19. Regresión logistica

Jacobo Hirsch Rodriguez

2024-11-05

Lectura de los datos

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

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## 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
```

Análisis de los datos

vemos como esta estructurado el conjunto de datos weekly del paquete ISLR

```
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
```

usamos el comando glimpse de tidyverse para visualizar una vista general del conjunto de datos

```
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~
```

vemos el resumen estadístico

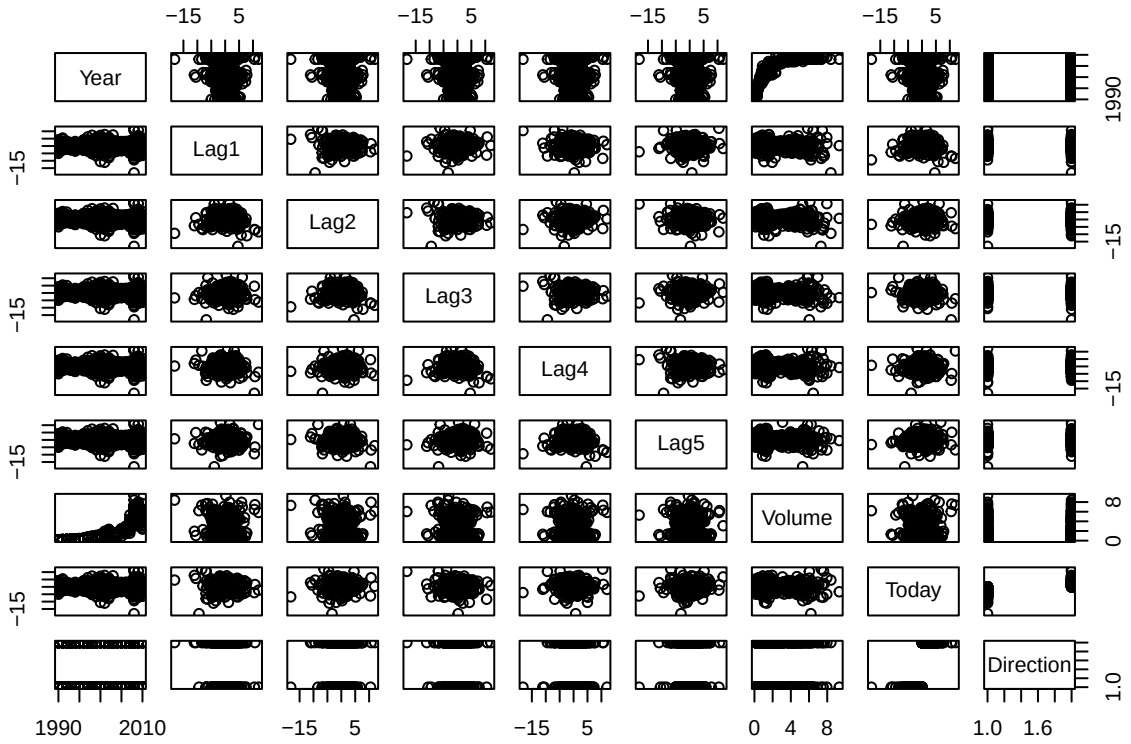
```
summary(Weekly)
```

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

La mayoría de las variables Lag1 a Lag5, así como Today, tienen distribuciones similares y reflejan los cambios porcentuales en el rendimiento del mercado en semanas anteriores. Volume muestra la variabilidad en el volumen de transacciones, mientras que Direction es una variable categórica que indica la dirección general del mercado.

genera un gráfico de pares, mostrando diagramas de dispersión entre todas las variables numéricas de Weekly

