi-cuadrado	gl	Sig.
Intersección	348.641	375.027	332.641[a]	.000	0	.
female	349.730	369.520	337.730	5.090	2	.078
science score	348.139	367.929	336.139	3.498	2	.174
social studies score	357.602	377.392	345.602	12.962	2	.002

El estadístico de chi-cuadrado es la diferencia de la log-verosimilitud -2 entre el modelo final y el modelo reducido. El modelo reducido se forma omitiendo un efecto del modelo final. La hipótesis nula es que todos los parámetros de dicho efecto son 0.

a. Este modelo reducido es equivalente al modelo final porque omitir el efecto no aumenta los grados de libertad.

Risk ratios del modelo estimado

Estimaciones de parámetro

								95% de intervalo de confianza para Exp(B)	
ses[a]		B	Desv. Error	Wald	gl	Sig.	Exp(B)	Límite inferior	Límite superior
middle	Intersección	-2.729	1.139	5.740	1	.017			
	[female=0]	.817	.391	4.362	1	.037	2.263	1.052	4.869
	[female=1]	0[b]	.	.	0	.	.	.	.
	science score	.024	.021	1.262	1	.261	1.024	.983	1.067
	social studies score	.039	.020	3.978	1	.046	1.040	1.001	1.080
high	Intersección	-6.819	1.442	22.351	1	.000			
	[female=0]	.849	.448	3.592	1	.058	2.338	.971	5.629
	[female=1]	0[b]	.	.	0	.	.	.	.
	science score	.046	.025	3.430	1	.064	1.048	.997	1.100
	social studies score	.082	.024	11.816	1	.001	1.085	1.036	1.137

a. La categoría de referencia es: low.

b. Este parámetro está establecido en cero porque es redundante.

A continuación se presenta la aplicación de la regresión logística multinomial en el lenguaje de programación Python. Primero se obtendrá toda la información y por último en el tercer apartado se presentaran los respectivos análisis e interpretaciones.

2.4 Python

Importanto paqueterías

```
import pandas as pd
import numpy as np
import pyreadstat as pr
from sklearn.metrics import accuracy_score
import statsmodels.api as sm
import warnings
warnings.filterwarnings('ignore')
```

Importando la base de datos y mostrando los nombres y tipos de las variables

```
datos = pd.read_spss(r"C:\Users\Fernando\OneDrive\Documentos\Ciclo II - 2023\Proyectos
datos.head()
```

```
      id  female     ses  schtyp  ... socst        honors  awards  cid
0   45.0  female     low  public  ...  26.0  not enrolled     0.0  1.0
1  108.0    male  middle  public  ...  36.0  not enrolled     0.0  1.0
2   15.0    male    high  public  ...  42.0  not enrolled     0.0  1.0
3   67.0    male     low  public  ...  32.0  not enrolled     0.0  1.0
4  153.0    male  middle  public  ...  51.0  not enrolled     0.0  1.0

[5 rows x 13 columns]
```

```
datos.dtypes
```

```
id            float64
female       category
ses          category
schtyp       category
prog         category
read          float64
write         float64
math          float64
science       float64
socst         float64
honors       category
awards        float64
cid           float64
dtype: object
```

Frecuencia de los datos para cada una de las variables categóricas que se utilizarán en la aplicación de la RLM

```
print(datos.ses.value_counts())
```

```
ses
middle    95
high      58
low       47
Name: count, dtype: int64
```

```
print(datos.female.value_counts())
```

```
female
female    109
male       91
Name: count, dtype: int64
```

Estadísticos descriptivos de las variables númericas del puntaje en la asignatura de ciencias y estudios sociales respectivamente

```
print(datos.science.describe())
```

```
count    200.000000
mean      51.850000
std        9.900891
min       26.000000
25%       44.000000
50%       53.000000
75%       58.000000
max       74.000000
Name: science, dtype: float64
```

```
print(datos.socst.describe())
```

```
count    200.000000
mean      52.405000
std       10.735793
min       26.000000
25%       46.000000
50%       52.000000
75%       61.000000
max       71.000000
Name: socst, dtype: float64
```

