A7-Regresión logística

Rodolfo Jesús Cruz Rebollar

2024-11-06

```
# Librerías ISLR y tidyverse para lectura de los datos y realización de
modelo logístico

library(ISLR)

library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.3.3
```

```
## ── Attaching core tidyverse packages ──────────────────────────
tidyverse 2.0.0 ──
## ✓ dplyr     1.1.3     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.4     ✓ tibble    3.2.1
## ✓ lubridate 1.9.2     ✓ tidyr     1.3.0
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────────
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force
all conflicts to become errors
```

```
# Leer la base de datos de weekly de la librería ISLR

indice_bursatil = ISLR::Weekly

head(indice_bursatil)
```

```
##   Year   Lag1   Lag2   Lag3   Lag4   Lag5    Volume  Today Direction
## 1 1990  0.816  1.572 -3.936 -0.229 -3.484 0.1549760 -0.270      Down
## 2 1990 -0.270  0.816  1.572 -3.936 -0.229 0.1485740 -2.576      Down
## 3 1990 -2.576 -0.270  0.816  1.572 -3.936 0.1598375  3.514        Up
## 4 1990  3.514 -2.576 -0.270  0.816  1.572 0.1616300  0.712        Up
## 5 1990  0.712  3.514 -2.576 -0.270  0.816 0.1537280  1.178        Up
## 6 1990  1.178  0.712  3.514 -2.576 -0.270 0.1544440 -1.372      Down
```

1. Análisis de datos

```
# Calcular diferentes estadísticos descriptivos por variable
```

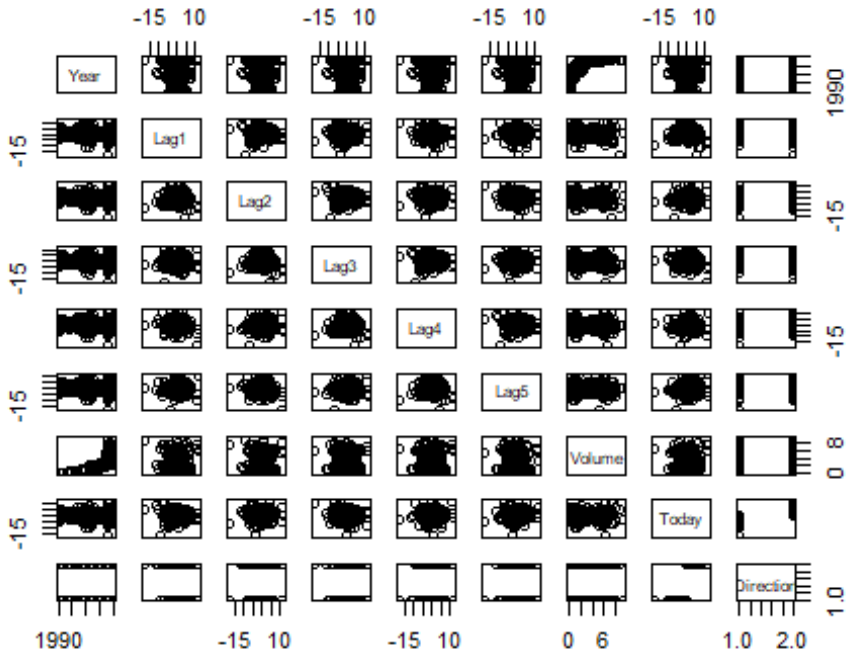
```r
desc_stat = summary(indice_bursatil[, c(-1, -9)])

desc_stat
```

```
##       Lag1                Lag2                Lag3                Lag4
##  Min.   :-18.1950   Min.   :-18.1950   Min.   :-18.1950   Min.   :-
18.1950
##  1st Qu.: -1.1540   1st Qu.: -1.1540   1st Qu.: -1.1580   1st Qu.: -
1.1580
##  Median :  0.2410   Median :  0.2410   Median :  0.2410   Median :
0.2380
##  Mean   :  0.1506   Mean   :  0.1511   Mean   :  0.1472   Mean   :
0.1458
##  3rd Qu.:  1.4050   3rd Qu.:  1.4090   3rd Qu.:  1.4090   3rd Qu.:
1.4090
##  Max.   : 12.0260   Max.   : 12.0260   Max.   : 12.0260   Max.   :
12.0260
##       Lag5              Volume            Today
##  Min.   :-18.1950   Min.   :0.08747   Min.   :-18.1950
##  1st Qu.: -1.1660   1st Qu.:0.33202   1st Qu.: -1.1540
##  Median :  0.2340   Median :1.00268   Median :  0.2410
##  Mean   :  0.1399   Mean   :1.57462   Mean   :  0.1499
##  3rd Qu.:  1.4050   3rd Qu.:2.05373   3rd Qu.:  1.4050
##  Max.   : 12.0260   Max.   :9.32821   Max.   : 12.0260
```

