# Regresion_Logistica

Arturo

2023-10-17

```
# Instalacion y carga de paquetes
if (!require(ISLR) || !require(tidyverse) || !require(caret)) {
  install.packages("ISLR")
  install.packages("tidyverse")
  install.packages("caret")
}
```

```
## Loading required package: ISLR
```

```
## Loading required package: tidyverse
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     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
## Loading required package: caret
##
## Loading required package: lattice
##
##
## Attaching package: 'caret'
##
##
## The following object is masked from 'package:purrr':
##
##     lift
```

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

**Nombre: Arturo Garza Campuzano**

**Matrícula: A00828096**

## 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.

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

```
rm(Weekly)
```

```
## Warning in rm(Weekly): object 'Weekly' not found
```

```
load(".RData")

# Cargamos el conjunto de datos 'Weekly' y mostramos las primeras filas
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 'glimpse' para obtener una descripcion mas detallada de 'Weekly'
glimpse(Weekly)
```

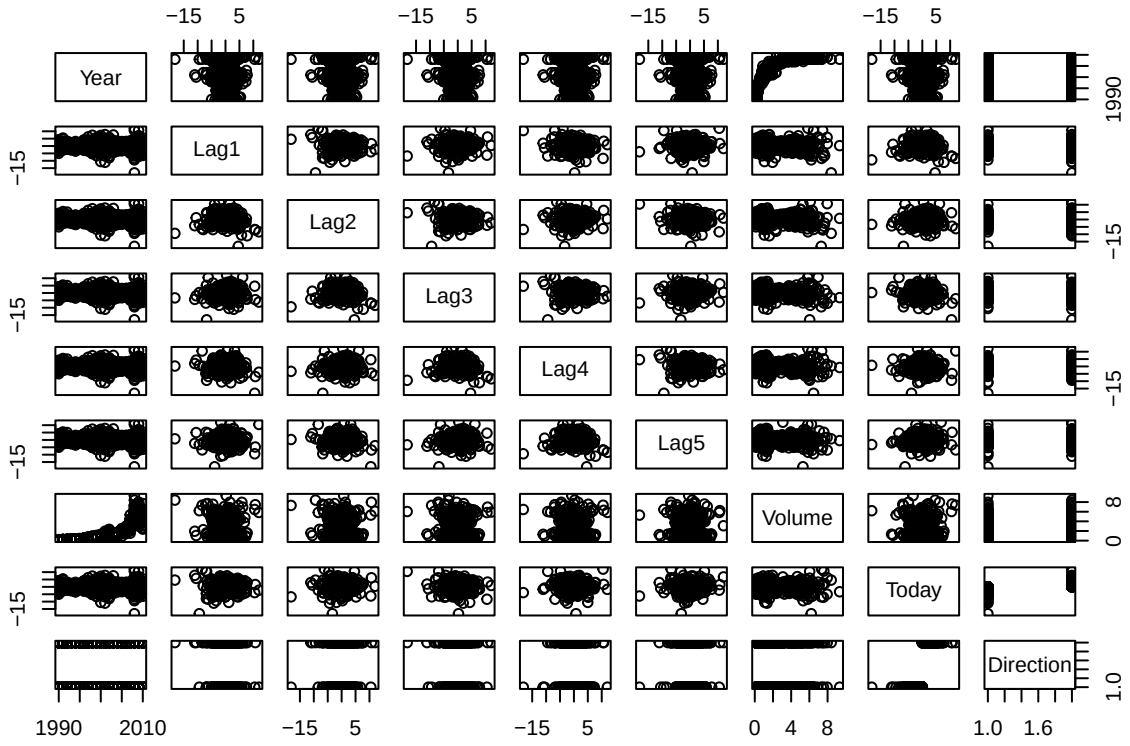
```
## Rows: 1,089
## Columns: 9
## $ Year      <dbl> 1990, 1990, 1990, 1990, 1990, 1990, 1990, 1990, 1990, 1990, ~
## $ Lag1      <dbl> 0.816, -0.270, -2.576, 3.514, 0.712, 1.178, -1.372, 0.807, 0~
## $ Lag2      <dbl> 1.572, 0.816, -0.270, -2.576, 3.514, 0.712, 1.178, -1.372, 0~
## $ Lag3      <dbl> -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~
```

```
# Resumimos estadisticamente el conjunto de datos 'Weekly'
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
##
##
##
##
```

```
# Creamos un grafico de pares para analizar las relaciones entre variables
pairs(Weekly)
```



