Act_7_A01742161

Rogelio Lizárraga

2024-11-05

Regresión logística

Trabaja con el set de datos Weekly, que forma parte de la librería ISLR. Este set de datos contiene información sobre el rendimiento porcentual semanal del índice bursátil S&P 500 entre los años 1990 y 2010. Se busca predecir el tendimiento (positivo o negativo) dependiendo del comportamiento previo de diversas variables de la bolsa bursátil S&P 500.

Encuentra un modelo logístico para encontrar el mejor conjunto de predictores que auxilien a clasificar la dirección de cada observación.

Se cuenta con un set de datos con 9 variables (8 numéricas y 1 categórica que será nuestra variable respuesta: Direction). Las variables Lag son los valores de mercado en semanas anteriores y el valor del día actual (Today). La variable volumen (Volume) se refiere al volumen de acciones. Realiza: cambia)

```
library(ISLR)
library(MASS)
data("Weekly")
str(Weekly)
```

```
## 'data.frame':    1089 obs. of  9 variables:
##  $ Year     : num  1990 1990 1990 1990 1990 1990 1990 1990 1990 1990 ...
##  $ Lag1     : num  0.816 -0.27 -2.576 3.514 0.712 ...
##  $ Lag2     : num  1.572 0.816 -0.27 -2.576 3.514 ...
##  $ Lag3     : num  -3.936 1.572 0.816 -0.27 -2.576 ...
##  $ Lag4     : num  -0.229 -3.936 1.572 0.816 -0.27 ...
##  $ Lag5     : num  -3.484 -0.229 -3.936 1.572 0.816 ...
##  $ Volume   : num  0.155 0.149 0.16 0.162 0.154 ...
##  $ Today    : num  -0.27 -2.576 3.514 0.712 1.178 ...
##  $ Direction: Factor w/ 2 levels "Down","Up": 1 1 2 2 2 1 2 2 2 1 ...
```

```
summary(Weekly)
```

```
##       Year           Lag1               Lag2               Lag3         
##  Min.   :1990   Min.   :-18.1950   Min.   :-18.1950   Min.   :-18.1950  
##  1st Qu.:1995   1st Qu.: -1.1540   1st Qu.: -1.1540   1st Qu.: -1.1580  
##  Median :2000   Median :  0.2410   Median :  0.2410   Median :  0.2410  
##  Mean   :2000   Mean   :  0.1506   Mean   :  0.1511   Mean   :  0.1472  
##  3rd Qu.:2005   3rd Qu.:  1.4050   3rd Qu.:  1.4090   3rd Qu.:  1.4090  
##  Max.   :2010   Max.   : 12.0260   Max.   : 12.0260   Max.   : 12.0260  
##       Lag4               Lag5              Volume            Today         
##  Min.   :-18.1950   Min.   :-18.1950   Min.   :0.08747   Min.   :-18.1950  
```

```
##  1st Qu.: -1.1580   1st Qu.: -1.1660   1st Qu.:0.33202   1st Qu.: -1.1540
##  Median :  0.2380   Median :  0.2340   Median :1.00268   Median :  0.2410
##  Mean   :  0.1458   Mean   :  0.1399   Mean   :1.57462   Mean   :  0.1499
##  3rd Qu.:  1.4090   3rd Qu.:  1.4050   3rd Qu.:2.05373   3rd Qu.:  1.4050
##  Max.   : 12.0260   Max.   : 12.0260   Max.   :9.32821   Max.   : 12.0260
##  Direction
##  Down:484
##  Up  :605
##
##
##
##
```

```
head(Weekly)
```

```
##   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. El análisis de datos. Estadísticas descriptivas y coeficiente de correlación entre las variables.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:MASS':
##
##     select

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
head(Weekly)
```

```
##   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
```

