A5-Regresión logística

Carlos David Lozano Sanguino - A01275316

2023-10-19

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 rendimiento (positivo o negativo) dependiendo del comportamiento previo de diversas variables de la bolsa bursátil S&P 500.

```
library(ISLR)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
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
```

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:

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

```
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
```

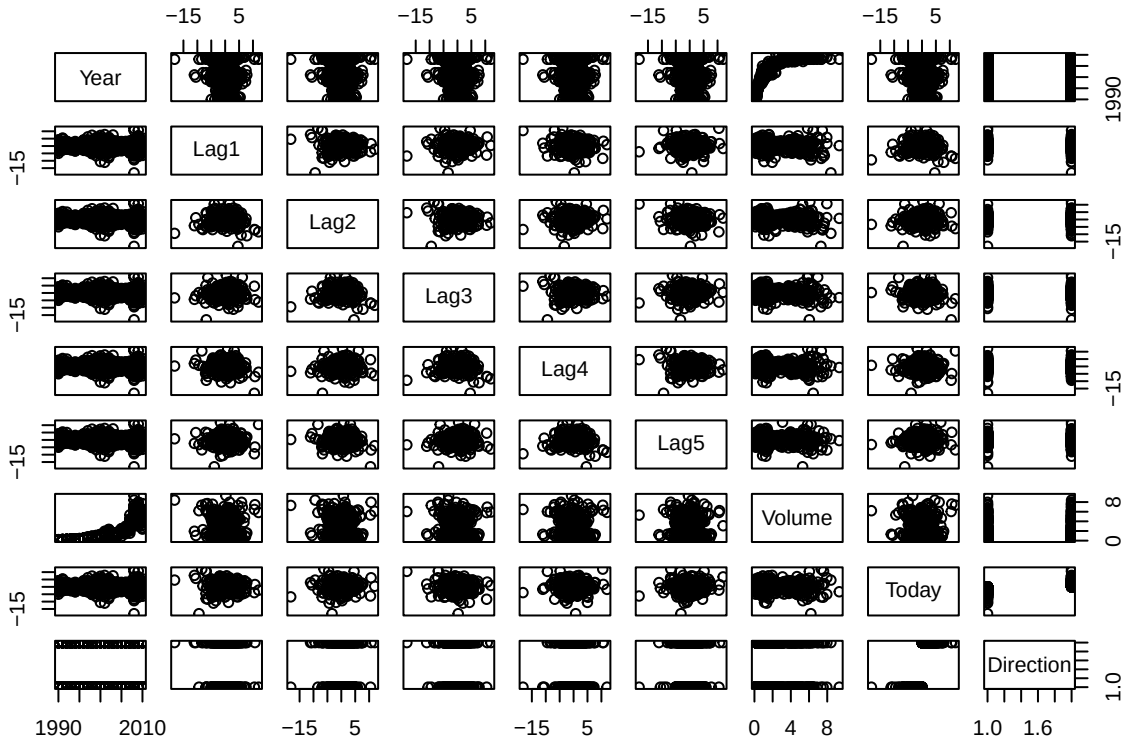
```
glimpse(Weekly)
```

```
## 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
##
##
##
##
```