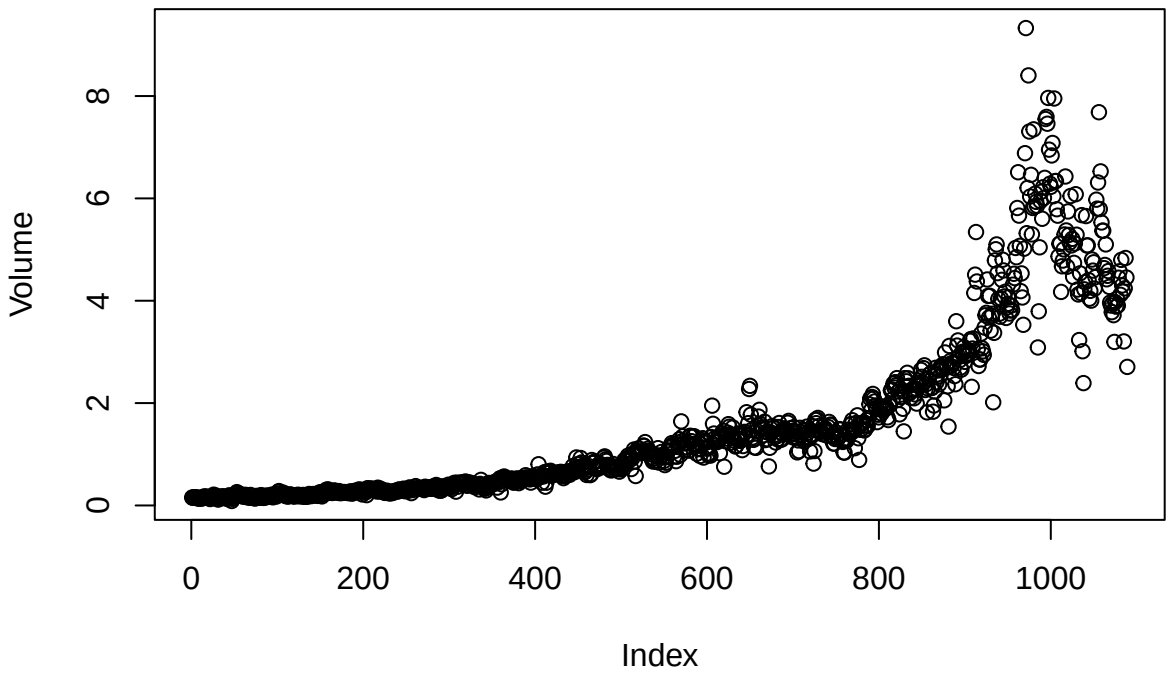

```
# Calculamos la matriz de correlacion de todas las columnas, excluyendo la ultima
cor(Weekly[, -9])
```

```
##                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
```

```
# Adjuntamos 'Weekly', lo que permite acceder a sus columnas sin especificar el nombre del conjunto de
attach(Weekly)

# Creamos un grafico de dispersion de la columna 'Volume'
plot(Volume)
```




Obervaciones:

- El set de datos Weekly cuenta con 1,089 registros. Nueve columnas (variables): Year, Lag1, Lag2, Lag3, Lag4, Lag5, Volume, Today y Direction.
- Las variables que cuentan con unidades similares son Lag1, Lag2, Lag3, Lag4, Lag5 y Today.
- Las variables cuyas unidades son muy diferentes son Year y Volume.
- Las variables que parecen correlacionarse más entre sí son Year y Volume, y Today y Direction.
- A medida que el índice de los registros aumenta el valor de la variable Volume aumenta.