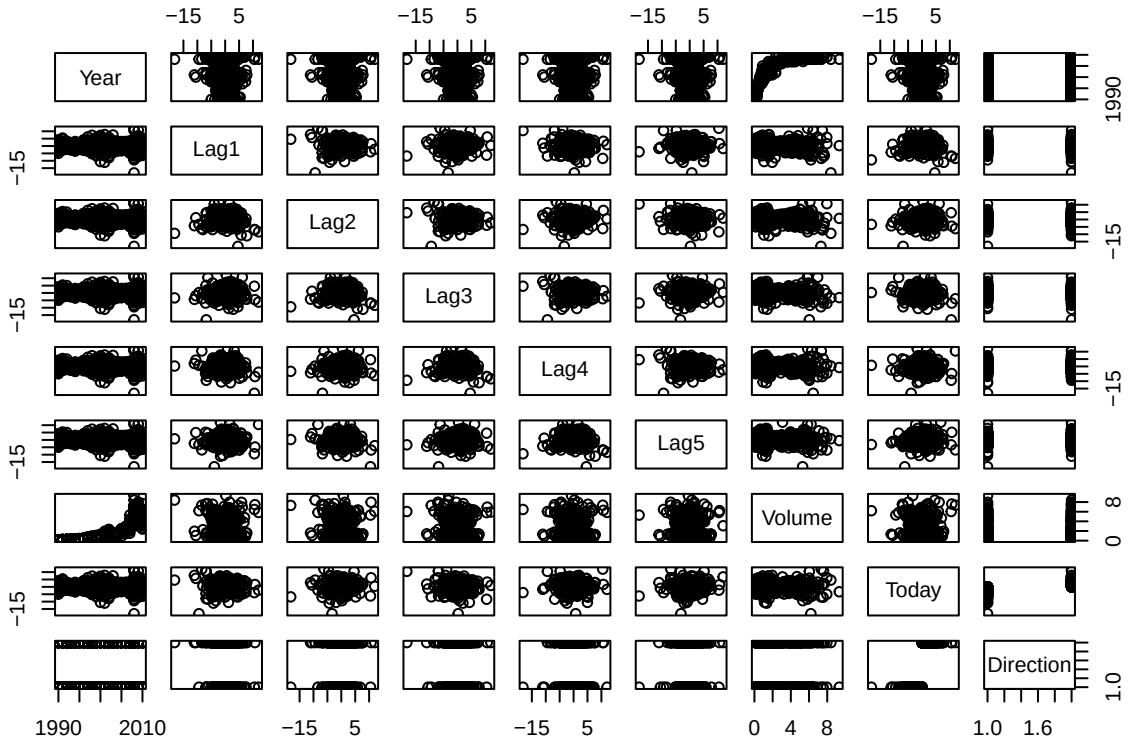
```
glimpse(Weekly)
```

```
## 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~
```

```
summary(Weekly)
```

```
##       Year           Lag1               Lag2               Lag3         
##  Min.   :1990   Min.   :-18.1950   Min.   :-18.1950   Min.   :-18.1950  
##  1st Qu.:1995   1st Qu.: -1.1540   1st Qu.: -1.1540   1st Qu.: -1.1580  
##  Median :2000   Median :  0.2410   Median :  0.2410   Median :  0.2410  
##  Mean   :2000   Mean   :  0.1506   Mean   :  0.1511   Mean   :  0.1472  
##  3rd Qu.:2005   3rd Qu.:  1.4050   3rd Qu.:  1.4090   3rd Qu.:  1.4090  
##  Max.   :2010   Max.   : 12.0260   Max.   : 12.0260   Max.   : 12.0260  
##       Lag4               Lag5              Volume            Today         
##  Min.   :-18.1950   Min.   :-18.1950   Min.   :0.08747   Min.   :-18.1950  
##  1st Qu.: -1.1580   1st Qu.: -1.1660   1st Qu.:0.33202   1st Qu.: -1.1540  
##  Median :  0.2380   Median :  0.2340   Median :1.00268   Median :  0.2410  
##  Mean   :  0.1458   Mean   :  0.1399   Mean   :1.57462   Mean   :  0.1499  
##  3rd Qu.:  1.4090   3rd Qu.:  1.4050   3rd Qu.:2.05373   3rd Qu.:  1.4050  
##  Max.   : 12.0260   Max.   : 12.0260   Max.   :9.32821   Max.   : 12.0260  
##  Direction
##  Down:484
##  Up  :605
##
##
##
##
```