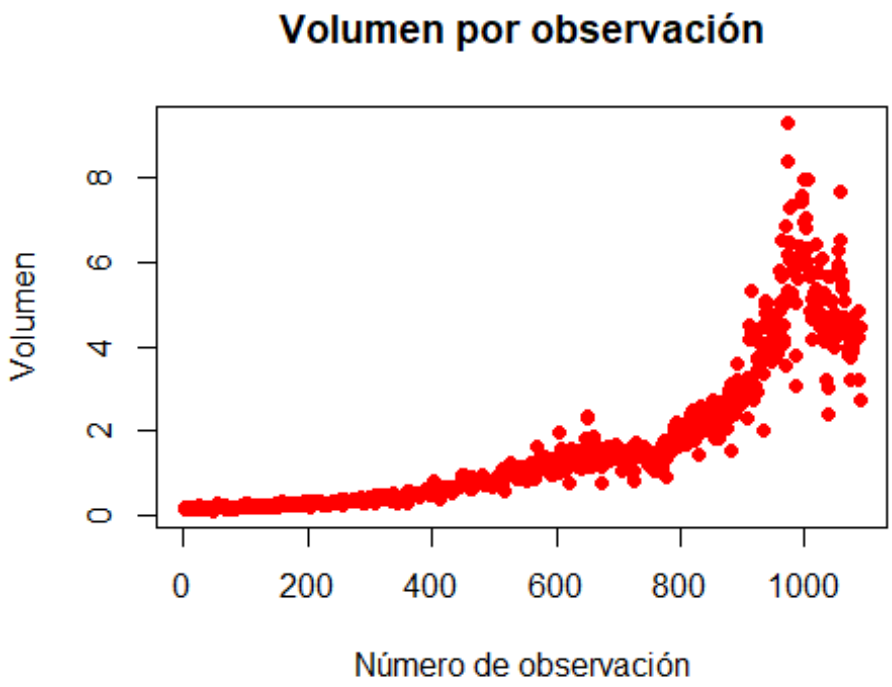
```r
# Forma alternativa de visualizar un resumen de los datos

glimpse(indice_bursatil)
```

```
## Rows: 1,089
## Columns: 9
## $ Year      <dbl> 1990, 1990, 1990, 1990, 1990, 1990, 1990, 1990,
1990, 1990, …
## $ Lag1      <dbl> 0.816, -0.270, -2.576, 3.514, 0.712, 1.178, -1.372,
0.807, 0…
## $ Lag2      <dbl> 1.572, 0.816, -0.270, -2.576, 3.514, 0.712, 1.178, -
1.372, 0…
## $ Lag3      <dbl> -3.936, 1.572, 0.816, -0.270, -2.576, 3.514, 0.712,
1.178, -…
## $ Lag4      <dbl> -0.229, -3.936, 1.572, 0.816, -0.270, -2.576, 3.514,
0.712, …
## $ Lag5      <dbl> -3.484, -0.229, -3.936, 1.572, 0.816, -0.270, -
2.576, 3.514,…
## $ Volume    <dbl> 0.1549760, 0.1485740, 0.1598375, 0.1616300,
0.1537280, 0.154…
## $ Today     <dbl> -0.270, -2.576, 3.514, 0.712, 1.178, -1.372, 0.807,
0.041, 1…
## $ Direction <fct> Down, Down, Up, Up, Up, Down, Up, Up, Up, Down,
Down, Up, Up…
```