```
pairs(Weekly)
```

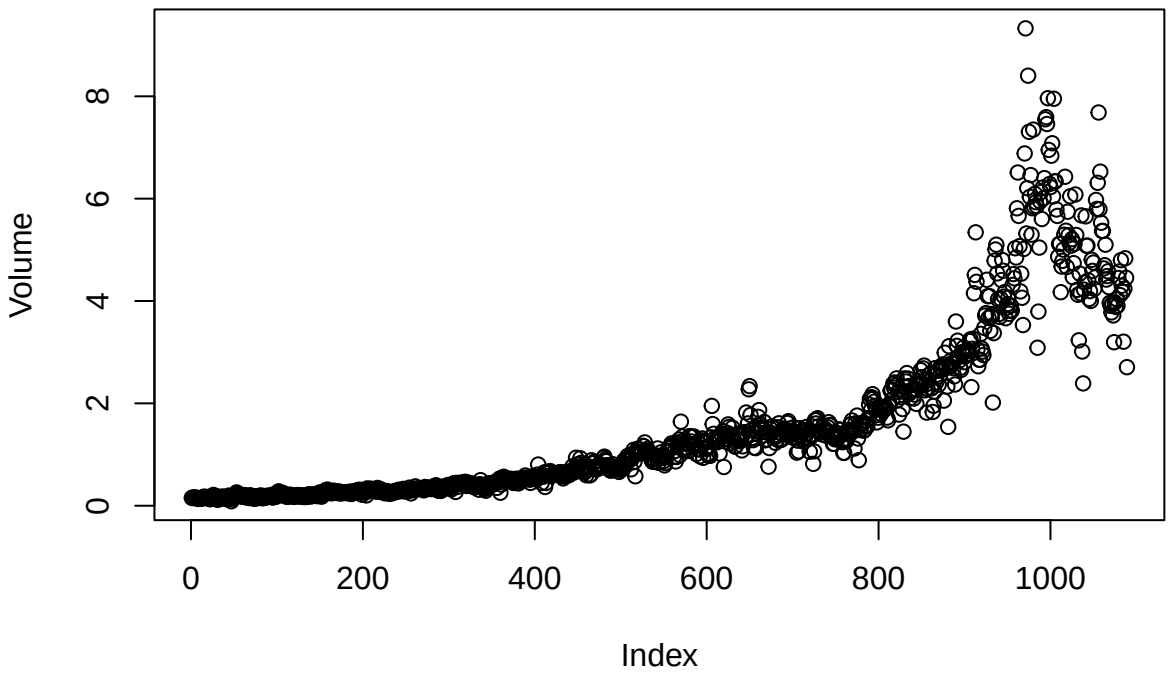


se calcula la matriz de correlación excluyendo la columna direction ya que esta es una variable categorica

```
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)
```



Esta gráfica muestra el numero de transacciones que se hicieron durante todas las semanas del año 1900 al año 2010, podemos observar que la cantidad de transacciones fue en aumento constante y tuvo una caída visible sin tendencia a recuperarse.

#Cálculo del modelo logistico

obtenemos un modelo lineal generalizado escogiendo a la variable direction como la dependiente, con todas las variables como las explicativas a excepción de la variable today se escoge la familia binomial ya que el problema es de clasificación binaria

escribimos el código para obtener el modelo de regresión logistica

```
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
```

log2 es el único predictor con un valor p menor a 0.05 (valor p = 0.0275). Esto indica que Lag2 es estadísticamente significativo y su coeficiente positivo (0.059449) sugiere que un incremento en el rendimiento de dos semanas atrás está asociado con una mayor probabilidad de que el mercado suba.

La función contrasts muestra cómo se codifica la variable de respuesta Direction en el modelo.

```
contrasts(Direction)
```

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

vemos que en el modelo se codifico down como cero y up como 1, que servirá al momento de interpretar el modelo en el futuro por que este modelo no será utilizado

calculamos los intervalos de confianza para los coeficientes del modelo. En este caso, estamos calculando intervalos de confianza al 95% para cada coeficiente en el modelo modelo.log.m