**2. Formula un modelo logístico con todas las variables menos la variable "Today". Calcula los intervalos de confianza para las beta i. 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"
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
```

```
# Intervalos de confianza para las betas i
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
```

Observaciones:

- La única variable significativa parece ser *Lag2*.
- La mayoría de los estimadores son cercanos a cero.
- La devianza residual es ligeramente menor que la nula, lo cual indica que el modelo ha capturado una cantidad significativa de la variabilidad de los datos y, por tanto, propociona un mejor ajuste.

Interpretación: Considerando que el nivel de significancia es $\alpha = 0.05$ entonces la única variable que influye en el modelo es *Lag2*. El efecto de las variables *Intercept*, *Lag2* y *Volume* en los momios es positivo, lo cual representa cuánto cambian las probabilidades de dirección "Up" por un aumento de una unidad en las variables.

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

División de la base de datos en un conjunto de entrenamiento y prueba.

```
# Training: observaciones desde 1990 hasta 2008
conjunto_entrenamiento <- Weekly[Weekly$Year >= 1990 & Weekly$Year <= 2008, ]
datos.entrenamiento <- (Year < 2009)
```

```r
# Test: observaciones de 2009 y 2010
conjunto_prueba <- Weekly[Weekly$Year >= 2009 & Weekly$Year <= 2010, ]

# Verifica
nrow(conjunto_entrenamiento) + nrow(conjunto_prueba)
```

```
## [1] 1089
```

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

A continuación se presenta el modelo logístico con las variables significativas (Lag_2) en base al entrenamiento, ajustando el modelo encontrado.

```r
# Ajuste del modelo encontrado
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
```

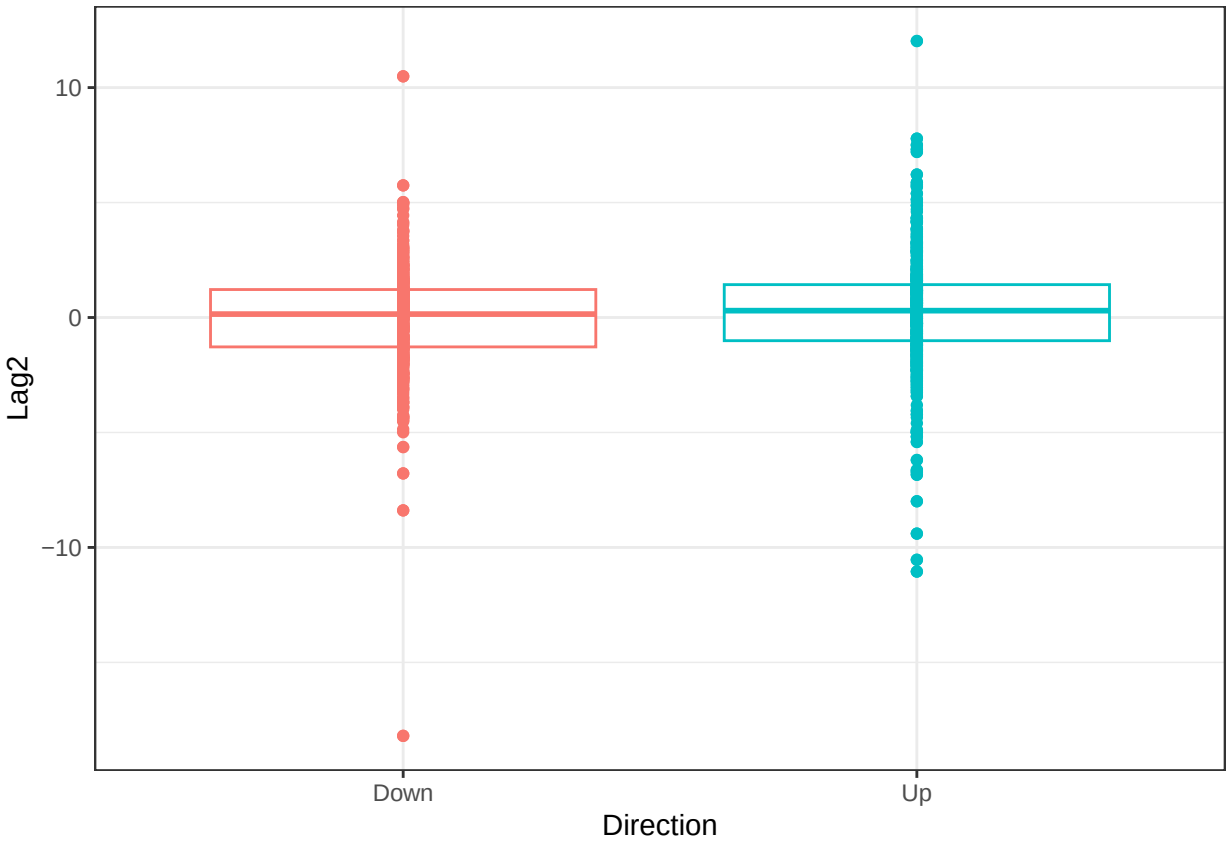
Observaciones:

- Todas las variables son significativas. *Intercept* es la más significa.
- Todos los estimadores son cercanos a cero.
- La devianza residual es ligeramente menor que la nula, lo cual indica que el modelo ha capturado una cantidad significativa de la variabilidad de los datos y, por tanto, propociona un mejor ajuste.

Interpretación: Considerando que el nivel de significancia es $\alpha = 0.05$ entonces todas las variables influyen en el modelo. El efecto de todas las variables en los momios es positivo, lo cual representa cuánto cambian las probabilidades de dirección "Up" por un aumento de una unidad en las variables.

**5. Representa gráficamente el modelo.**