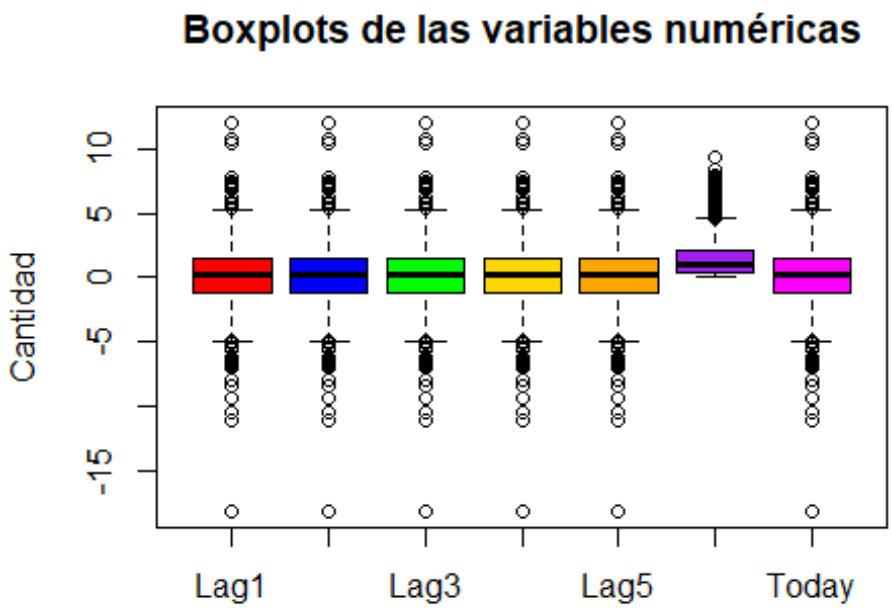
```r
# Graficar matriz de gráficos de dispersión de todas las variables del dataset

pairs(indice_bursatil)
```




```r
# Realizar un gráfico de los valores de volumen del dataset

plot(indice_bursatil$Volume, xlab = "Número de observación", ylab = "Volumen",
     main = "Volumen por observación", col = "red", pch = 19)
```




```r
# Graficar boxplot por cada variable numérica del dataset

boxplot(indice_bursatil[, c(-1, -9)], main = "Boxplots de las variables numéricas",
        ylab = "Cantidad",
        col = c("red", "blue", "green", "gold", "orange", "purple", "magenta"))
```




```r
# Calcular coeficientes de correlación entre variables numéricas

r_coefs = cor(indice_bursatil[, c(-1, -9)])

r_coefs
```

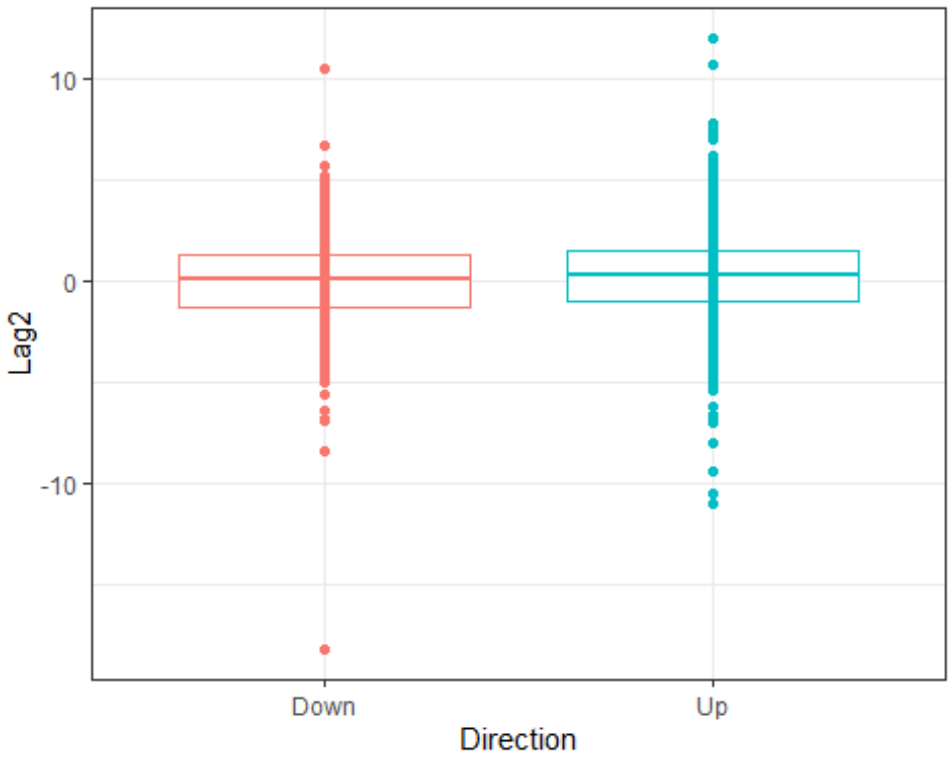
```
##                Lag1        Lag2        Lag3         Lag4         Lag5
## Lag1    1.000000000 -0.07485305  0.05863568 -0.071273876 -0.008183096
## Lag2   -0.074853051  1.00000000 -0.07572091  0.058381535 -0.072499482
## Lag3    0.058635682 -0.07572091  1.00000000 -0.075395865  0.060657175
## Lag4   -0.071273876  0.05838153 -0.07539587  1.000000000 -0.075675027
## Lag5   -0.008183096 -0.07249948  0.06065717 -0.075675027  1.000000000
## Volume -0.064951313 -0.08551314 -0.06928771 -0.061074617 -0.058517414
## Today  -0.075031842  0.05916672 -0.07124364 -0.007825873  0.011012698
##              Volume        Today
## Lag1   -0.06495131 -0.075031842
## Lag2   -0.08551314  0.059166717
## Lag3   -0.06928771 -0.071243639
## Lag4   -0.06107462 -0.007825873
## Lag5   -0.05851741  0.011012698
## Volume  1.00000000 -0.033077783
## Today  -0.03307778  1.000000000
```