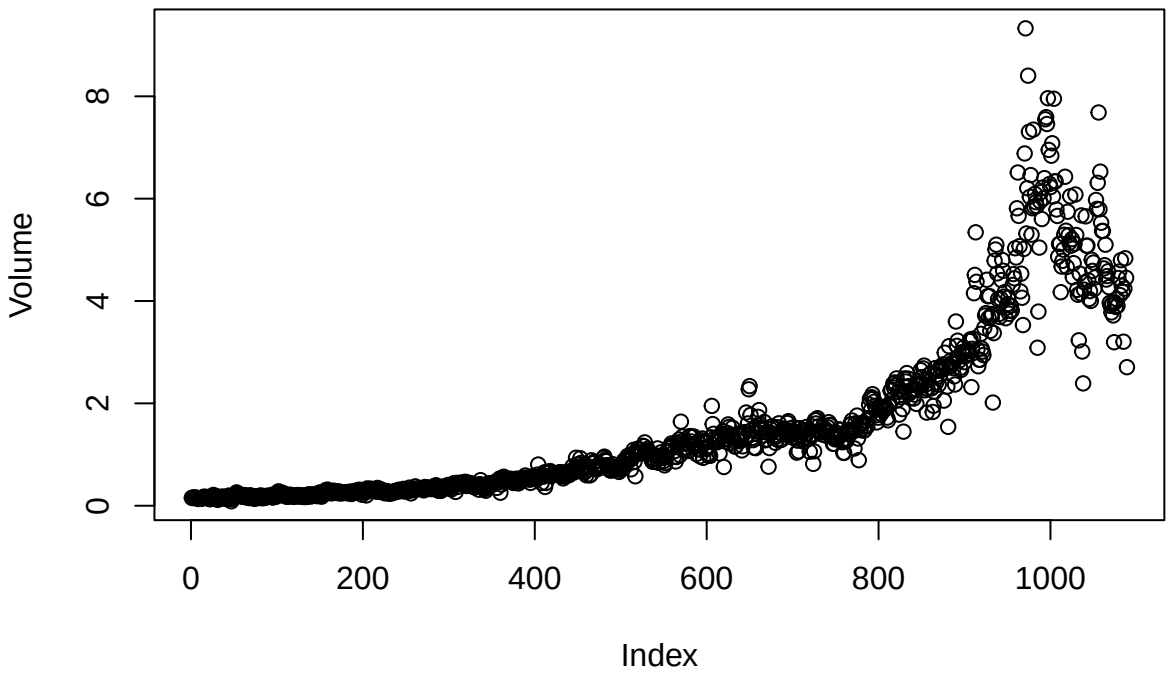
```
pairs(Weekly)
```




```
cor(Weekly[, -9])
```

```
##                Year         Lag1        Lag2        Lag3         Lag4
## Year     1.00000000 -0.032289274 -0.03339001 -0.03000649 -0.031127923
## Lag1    -0.03228927  1.000000000 -0.07485305  0.05863568 -0.071273876
## Lag2    -0.03339001 -0.074853051  1.00000000 -0.07572091  0.058381535
## Lag3    -0.03000649  0.058635682 -0.07572091  1.00000000 -0.075395865
## Lag4    -0.03112792 -0.071273876  0.05838153 -0.07539587  1.000000000
## Lag5    -0.03051910 -0.008183096 -0.07249948  0.06065717 -0.075675027
## Volume   0.84194162 -0.064951313 -0.08551314 -0.06928771 -0.061074617
## Today   -0.03245989 -0.075031842  0.05916672 -0.07124364 -0.007825873
##                Lag5      Volume        Today
## Year    -0.030519101  0.84194162 -0.032459894
## Lag1    -0.008183096 -0.06495131 -0.075031842
## Lag2    -0.072499482 -0.08551314  0.059166717
## Lag3     0.060657175 -0.06928771 -0.071243639
## Lag4    -0.075675027 -0.06107462 -0.007825873
## Lag5     1.000000000 -0.05851741  0.011012698
## Volume  -0.058517414  1.00000000 -0.033077783
## Today    0.011012698 -0.03307778  1.000000000
```

```
attach(Weekly)
plot(Volume)
```



En la matriz de correlación entre las variables, se observa que Volume tiene una correlación moderada con el año (Year), indicando un posible crecimiento en el volumen de acciones con el tiempo.

La matriz de gráficos de dispersión permite verificar visualmente las relaciones entre las variables de rezago (Lag1, Lag2, etc.) y la variable de respuesta (Direction), aunque, como se puede observar, las correlaciones visuales entre estas variables parecen ser bajas.