```r
# Gráfico de las variables significativas (boxplot), ejemplo: Lag2):
ggplot(data = conjunto_entrenamiento, mapping = aes(x = Direction, y = Lag2)) +
geom_boxplot(aes(color = Direction)) +
geom_point(aes(color = Direction)) +
theme_bw() +
theme(legend.position = "null")
```




Observaciones:

- Los valores de la variable *Lag_2* que cuentan con una driección "Up" ocupan un mayor rango que los valores con dirección "Down".
- Las cajas de ambos valores se traslapan.

Interpretación: No hay una distinción significativa entre los registros de la variable Lag_2 considerando ambas direcciones.

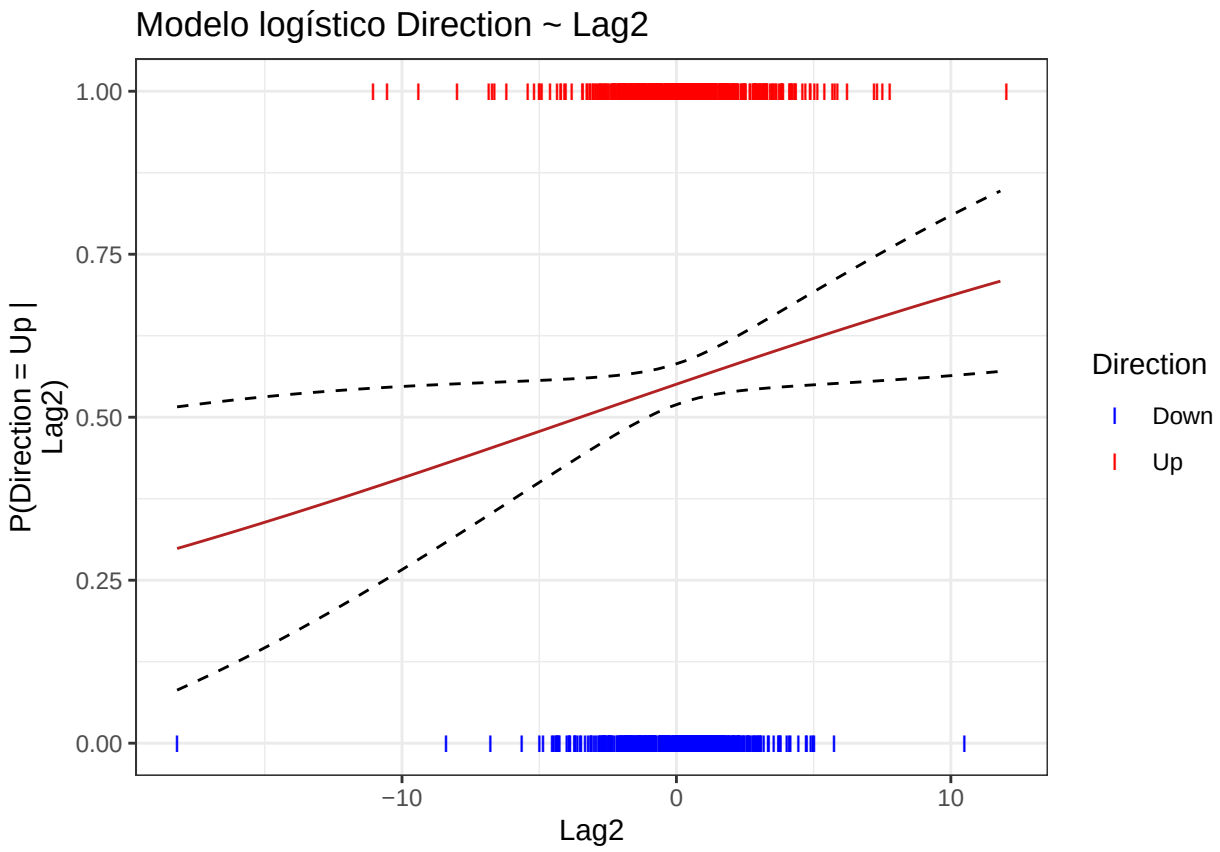
```
# Representación gráfica del modelo