2. Formulación de modelo logístico

```
# Formular el modelo de regresión logística con todas las variable menos
"today"

log.model = glm(Direction ~. - Today, data = indice_bursatil, family =
"binomial")

# Resumen del modelo logístico

summary(log.model)

##
## Call:
## glm(formula = Direction ~ . - Today, family = "binomial", data =
indice_bursatil)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 17.225822  37.890522   0.455   0.6494
## Year        -0.008500   0.018991  -0.448   0.6545
## Lag1        -0.040688   0.026447  -1.538   0.1239
## Lag2         0.059449   0.026970   2.204   0.0275 *
## Lag3        -0.015478   0.026703  -0.580   0.5622
## Lag4        -0.027316   0.026485  -1.031   0.3024
## Lag5        -0.014022   0.026409  -0.531   0.5955
## Volume       0.003256   0.068836   0.047   0.9623
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1496.2  on 1088  degrees of freedom
## Residual deviance: 1486.2  on 1081  degrees of freedom
## AIC: 1502.2
##
## Number of Fisher Scoring iterations: 4

# Visualizar contrastes asociados con un factor en particular

contrasts(indice_bursatil$Direction)

##      Up
## Down  0
## Up    1

# Calcular intervalos de confianza para los coeficientes del modelo
logístico

confint(object = log.model, level = 0.95)
```

```
## Waiting for profiling to be done...

##                     2.5 %      97.5 %
## (Intercept) -56.985558236 91.66680901
## Year          -0.045809580  0.02869546
## Lag1          -0.092972584  0.01093101
## Lag2           0.007001418  0.11291264
## Lag3          -0.068140141  0.03671410
## Lag4          -0.079519582  0.02453326
## Lag5          -0.066090145  0.03762099
## Volume        -0.131576309  0.13884038
```

```
# Gráfico de la variable significativa Lag2 encontrada en el modelo previo

ggplot(data = indice_bursatil, mapping = aes(x = Direction, y = Lag2)) +
geom_boxplot(aes(color = Direction)) +
geom_point(aes(color = Direction)) +
theme_bw() +
theme(legend.position = "null")
```



Prueba de hipótesis para significancia de coeficientes β_i:

H_0: $\beta_i = 0$ (el predictor β_i no es estadísticamente significativo).

H_1: $\beta_i \neq 0$ (el predictor β_i sí es estadísticamente significativo).

En base a los resultados anteriores, es posible observar que en el resumen del modelo logístico, todas las variables predictoras sin tomar en cuenta a la variable de respuesta `today` y a la variable predictora `Lag2`, poseen un p valor bastante superior a 0.05, lo cual significa que en el caso de todas las variables a excepción de `today` y `Lag2`, no se rechaza la hipótesis nula H_0 de la prueba de hipótesis para verificación de significancia de los coeficientes del modelo logístico, motivo por el cual, se afirma que todas esas variables predictoras no son estadísticamente significativas, lo cual a su vez implica que dichas variables no influyen en el modelo logístico, no obstante, para el caso específico de la variable `Lag2`, se observa que su valor p es igual a 0.0275, lo cual al ser inferior a 0.05, provoca que se rechaze H_0, motivo por el cual, eso indica que la variable `Lag2` sí influye en el modelo logístico estadísticamente hablando, mientras que el resto de variables no.