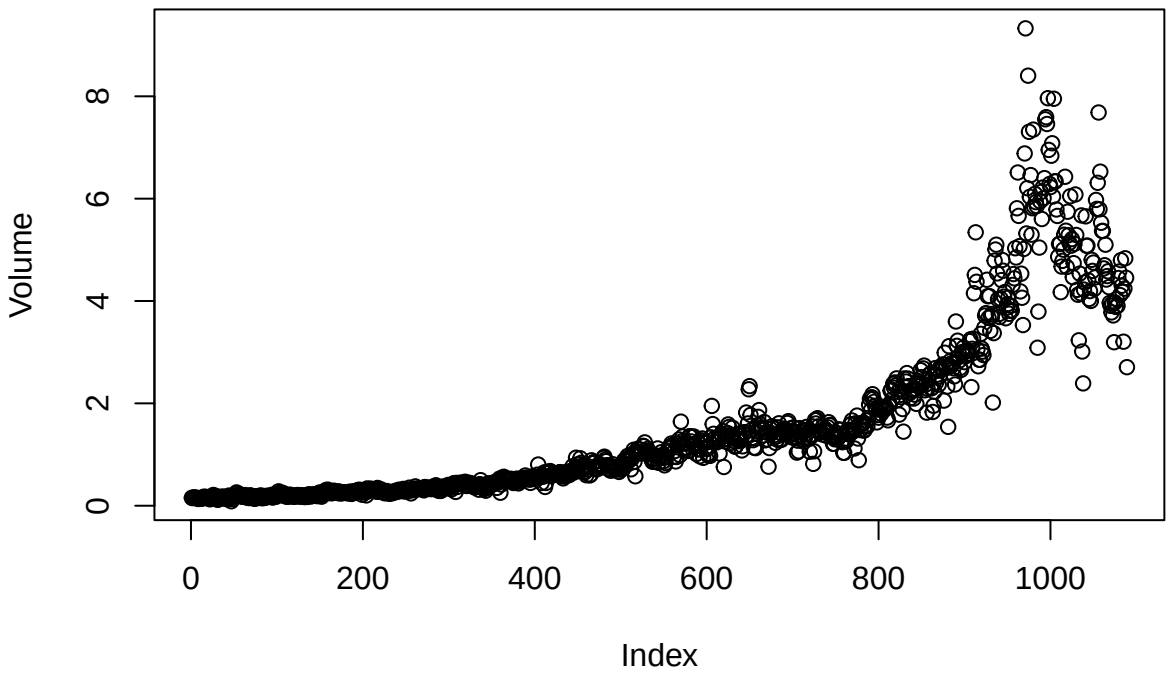
```
pairs(Weekly)
```




```
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)
```



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

Modelo con todos los predictores, excluyendo "Today", obtenemos el modelo con todas las variables en la que se puede apreciar que las variables que no influyen son todas aquellas que no tengan asterisco junto a datos como el error estandar, valor z entre otras, al final la unica variable que influye es Lag2 y tiene un efecto positivo en la dirección "Up" mientras que las demas no influyen.

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

```
##
## Call:
## glm(formula = Direction ~ . - Today, family = binomial, data = Weekly)
##
## 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)
```

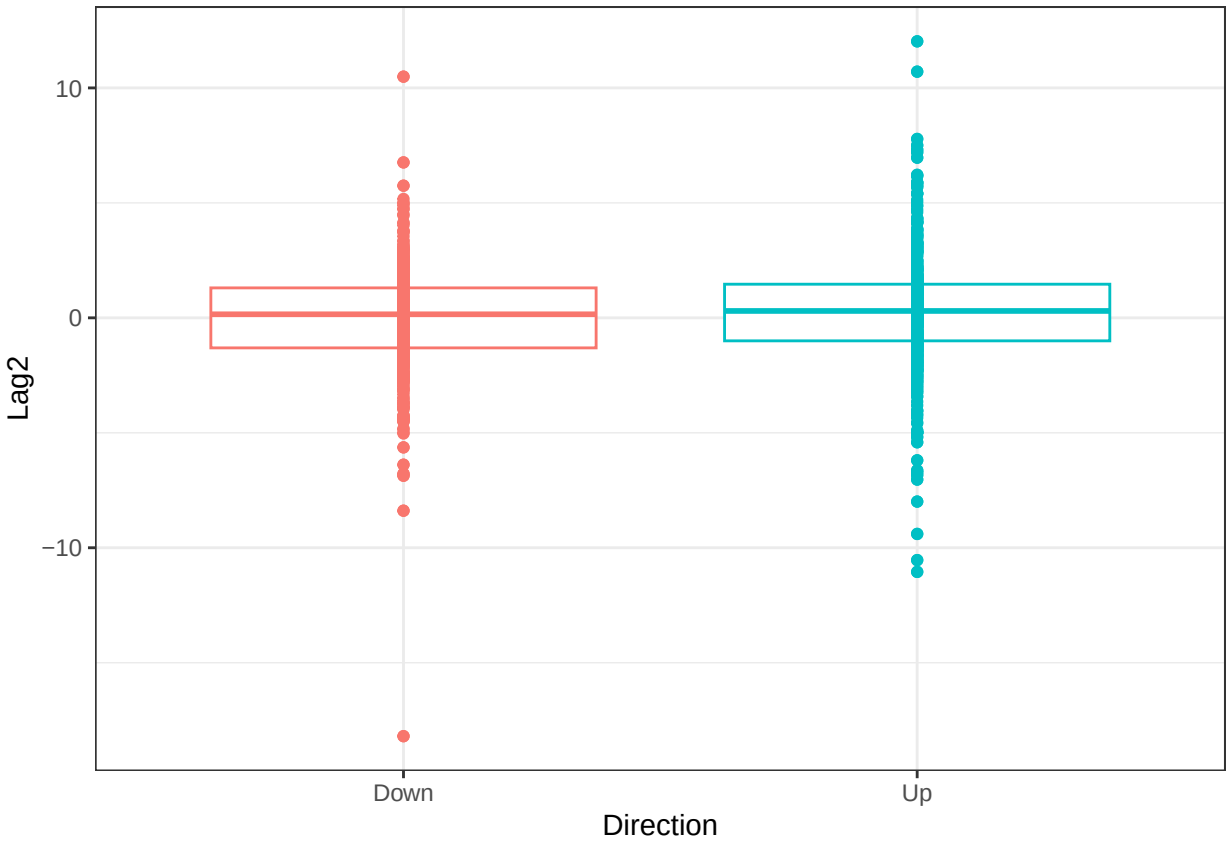
```
## Waiting for profiling to be done...
```

```
##                     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
```