```
confint(object = modelo.log.m, level = 0.95)
```

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

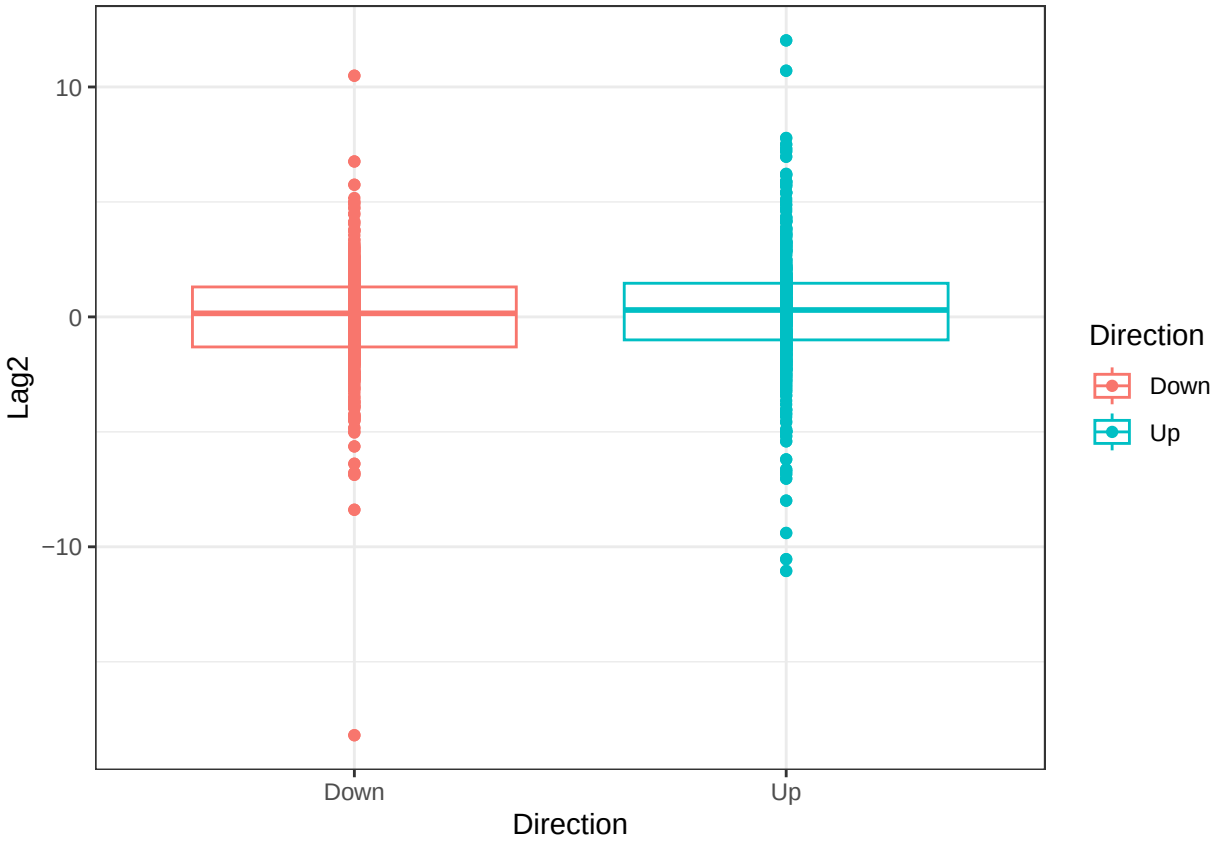
```
##                     2.5 %      97.5 %
## (Intercept) -56.985558236 91.66680901
## Year         -0.045809580  0.02869546
## Lag1         -0.092972584  0.01093101
## Lag2          0.007001418  0.11291264
## Lag3         -0.068140141  0.03671410
## Lag4         -0.079519582  0.02453326
## Lag5  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

el intervalo de confianza del intercepto es demasiado amplio por lo que no es muy preciso, además de que incluye el 0 por lo que no es estadisticamente significativo. Las demás variables tienen coeficientes con intervalos de confianza que incluyen el cero por lo que se vuelve a indicar su nula significancia en el modelo. El único valor que fue significativo (al igual que vimos en el summary) es lag2.

#Modelo de regresión logistica con variables significativas

vamos a gráficar como se comportó la variable independiente de nuestro modelo las semanas que el mercado estuvo arriba y que estuvo abajo, hacemos esto con un boxplot

```r
library(ggplot2)

# Gráfico
ggplot(data = Weekly, mapping = aes(x = Direction, y = Lag2)) + #hacemos un plano donde especificamos q
  geom_boxplot(aes(color = Direction)) +  #se crea un boxplot en donde se escoge el color automaticamen
  geom_point(aes(color = Direction)) + #agrega puntos individuales al gráfico, representando cada obser
  theme_bw()  #Aplica un tema de fondo blanco y líneas en negro al gráfico
```



Las distribuciones de Lag2 para Direction = “Down” y Direction = “Up” son muy similares. Ambas categorías tienen una mediana cercana a 0, lo que indica que no hay una diferencia clara en el valor central de Lag2 entre las semanas en las que el mercado sube y baja.

los boxplots son casi simetricos, esto puede indicar que la variable log2 podría no ser un buen indicador para saber si el mercado sube o baja, aún así lo utilizaremos para el siguiente modelo por que resultó ser el más significativo de las demás variables.