Por otra parte, es importante mencionar que tomando en cuenta solamente aquellas variables predictoras que sí tienen influencia en el modelo desde la perspectiva estadística (`Lag2`), el coeficiete de dicha variable en el modelo logístico es igual a 0.0594486, lo cual hace referencia a que por cada unidad que cambia (aumenta o disminuye) el valor de la variable `Lag2`, el logaritmo de los momios u odds experimenta un cambio (incremento o disminución) de 0.0594486.

3. Creación de conjunto de entrenamiento y prueba y reajuste del modelo

```
# Crear conjunto de datos de entrenamiento (datos desde 1990 hasta 2008)

dataset_train = (indice_bursatil$Year >= 1990) & (indice_bursatil$Year <=
2008)

# Crear conjunto de datos de prueba (datos de 2009 y 2010)

dataset_test = indice_bursatil[(indice_bursatil$Year == 2009) |
                                 (indice_bursatil$Year == 2010), ]
```

Nota: el modelo ajustado a la única variable significativa encontrada se realiza en la siguiente parte.

4. Formulación de modelo logístico sólo con variables significativas en entrenamiento

En los resultados del modelo reajustado, se aprecia que la única variable predictora que resulta ser estadísticamente significativa en el modelo es `Lag2`, por lo que a continuación se procederá a formular otro modelo de regresión logística que involucre solamente a la única variable predictora `Lag2` que resultó ser significativa en el dataset de entrenamiento.

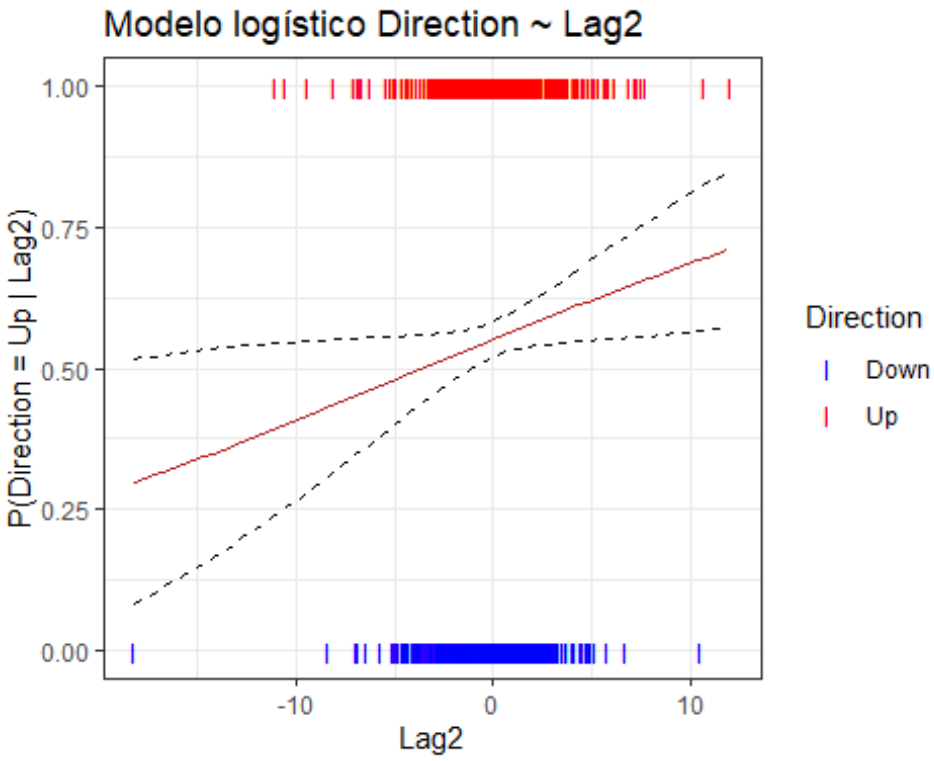
```r
# Formulación de nuevo modelo logístico solo con variables significativas
en entrenamiento
# Nota: la única variable significativa detectada es Lag2, por lo que se
procederá a
# reformular el modelo logístico solamente con esa variable predictora

lmodel_sig = glm(Direction ~ Lag2, data = indice_bursatil, family =
"binomial",
                 subset = dataset_train)

summary(lmodel_sig)
```

```
##
## Call:
## glm(formula = Direction ~ Lag2, family = "binomial", data =
indice_bursatil,
##     subset = dataset_train)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.20326    0.06428   3.162  0.00157 **
## Lag2         0.05810    0.02870   2.024  0.04298 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1354.7  on 984  degrees of freedom
## Residual deviance: 1350.5  on 983  degrees of freedom
## AIC: 1354.5
##
## Number of Fisher Scoring iterations: 4
```