Gráfico de las variables significativas (boxplot), ejemplo: Lag2):

Posteriormente se hizo la boxplot para obtener los intervalos de confianza para las betha o variables significativas que seria solo Lag2

```
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")
```



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.

El Intercept (Intercepción) tiene un valor de 0.20326, lo que significa que el logaritmo de odds de la dirección es positivo cuando Lag2 es igual a cero. Además, es significativo con un p-valor de 0.00157. Lag2 tiene un coeficiente de 0.05810, lo que significa que un aumento de una unidad en Lag2 aumenta los logaritmos de odds en un 5.81%. Es significativo con un p-valor de 0.04298. En resumen, el valor de Lag2 tiene un efecto significativo en la dirección del evento, y un aumento en Lag2 se asocia con un mayor logaritmo de odds de que la dirección sea "Up".

```
# 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)
```

```
# 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)
```

```
##
## 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
```

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

El modelo devuelve las predicciones del logarítmo de Odds. La predicción sedebe convertir en probabilidad. Esto se logra con el comando 'predict' y el 'type="response"' a continuacion:

Se crea un vector nuevos_puntos que contiene nuevos valores interpolados en el rango del predictor Lag2. Los valores se generan utilizando la función seq(). El from y to indican el rango mínimo y máximo de los valores de Lag2 en el conjunto de datos Weekly. El argumento by establece el incremento entre los valores interpolados, que en este caso es 0.5. Luego, utiliza el comando predict() para realizar predicciones en estos nuevos puntos. Estas predicciones se realizan utilizando un modelo de regresión logística binaria (modelo.log.s) previamente ajustado. El argumento newdata se utiliza para proporcionar los nuevos puntos en los que se desea realizar las predicciones. En este caso, se crea un nuevo conjunto de datos con una sola columna llamada "Lag2" que contiene los valores interpolados.

El argumento se.fit se establece en TRUE, lo que significa que se calcularán los errores estándar de las predicciones. El argumento type se establece en "response", lo que significa que se calcularán las probabilidades de que la variable respuesta pertenezca al nivel de referencia, que en este caso es "Up" en el modelo de regresión logística. En resumen, este código genera una serie de nuevos puntos dentro del rango de valores observados para el predictor Lag2 y utiliza el modelo de regresión logística para predecir las probabilidades de que la variable de respuesta sea "Up" en estos nuevos puntos. Las predicciones se realizan junto con sus errores estándar.

```
# 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
predicciones <- predict(modelo.log.s, newdata = data.frame(Lag2 =
nuevos_puntos),se.fit = TRUE, type = "response")
```