##Creación del modelo con variable significativa

```r
# Dividir el conjunto de datos en entrenamiento y prueba
datos.entrenamiento <- Weekly$Year < 2009 #seleccionamos todos los datos donde el año fue menor al 2009
datos.test <- Weekly[!datos.entrenamiento, ] #se hace un conjunto de prueba donde se seleccionan todos

# Verifica el número de filas en cada conjunto
cat("Número de observaciones en entrenamiento:", sum(datos.entrenamiento), "\n")
```

```
## Número de observaciones en entrenamiento: 985
```

```r
cat("Número de observaciones en test:", nrow(datos.test), "\n")
```

```
## Número de observaciones en test: 104
```

```r
# Ajuste del modelo logístico con Lag2 como variable significativa
modelo.log.s <- glm(Direction ~ Lag2, data = Weekly,
                    family = binomial, subset = datos.entrenamiento) #es importante aclarar que se util

# Resumen del modelo
summary(modelo.log.s)
```

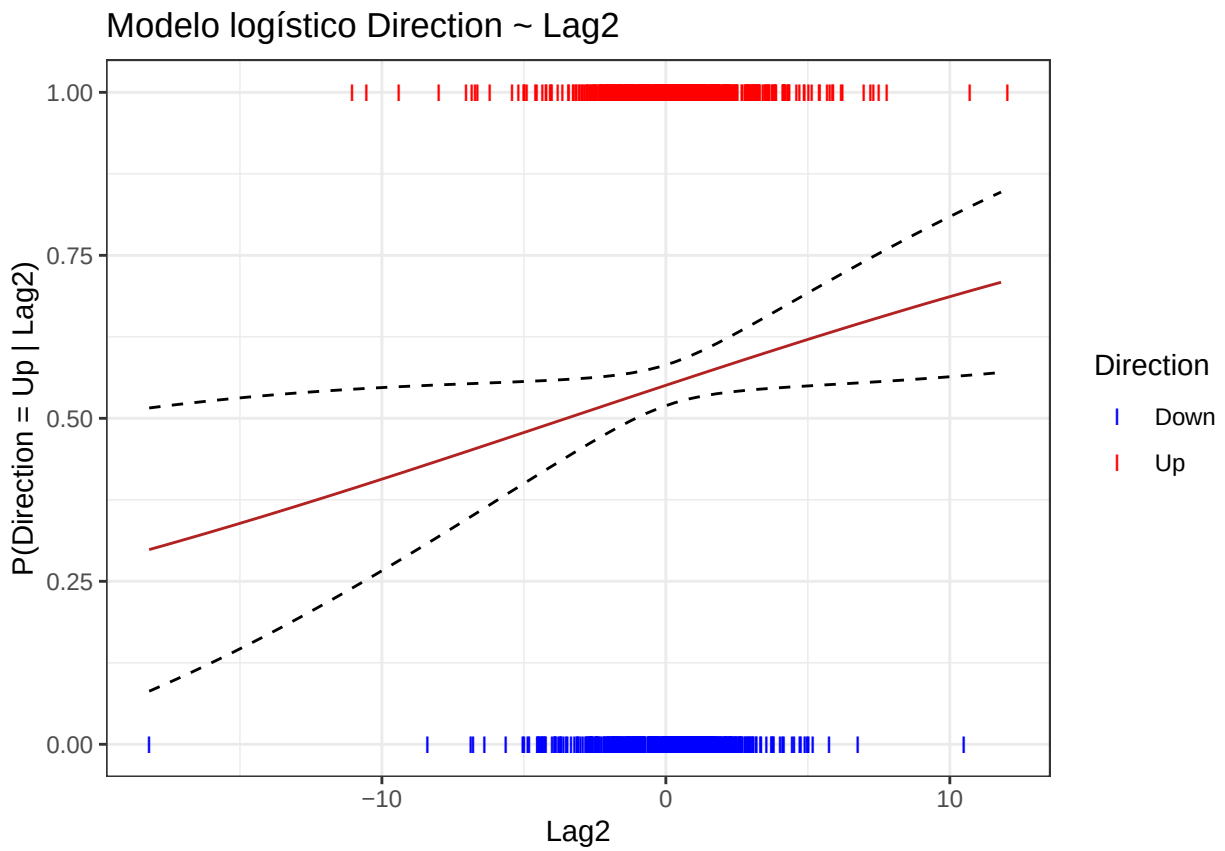
```
##
## Call:
## glm(formula = Direction ~ Lag2, family = binomial, data = Weekly,
##     subset = datos.entrenamiento)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.20326    0.06428   3.162  0.00157 **
## Lag2         0.05810    0.02870   2.024  0.04298 *
## ---
## 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
```