5. Representación gráfica del modelo

```r
# Vector con nuevos valores interpolados dentro del rango del predictor
Lag2

new_points = seq(from = min(indice_bursatil$Lag2), to =
max(indice_bursatil$Lag2),
                 by = 0.5)

# Predicción de los nuevos puntos según el modelo con el comando
predict() se
# calcula la probabilidad de que la variable respuesta pertenezca al
nivel de
# referencia (en este caso “Up”)
```

```r
forecast = predict(lmodel_sig, newdata = data.frame(Lag2 = new_points),
                   se.fit = TRUE, type = "response")

# Límites del intervalo de confianza (95%) de las predicciones

CI_inferior = forecast$fit - 1.96 * forecast$se.fit

CI_superior = forecast$fit + 1.96 * forecast$se.fit

# Matriz de datos con los nuevos puntos y sus predicciones

curve_data =  data.frame(Lag2 = new_points,
                         probabilidad = forecast$fit,
                         CI.inferior = CI_inferior,
                         CI.superior = CI_superior)

# Codificación 0,1 de la variable respuesta Direction

indice_bursatil$Direction = ifelse(indice_bursatil$Direction == "Up", yes
= 1, no = 0)

ggplot(indice_bursatil, aes(x = Lag2, y = Direction)) +

geom_point(aes(color = as.factor(Direction)), shape = "I", size = 3) +

geom_line(data = curve_data, aes(y = probabilidad), color = "firebrick")
+

geom_line(data = curve_data, aes(y = CI.superior), linetype = "dashed") +

geom_line(data = curve_data, aes(y = CI.inferior), linetype = "dashed") +
labs(title = "Modelo logístico Direction ~ Lag2", y = "P(Direction = Up |
Lag2)",
     x = "Lag2") +
scale_color_manual(labels = c("Down", "Up"), values = c("blue", "red")) +

guides(color=guide_legend("Direction")) +

theme(plot.title = element_text(hjust = 0.5)) +

theme_bw()
```



6. Evaluación del modelo

Prueba de χ^2

Prueba de hipótesis:

H_0: El modelo nulo explica adeucadamente el log de los nomios (la probabilidad es constante).

H_1: El modelo con predictores explica mejor el log de los nomios (la probabilidad está sujeta a los predictores).

```
# Realizar una prueba de chi2 para evaluar significancia del modelo con
predictores con
# respecto al modelo nulo

prueba_chi2 = anova(lmodel_sig, test = "Chisq")

prueba_chi2
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Direction
##
```

```
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                  984     1354.7
## Lag2  1   4.1666       983     1350.5  0.04123 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En los resultados de la prueba χ^2 se aprecia que el p valor del test es igual a 0.04123, lo cual al ser menor que 0.05, provoca que se rechaze H_0, motivo por el cual, se puede afirmar que el modelo con predictores explica de mejor manera el log de los nomios, por lo que en conclusión, el modelo que sí posee predictores, sí resulta ser útil para explicar los datos del índice bursátil en función de la variable predictora `Lag2` (valores del mercado en las 2 semanas anteriores).

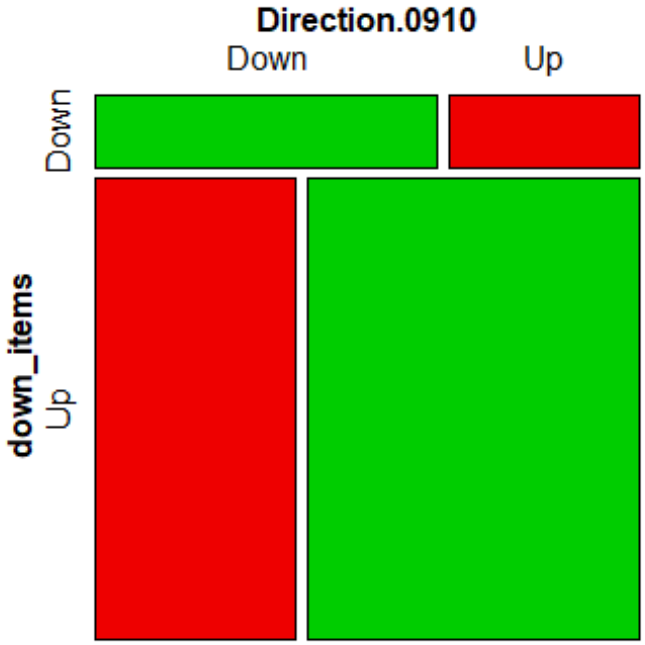
Matriz de confusión