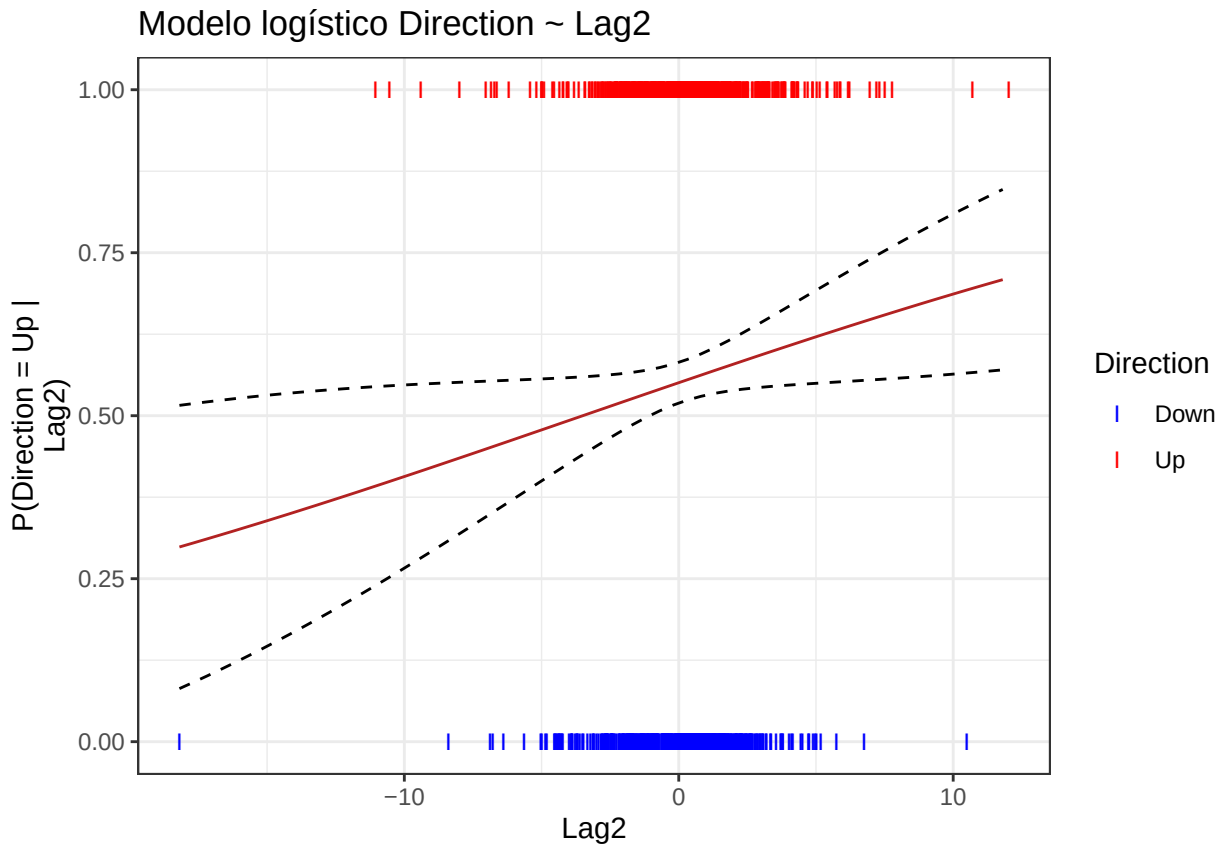
5. Representa gráficamente el modelo:

Primero se crean los limites de los intervalos de confianza para posteriormente graficarlos

```
# Límites del intervalo de confianza (95%) de las predicciones
CI_inferior <- predicciones$fit - 1.96 * predicciones$se.fit
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)
```

Con lo anterior esstablecodp se usa los diferentes elementos como ggplot,geom_point,geom_line, etc para obtener la curva de probabilidad de el modelo logístico y los intervalos de confianza asociados. El gráfico es de útilidad para visualizar cómo el modelo logístico se ajusta a los datos y cómo las probabilidades de "Up" varían con "Lag2".

```
# 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()
```



6. Evalúa el modelo con las pruebas de verificación correspondientes (Prueba de chi cuadrada, matriz de confusión).

Chi cuadrada: Se evalúa la significancia del modelo con predictores con respecto al modelo nulo ("Residual deviance" vs "Null deviance"). Si valor p es menor que alfa será significativo.