muestran que Lag2 es un predictor estadísticamente significativo ($p < 0.05$), lo que indica que es poco probable que su relación con Direction sea producto del azar. Sin embargo, su tamaño del efecto es pequeño, con un coeficiente de 0.058, lo cual implica que un aumento de una unidad en Lag2 incrementa las odds de que el mercado suba (Direction = "Up") en solo un 6%. Aunque significativo, el impacto de Lag2 es modesto, sugiriendo que, por sí solo, no tiene un gran poder predictivo.

#Representación grafica del modelo

```r
# Vector con nuevos valores interpolados en el rango del predictor Lag2:
# Este vector contiene valores de Lag2 desde su mínimo hasta su máximo, con un paso de 0.5.
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 de que la variable respuesta pertenezca al nivel de referencia (en este ca
predicciones <- predict(modelo.log.s,
                        newdata = data.frame(Lag2 = nuevos_puntos), # Crea un data frame con Lag2 = nue
                        se.fit = TRUE,          # Calcula el error estándar de las predicciones
                        type = "response")      # Retorna las predicciones como probabilidades (entre 0

# Límites del intervalo de confianza (95%) de las predicciones
# Calcula el intervalo de confianza del 95% para cada predicción usando el error estándar.
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
# Crea un data frame que contiene los valores interpolados de Lag2, las probabilidades predichas
# y los intervalos de confianza superior e inferior para cada valor de Lag2.
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
# Convierte la variable Direction en valores numéricos para el gráfico, donde "Down" = 0 y "Up" = 1.
Weekly$Direction <- ifelse(Weekly$Direction == "Down", yes = 0, no = 1)

# Crear el gráfico usando ggplot2
ggplot(Weekly, aes(x = Lag2, y = Direction)) +
  # Agrega puntos de las observaciones originales de Weekly, coloreados por Direction.
  geom_point(aes(color = as.factor(Direction)), shape = "I", size = 3) +

  # Agrega una línea con la curva de probabilidad predicha por el modelo
  geom_line(data = datos_curva, aes(y = probabilidad), color = "firebrick") +

  # Agrega las líneas de los límites superior e inferior del intervalo de confianza
  geom_line(data = datos_curva, aes(y = CI.superior), linetype = "dashed") +
  geom_line(data = datos_curva, aes(y = CI.inferior), linetype = "dashed") +

  # Etiquetas del gráfico
  labs(title = "Modelo logístico Direction ~ Lag2",
       y = "P(Direction = Up | Lag2)",
       x = "Lag2") +

  # Colores personalizados para los puntos de Direction: azul para "Down" y rojo para "Up"
  scale_color_manual(labels = c("Down", "Up"), values = c("blue", "red")) +

  # Título de la leyenda para indicar los niveles de Direction
  guides(color = guide_legend("Direction")) +

  # Centrado del título del gráfico
  theme(plot.title = element_text(hjust = 0.5)) +

  # Tema de fondo blanco y negro
  theme_bw()
```