```
# Cálculo de probabilidad predicha por el modelo con los datos de prueba
o test

prob_test = predict(lmodel_sig, newdata = dataset_test, type =
"response")

# Vector de elementos "Down"

down_items = rep("Down", length(prob_test))

# Sustitución de "Down" por "Up" si la p > 0.5

down_items[prob_test > 0.5] = "Up"

Direction = ifelse(indice_bursatil$Direction == 1, yes = "Up", no =
"Down")

Direction.0910 = Direction[!dataset_train]

# Matriz de confusión

conf_matrix = table(down_items, Direction.0910)

conf_matrix
```

```
##            Direction.0910
## down_items Down Up
##       Down    9  5
##       Up     34 56
```

```
library(vcd)
```

```
## Loading required package: grid

##
## Attaching package: 'vcd'

## The following object is masked from 'package:ISLR':
##
##     Hitters

mosaic(conf_matrix, shade = TRUE, colorize = TRUE,
gp = gpar(fill = matrix(c("green3", "red2", "red2", "green3"), 2, 2)))
```




```
mean(down_items == Direction.0910)

## [1] 0.625
```

En los resultados de evaluación previos, se observa que en la matriz de confusión del modelo, el modelo logístico fue capaz de predecir correctamente 65 de un total de 104 registros del dataset de prueba, lo cual da como resultado un nivel de exactitud del 62.5%, lo cual hace referencia al hecho de que el modelo logístico fue capaz de clasificar correctamente el 62.5% de los registros del dataset de prueba, por lo que el modelo resulta ser mayormente adecuado para predecir la tendencia del índice bursátil en función de los valores del mercado de las últimas 2 semanas (`Lag2`).

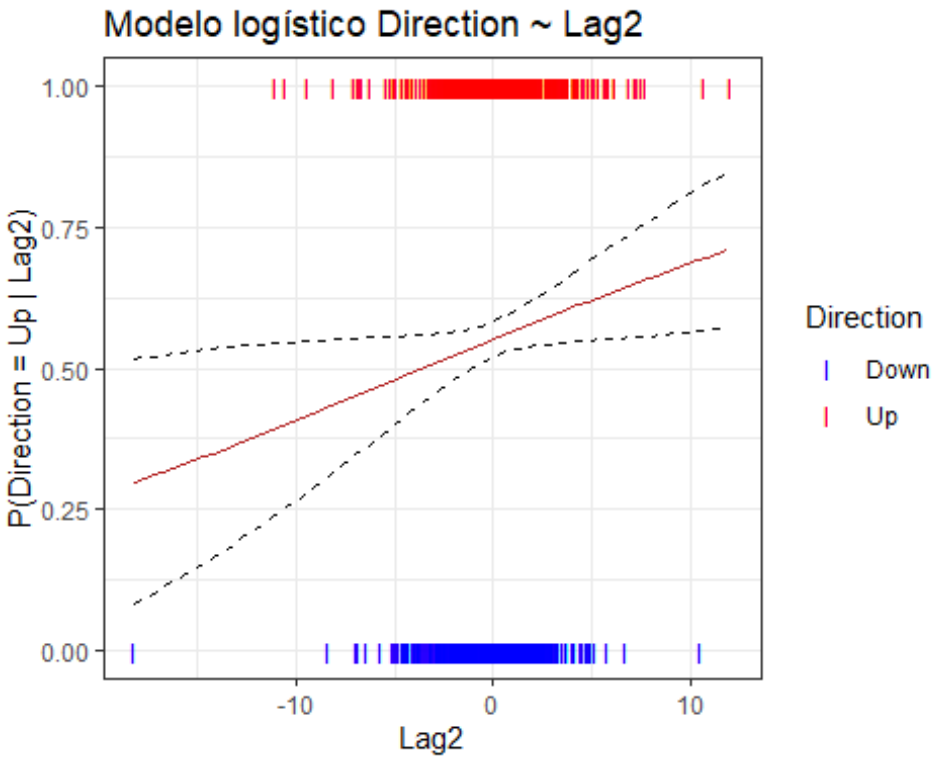
7. Ecuación, gráfica e interpretación del modelo logístico significativo

Ecuación del modelo logístico significativo

$$log\left(\frac{p(Direction = Up|Lag2)}{1 - p(Direction = Up|Lag2)}\right) = 0.2033 + 0.0581Lag_2$$

Gráfica del modelo logístico significativo