La tabla de Análisis de Devianza se utiliza para evaluar la contribución estadísticamente significativa de un predictor en un modelo de regresión logística binaria. El modelo es "binomial" con una función de enlace

logit, y la variable de respuesta se define como antes que seria “Direction.”

En la fila “NULL,” no se incluye ningún predictor en el modelo, lo que significa que se parte de un modelo sin variables predictoras (intercepto solo). El valor de “Df” (grados de libertad) para este modelo es 984, y la devianza residual (Resid. Dev) es 1354.7.

En la siguiente fila, se agrega el predictor “Lag2” al modelo. El valor de “Df” para este predictor es 1, lo que significa que se agrega un grado de libertad al modelo. La devianza residual después de agregar “Lag2” se reduce a 1350.5.

El valor “0.04123” en la columna “Pr(>Chi)” indica el p-valor asociado con la comparación entre el modelo nulo (sin predictor) y el modelo con “Lag2.” Este p-valor sugiere que la adición de “Lag2” al modelo es estadísticamente significativa. Como el p-valor (0.04123) es menor que el nivel de significancia comúnmente utilizado, como 0.05, se considera que la variable “Lag2” tiene un efecto estadísticamente significativo en el modelo.

Los asteriscos () junto al p-valor indican el nivel de significancia. En este caso, se muestra un asterisco (), lo que indica que el p-valor es menor que 0.05, lo que es comúnmente aceptado como un nivel de significancia.

Por lo tanto, se muestra que la adición del predictor “Lag2” al modelo es estadísticamente significativa, lo que significa que “Lag2” contribuye de manera significativa a la explicación de la variabilidad en la variable de respuesta “Direction” en el modelo de regresión logística binaria.

```
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
```

Cálculo de las predicciones correctas así como de los falsos negativos y positivos. Normalmente se usa un límite de 0.5.

```
# Cálculo de la probabilidad predicha por el modelo con los datos de test
library(vcd)
```

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

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

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

```
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"
```

```r
Direction.0910 = Direction[!datos.entrenamiento]
# Matriz de confusión
matriz.confusion <- table(pred.modelo, Direction.0910)
matriz.confusion
```

```
##            Direction.0910
## pred.modelo Down Up
##        Down    9  5
##        Up     34 56
```

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


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

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

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 cambia)

Se puede concluir e interpretar lo siguiente: Dentro del contexto del problema, donde se trabaja con el conjunto de datos "Weekly" que contiene información sobre el rendimiento semanal del índice bursátil S&P 500, los resultados indican que el único predictor relevante en el modelo de regresión logística es "Lag2". Esto se puede inferir del análisis de los coeficientes y los p-valores en el modelo. Los otros predictores, como "Year," "Lag1," "Lag3," "Lag4," "Lag5," y "Volume," no muestran una influencia estadísticamente significativa en la dirección del mercado.

En este contexto, el resultado más relevante es la significancia de "Lag2" en la dirección del mercado. El

coeficiente positivo para “Lag2” sugiere que un aumento en “Lag2” está asociado con un aumento en la probabilidad de que la dirección del mercado sea “Up.” Esto significa que el rendimiento del mercado dos semanas atrás (Lag2) parece influir en la dirección del mercado actual.

Por lo tanto, “Lag2” podría ser un buen predictor para predecir la dirección del mercado, al menos en este modelo. Sin embargo, para determinar si este es un “buen” modelo, se deben considerar otros factores, como la precisión general del modelo en términos de métricas de evaluación de clasificación, como la precisión, la sensibilidad y la especificidad.

Para determinar si el modelo es adecuado, se deben realizar evaluaciones adicionales, como la validación cruzada y el análisis de métricas de desempeño, para determinar cuán bien el modelo puede generalizarse a nuevos datos. También es importante considerar el contexto y el propósito de la predicción. En el ámbito financiero, las predicciones precisas son cruciales, y la elección de variables predictoras y modelos adecuados puede requerir más análisis y refinamiento.

En resumen, en el contexto del problema, “Lag2” parece ser un predictor relevante para predecir la dirección del mercado, pero se necesitan evaluaciones adicionales y consideraciones para determinar si este es un “buen” modelo en términos de precisión y generalización a nuevos datos.