Análisis bivariado de las variables dependiente e independientes

```
Cross_Table=pd.crosstab(datos["female"],datos["ses"])
Cross_Table
```

```
ses      high  low  middle
female
female     29   32      48
male       29   15      47
```

Estimación y ajuste del modelo

```
#Seleccionando variables de estudio
y = datos['ses']
x = datos[["science", "socst"]]

# Haciendo female variable dummie
recod = pd.get_dummies(datos['female'], drop_first = True)
recod = pd.DataFrame(np.where(recod,1,0), columns = recod.columns)
x = pd.concat((x, recod), axis = 1)

# Estimando el modelo
mod = sm.MNLogit(y, sm.add_constant(x))
Resultado = mod.fit()
```

```
Optimization terminated successfully.
         Current function value: 0.970174
         Iterations 6
```

```
print(Resultado.summary2())
```

```
                        Results: MNLogit
=================================================================
Model:              MNLogit          Method:            MLE
Dependent Variable: ses              Pseudo R-squared: 0.079
Date:               2023-09-24 13:43 AIC:               404.0697
```

```
No. Observations:   200              BIC:              430.4562
Df Model:           6                Log-Likelihood:   -194.03
Df Residuals:       192              LL-Null:          -210.58
Converged:          1.0000           LLR p-value:      1.0052e-05
No. Iterations:     6.0000           Scale:            1.0000
-----------------------------------------------------------------
  ses = 0     Coef.   Std.Err.     t     P>|t|    [0.025   0.975]
-----------------------------------------------------------------
  const       6.8191    1.4424   4.7277  0.0000   3.9921   9.6461
science      -0.0465    0.0251  -1.8521  0.0640  -0.0957   0.0027
  socst      -0.0819    0.0238  -3.4375  0.0006  -0.1286  -0.0352
   male      -0.8495    0.4482  -1.8953  0.0581  -1.7280   0.0290
-----------------------------------------------------------------
  ses = 1     Coef.   Std.Err.     t     P>|t|    [0.025   0.975]
-----------------------------------------------------------------
  const       4.0902    1.2088   3.3835  0.0007   1.7209   6.4595
science      -0.0229    0.0209  -1.0982  0.2721  -0.0638   0.0180
  socst      -0.0430    0.0199  -2.1621  0.0306  -0.0820  -0.0040
   male      -0.0329    0.3500  -0.0939  0.9252  -0.7189   0.6532
=================================================================
```

Precisión del modelo estimado

```python
Clasificacion = pd.DataFrame(Resultado.pred_table(), columns = ['Alto', 'Bajo', 'Medio
index = ['Alto', 'Bajo', 'Medio'], dtype = np.int8)
print(Clasificacion)
```

```
       Alto  Bajo  Medio
Alto     19     2     37
Bajo      1    11     35
Medio    12     7     76
```

Con toda la información que se ha obtenido, estamos en condiciones de interpretar los resultados. A continuación se presenta el análisis de los resultados de forma general.

3. Análisis de los resultados

La tabla de contingencia o tabla cruzada, nos muestra a nivel bivariado, la relación que existe entre la variable dependiente, clase social y las variables independientes. El porcentaje de hombres en la clase media y alta es mayor que el de mujeres y, también, el desempeño de los estudios ha sido mayor cuanto mayor es la clase social.

Estando en condiciones de estimar la regresión logística, y tomando como referencia la categoría de clase social baja, debido a que si un individuo pertenece a ella se considera como éxito ascender a clase media o alta, entonces se estima el modelo, obteniendo como un primer vistazo que el modelo final se ajusta mejor en relación a los criterios AIC versus BIC, dado que este último es mayor. Sin embargo, es necesario realizar una prueba de bondad de ajuste para verificar si el modelo estimado se ajusta adecuadamente, por ellos se propone la bondad de ajuste de Pearson y la desviación, obteniendo como resultado 0.24 y 0.33 respectivamente. Dado que ambos valores son mayores que 0.05 no se rechaza la hipótesis nula de que los valores predichos por el modelo no difieren significativamente de los valores observados, es decir, existe un buen ajuste del modelo.