Finalmente, en el gráfico de dispersión del volumen (Volume) en función de los índices de las observaciones observmamos cómo la tendencia creciente hacia el final del periodo sugiere que el volumen de acciones aumentó durante los años.

2. Formula un modelo logístico con todas las variables menos la variable "Today". Calcula los intervalos de confianza para las β_i. Detecta variables que influyen y no influyen en el modelo. Interpreta el efecto de la variables en los odds (momios).

$H_0 = \beta_i$ no es significativo. $H_1 = \beta_i$ sí es significativo.

```
modelo.log.m <- glm(Direction ~ . -Today, data
= Weekly, family = binomial)
summary(modelo.log.m)
```

```
##
## Call:
```

```
## glm(formula = Direction ~ . - Today, family = binomial, data = Weekly)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7071  -1.2578   0.9941   1.0873   1.4665
##
## 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
```

```
contrasts(Direction)
```

```
##      Up
## Down  0
## Up    1
```

```
confint(object = modelo.log.m, 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
```

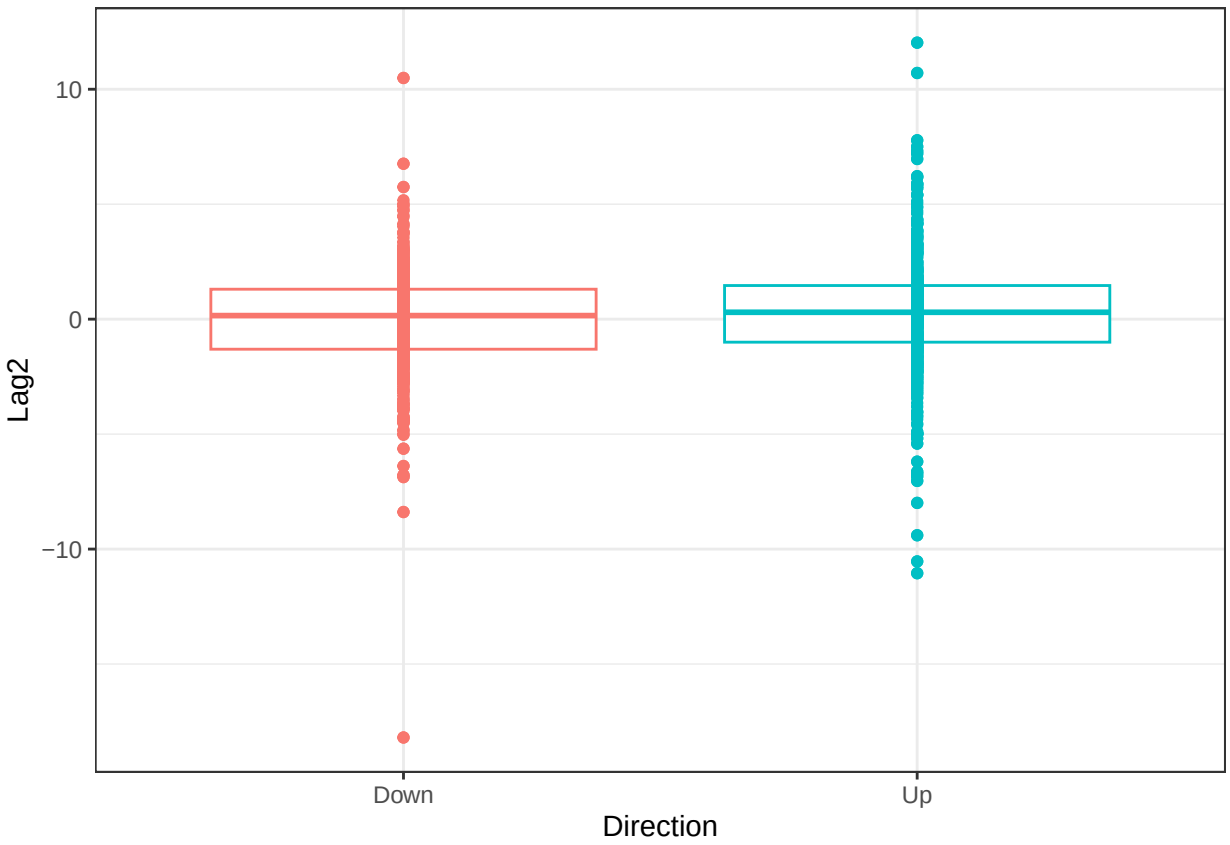
Observando los resultados de nuestro estadístico z y nuestros valores p, vemos que solo Lag2 rechaza la hipótesis inicial con un nivel de significancia de 0.05, por lo que solo la variable Lag2 es significativa para el modelo. Para el resto, no se rechaza la hipótesis inicial, por lo que no son estadísticamente significativas para el modelo.