```
# Graficar el modelo logístico significativo (aquel solamente con Lag2
como predictora)

ggplot(indice_bursatil, aes(x = Lag2, y = Direction)) +

geom_point(aes(color = as.factor(Direction)), shape = "I", size = 3) +

geom_line(data = curve_data, aes(y = probabilidad), color = "firebrick")
+

geom_line(data = curve_data, aes(y = CI.superior), linetype = "dashed") +

geom_line(data = curve_data, aes(y = CI.inferior), linetype = "dashed") +
labs(title = "Modelo logístico Direction ~ Lag2", y = "P(Direction = Up |
Lag2)",
     x = "Lag2") +
scale_color_manual(labels = c("Down", "Up"), values = c("blue", "red")) +

guides(color=guide_legend("Direction")) +

theme(plot.title = element_text(hjust = 0.5)) +

theme_bw()
```



Interpretación del modelo logístico significativo

En cuanto a la interpretación del modelo logístico significativo dentro del contexto del problema, es posible afirmar que el modelo calcula la probabilidad de que el índice bursátil suba o baje, esto de acuerdo con el valor del mercado en las 2 semanas previas al momento actual, lo cual significa que dependiendo de cuál haya sido el valor del mercado en las últimas 2 semanas, se calculará la probabilidad de la tendencia del índice bursátil, motivo por el cual, en caso de que el valor del mercado en las 2 semanas más recientes resulte ser un valor negativo mayormente grande, entonces la probabilidad del índice bursátil será mayormente baja, por lo que la tendencia de dicho índice será de tipo "Down" (tendencia bajista), en caso contrario, si el valor del mercado en las últimas 2 semanas es un valor positivo grande, entonces la probabilidad del índice bursátil será en su mayoría alta, por lo que la tendencia del índice bursátil será de tipo "Up" (tendencia alcista).

Por otro lado, también es importante mencionar que el modelo logístico con la variable significativa resulta ser en su mayoría un buen modelo, especialmente para predecir tendencias de tipo "Up" (alcistas), dado que el modelo fue capaz de predecir correctamente 56 de las 61 tendencias alcistas que hay en total en el dataset de prueba, lo que representa el 91.8032787% de todos los registros de tendencias alcistas del dataset de prueba, mientras que dicho modelo, por el contrario, solamente fue capaz de predecir de forma acertada 9 tendencias bajistas ("Down") de las 43 que hay en total en el dataset de prueba, lo cual representa el 20.9302326% de todos los registros de tendencias bajistas del dataset de prueba, por lo que esto señala que el modelo es bueno para predecir específicamente tendencias alcistas del índice bursátil,

mientras que no es bueno para predecir tendencias bajistas de dicho índice bursátil. Finalmente, al momento de predecir el tipo de tendencia del índice bursátil, se aprecia que la capacidad predictiva del modelo cambia significativamente en el sentido de que pasa de tener un alto nivel de exactitud cuando se predicen tendencias alcistas, a tener una baja exactitud al momento de tener que predecir tendencias bajistas del índice bursátil del mercado, por lo cual, la capacidad predictiva del modelo cambia notablemente al predecir ambos tipos de tendencias porque básicamente existe un sesgo en los datos de entrenamiento, dado que el modelo se entrena con un total de 488 datos con tendencia bajista ("Low") y 601 datos con tendencia alcista ("Up"), motivo por el cual, se aprecia que el modelo se entrena más con datos clasificados como "Up", por lo que hay menos datos clasificados como "Down" para entrenar el modelo, provocando así un sesgo en las clasificaciones derivadas del modelo, de forma que hay una mayor cantidad de predicciones de tendencia alcista ("Up") que bajista ("Down").