El R^2 de Nagelkerke es de 0.174, es decir, el modelo explica un 17.4% del cambio de la varianza de la variable dependiente (clase social). El pseudo R^2 de McFadden es de 0.079, lo que indica que el modelo solo predice el 7.9% del cambio de la clase social, ambas medidas correctoras son a priori demasiado bajos para el ajuste del modelo, es necesario recordar que estos pseudos R^2 no son un equivalente al R^2 que se presenta en la regresión múltiple que, es una medida muy intuitiva de lo bien que el modelo predice la variable dependiente. En conclusión, la capacidad de ajuste del modelo no es adecuada, sin embargo, para efectos prácticos de la aplicación de la regresión logística multinomial se continuará con el análisis de la importancia de los predictores en el modelo.

Recordando que la categoría de referencia de la variable dependiente es la clase social baja, con base a los coeficientes de regresión del modelo, el coeficiente de regresión de las notas en ciencias sociales es significativo y positivo para las clases medias y altas en relación con la clase baja. El signo de los coeficientes nos señala que un incremento en esta área de conocimiento incrementa la probabilidad de estar en la clase media y en la clase alta respecto a estar en la baja. Sin embargo, que el signo de la variable sexo es negativo para las dos clases sociales y que esta codificado como 1 (mujer), señala que, permaneciendo todos las demás constantes, ser mujer disminuye la probabilidad de estar en la las clases media y alta.

Para analizar el impacto relativo de las variables nos fijamos en los risk ratio superiores a la unidad entonces podemos ver que el mayor impacto sobre la movilidad social la tiene la variable sexo, puesto que ser mujer hace 0.87 veces más probable estar en la clase media respecto a estar en la baja, pero claro, multiplicar por 0.87 es una disminución de la probabilidad de ocurrencia.

A medida que el valor del puntaje de las calificaciones en la asignatura de ciencias aumenta en una unidad, un individuo tiene 1.024 veces más de probabilidades de pertenecer a la clase social media y 1.048 veces más de probabilidades de pertenecer a la clase alta. Además, a medida que el valor del puntaje de las calificaciones en la asignatura de estudios sociales aumenta en una unidad, un individuo tiene 1.040 veces más de probabilidades de pertenecer a la clase social media y 1.085 veces más de probabilidades de pertenecer a la clase alta.

Por último, la precisión del modelo solamente 19 de los 58 casos para la categoría de clase social alta fueron predichos correctamente; también 11 de los 47 casos para la categoría de clase social baja fueron predichos correctamente, y para la clase social media los valores predichos correctamente fueron 76 de 95.

4. Conclusiones

La regresión logística resulta útil para los casos en los que se desea predecir la presencia o ausencia de una característica o resultado según los valores de un conjunto de predictores. Es similar a un modelo de regresión lineal, pero está adaptado para modelos en los que la variable dependiente es dicotómica o politómica. Los coeficientes de regresión logística pueden utilizarse para estimar la razón de probabilidad de cada variable independiente del modelo. La regresión logística se puede aplicar a un rango más amplio de situaciones de investigación que el análisis discriminante. Para el caso de aplicación sobre la predicción de la clase social, con base al Pseudo R^2 de McFadden la estimación del modelo no se ajustó adecuadamente, y únicamente la categoría del puntaje en ciencia sociales predice significativamente para el nivel de clases sociales media y alta en relación a la clase baja. Con base a los risk ratios el mayor impacto sobre la movilidad social la tiene la variable sexo, puesto que ser mujer hace 0.87 veces más probable estar en la clase media respecto a estar en la baja. Finalmente, la capacidad predictora del modelo fue adecuada para la clase social media dado que los valores predichos correctamente fueron 76 de 95.