Por ello, haremos solo un modelo con la variable Lag 2.

Gráfico de la variable significativa Lag2 (boxplot):

```
library(ggplot2)

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



Como podemos observar, Lag2 parece comportarse igual para la dirección Arriba y Abajo. Es decir, sus medianas están prácticamente en el mismo punto, aunque Down tiene datos atípicos con desviaciones mayores, mientras Up tiene más datos que se salen dentro del rango de la caja.

3. Divide la base de datos en un conjunto de entrenamiento (datos desde 1990 hasta 2008) y de prueba (2009 y 2010). Ajusta el modelo encontrado.

4. Formula el modelo logístico sólo con las variables significativas en la base de entrenamiento.

$H_0 = \beta_i$ no es significativo. $H_1 = \beta_i$ sí es significativo.

```r
#Training: observaciones desde 1990 hasta 2008
datos.entrenamiento <- (Year < 2009)
# Test: observaciones de 2009 y 2010
datos.test <- Weekly[!datos.entrenamiento, ]
# Verifica:
nrow(datos.entrenamiento) + nrow(datos.test)
```

```
## integer(0)
```

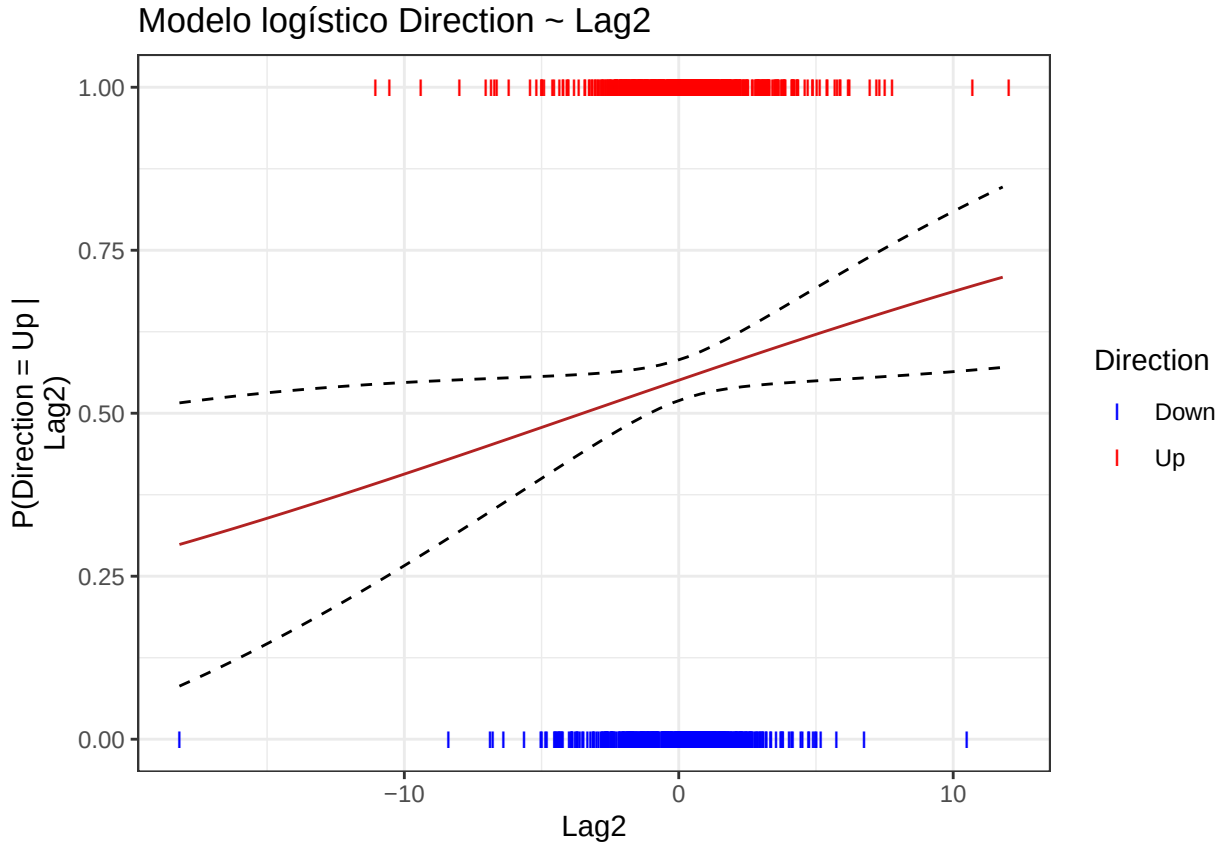
```r
# Ajuste del modelo logístico con variables significativas
modelo.log.s <- glm(Direction ~ Lag2, data = Weekly,
family = binomial, subset = datos.entrenamiento)
summary(modelo.log.s)
```