La pendiente suave de la línea roja en el gráfico indica que Lag2 tiene una influencia leve sobre la probabilidad de Direction = "Up". Esto implica que, aunque Lag2 afecta la probabilidad predicha, su poder de discriminación entre up y down es bajo. Esta falta de pronunciación sugiere que Lag2 por sí solo no es suficiente para hacer una predicción clara de la dirección del mercado Los intervalos son más estrechos alrededor de valores centrales de Lag2, donde el modelo es más confiable, y más amplios en los extremos, indicando mayor incertidumbre en esos rangos. #Evaluación del modelo

evaluacion mediante chi cuadrada

```
# 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.
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
```

El análisis de deviance muestra que Lag2 es un predictor estadísticamente significativo para Direction, pero su contribución al modelo es moderada. Esto concuerda con el análisis previo, donde observamos que la pendiente de la curva de probabilidad no era muy pronunciada, lo cual también indicaba que el poder predictivo de Lag2 es limitado.

```
 #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)
matriz.confusion
```

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

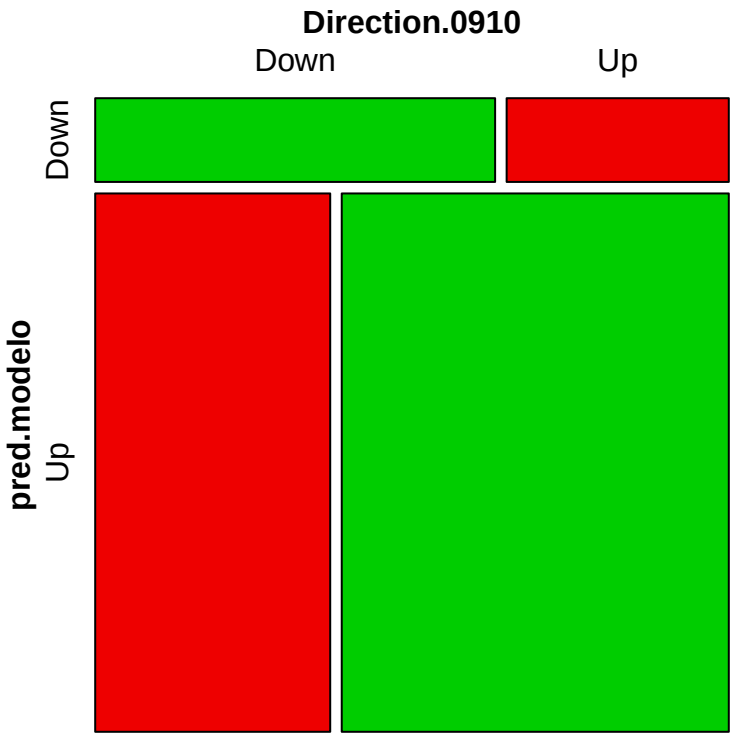
```
library(vcd)
```

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

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

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

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




```r
mean(pred.modelo == Direction.0910) #comparación de los datos reales con los predichos
```

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

De la matriz de confusión se obtuvo que:

9 observaciones fueron correctamente clasificadas como “Down” (predicción y valor real coinciden en “Down”). 56 observaciones fueron correctamente clasificadas como “Up” (predicción y valor real coinciden en “Up”). 5 observaciones fueron incorrectamente clasificadas como “Down” cuando realmente eran “Up” (falsos negativos). 34 observaciones fueron incorrectamente clasificadas como “Up” cuando realmente eran “Down” (falsos positivos).

En resumen, el modelo de regresión logística utilizando Lag2 tiene un rendimiento moderado con una exactitud de 62.5%. Aunque es capaz de predecir correctamente la clase “Up” en la mayoría de los casos, tiene problemas con las predicciones de “Down”, lo que se refleja en el alto número de falsos positivos. Esto sugiere que Lag2 no es un predictor suficientemente fuerte por sí solo, y que el modelo se beneficiaría de incluir más variables que claramento no se presentan en el dataset anteior, o ajustar el método para mejorar la precisión y reducir los errores de clasificación, siendo una idea de ajuste el balanceo de las clases para evitar sesgos en el modelo.

#Ecuación del modelo significativo

al final se obtuvo la siguiente ecuación: log(P(Direction = “up”) / (1 - P(Direction = “up”))) = 0.20326 + 0.05810 * Lag2

#Codigo para borrar todas las variables de ambiente

```r
#rm(list = ls()) #obvio esta comentado
```