```
##
## Call:
## glm(formula = Direction ~ Lag2, family = binomial, data = Weekly,
##     subset = datos.entrenamiento)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.536  -1.264   1.021   1.091   1.368
##
## 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
```

Como podemos observar, tanto el intercepto y Lag2 rechazan la hipótesis inicial, por lo que ambos son significativos para el modelo. El modelo anterior tenía un AIC: 1502.2, mientras que este modelo tiene un AIC: 1354.5, un AIC que ha disminuido bastante. Por lo tanto, la variable Lag2 sí tiene mayor influencia dentro del modelo # 5. Representa gráficamente el modelo

```r
#Vector con nuevos valores interpolados en el rango del predictor Lag2:
nuevos_puntos <- seq(from = min(Weekly$Lag2), to = max(Weekly$Lag2),
by = 0.5)
# Predicción de los nuevos puntos según el modelo con el comando predict() se calcula la probabilidad d
# referencia (en este caso "Up")
predicciones <- predict(modelo.log.s, newdata = data.frame(Lag2 =
nuevos_puntos),se.fit = TRUE, type = "response")
# Límites del intervalo de confianza (95%) de las predicciones
CI_inferior <- predicciones$fit - 1.96 * predicciones$se.fit
```

```r
CI_superior <- predicciones$fit + 1.96 * predicciones$se.fit
# Matriz de datos con los nuevos puntos y sus predicciones
datos_curva <- data.frame(Lag2 = nuevos_puntos, probabilidad =
predicciones$fit, CI.inferior = CI_inferior, CI.superior = CI_superior)
# Codificación 0,1 de la variable respuesta Direction
Weekly$Direction <- ifelse(Weekly$Direction == "Down", yes = 0, no = 1)
ggplot(Weekly, aes(x = Lag2, y = Direction)) +
geom_point(aes(color = as.factor(Direction)), shape = "I", size = 3) +
geom_line(data = datos_curva, aes(y = probabilidad), color = "firebrick") +
geom_line(data = datos_curva, aes(y = CI.superior), linetype = "dashed") +
geom_line(data = datos_curva, 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()
```



En la gráfica anterior se muestra el ajuste del modelo de regresión logística. A medida que el valor de Lag2 aumenta, también aumenta la probabilidad de que Direction sea "Up". La línea central roja representa esta probabilidad estimada, y las líneas punteadas son intervalos de confianza. Evaluaremos el modelo con chisq y matriz de confusión # 6. Evalúa el modelo con las pruebas de verificación correspondientes (Prueba de chi cuadrada, matriz de confusión). H_0: La variable Lag2 no mejora el modelo ($\beta_i = 0$) H_1: La variable Lag2 mejora el modelo de manera significativa ($\beta_i \neq 0$)

```r
# Chi cuadrada: Se evalúa la significancia del modelo con predictores con respecto al
anova(modelo.log.s, test ='Chisq')
```

```
## 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
```

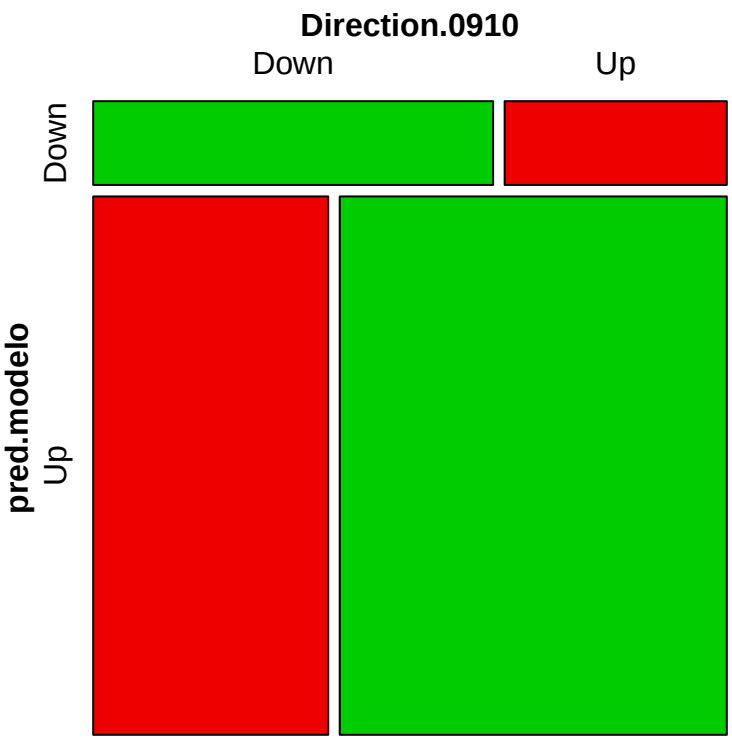
```r
# Cálculo de la probabilidad predicha por el modelo con los datos de test
prob.modelo <- predict(modelo.log.s, newdata = datos.test, type = "response")
# Vector de elementos "Down"
pred.modelo <- rep("Down", length(prob.modelo))
# Sustitución de "Down" por "Up" si la p > 0.5
pred.modelo[prob.modelo > 0.5] <- "Up"
Direction.0910 = Direction[!datos.entrenamiento]
# Matriz de confusión
matriz.confusion <- table(pred.modelo, Direction.0910)
library(vcd)
```

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

```
##
## Attaching package: 'vcd'
```

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

```r
mosaic(matriz.confusion, shade = T, colorize = T,
gp = gpar(fill = matrix(c("green3", "red2", "red2", "green3"), 2, 2)))
```




```
mean(pred.modelo == Direction.0910)
```

```
## [1] 0.625
```

Como podemos observar, nuestro valor p es menor a 0.05, por lo que la hipótesis inicial se rechaza y concluimos que Lag2 sí es significativa para nuestro modelo (es decir, para predecir direction. Además, vemos que la deviación de los residuos disminuye en 4 puntos al implementar Lag2 vs el modelo sin predictores.

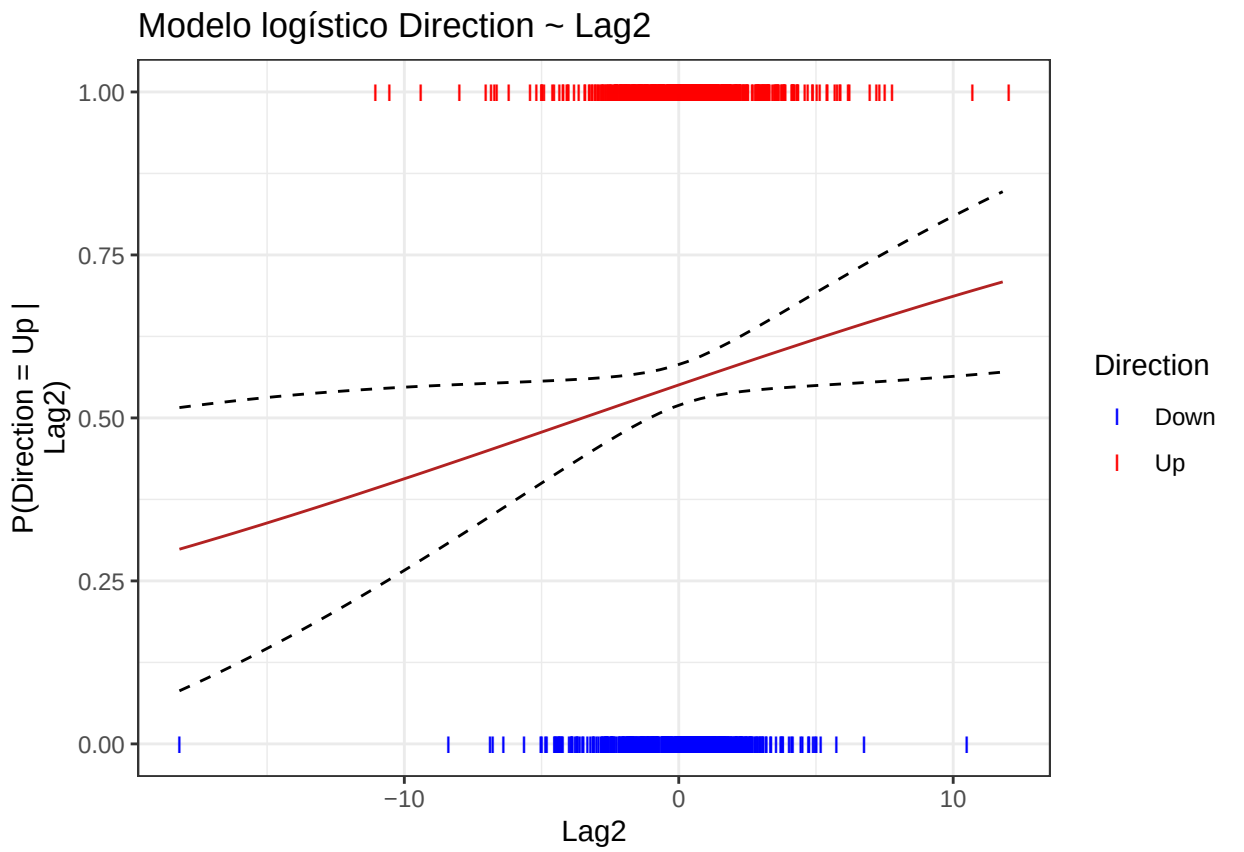
Por otro lado, en la matriz de confusión se muestran los resultados correctos de color verde, mientras los resultados incorrectos de color rojo. Observamos cómo la mayoría de las observaciones "Abajo" fueron clasificadas correctamente, pero el modelo cometió algunos errores en la predicción de "Arriba". El modelo tuvo un 62.5% de exactitud en la clasificación de Direction para los años 2009 y 2010. # 7. Escribe (ecuación), grafica el modelo significativo e interprétalo en el contexto del problema. Añade posibles es buen modelo, en qué no lo es, cuánto

La ecuación de nuestro modelo es: $\log\left(\frac{P(\text{Direction=Up})}{P(\text{Direction=Down})}\right) = 0.20326 + 0.05810 \cdot \text{Lag2}$.

Esto significa que, por cada unidad de incremento en Lag2, los momios de que el mercado suba ('Arriba') aumentan en un factor de $e^{0.05810} \approx 1.06$, lo cual equivale a aproximadamente un incremento del 6%.

```
ggplot(Weekly, aes(x = Lag2, y = Direction)) +
geom_point(aes(color = as.factor(Direction)), shape = "I", size = 3) +
geom_line(data = datos_curva, aes(y = probabilidad), color = "firebrick") +
geom_line(data = datos_curva, aes(y = CI.superior), linetype = "dashed") +
geom_line(data = datos_curva, 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()
```



Como se había mencionado anteriormente, en la gráfica se muestra que a medida que el valor de Lag2 aumenta, también aumenta la probabilidad de que Direction sea "Up". Significado del Modelo: Este modelo sugiere que un mayor rendimiento en la semana anterior (Lag2) incrementa ligeramente la probabilidad de que el mercado suba la semana siguiente. Moderado Impacto de Lag2: El impacto de Lag2 es pequeño (incremento de 6% en los odds por cada unidad adicional), lo que indica que Lag2 no es un predictor fuerte por sí solo. Esto se ve en la exactitud del modelo (62.5%), lo cual no es tan bueno.

En conclusión, este modelo logístico básico proporciona una idea inicial de cómo el rendimiento de dos pasos atrás (Lag2) puede influir en el mercado, pero el modelo tiene una precisión limitada, lo que sugiere que hay otros factores importantes no capturados por Lag2. Debido a lo anterior, es crucial explorar modelos más complejos para lograr mejores resultados predictivos.