# Vector con nuevos valores interpolados en el rango del predictor Lag2:
nuevos_puntos <- seq(from = min(conjunto_entrenamiento$Lag2), to = max(conjunto_entrenamiento$Lag2), by
# 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 =
```

```
# 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_infer
```

```
# Codificación 0,1 de la variable respuesta Direction
conjunto_entrenamiento$Direction <- ifelse(conjunto_entrenamiento$Direction == "Down", yes = 0, no = 1)
ggplot(conjunto_entrenamiento, 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()
```




Observaciones:

- Las observaciones "Up" cuentan con un mayor rango de valores que las observaciones "Down", en términos de la variable Lag_2.
- La línea roja representa la curva de probabilidades del modelo logístico en función de *Lag_2*; aunque, en este caso no sigue la forma característica de "S" de un modelo logístico.
- Las líneas punteadas son los intervalos de confianza al 95% de las predicciones; estas cuentan con una ligera curvatura.

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

Evaluación del modelo con la prueba de chi cuadrada

```
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
```

Interpretación: El análisis sugiere que Lag2 es un predictor significativo para determinar la Dirección según el modelo logístico ajustado. El valor de p es pequeño, lo cual indica que *Lag_2* esta relacionada con la probabilidad de Dirección = "Up".

*Evaluación del modelo con la matriz de confusión: Entrenamiento*

```
predicciones <- predict(modelo.log.s, newdata = data.frame(Lag2 = conjunto_entrenamiento$Lag2),se.fit =
valores_redondeados <- ifelse(predicciones <= 0.5, 0, 1)

# Asegurémonos de que los dos vectores sean factores con los mismos niveles
valores_redondeados <- factor(valores_redondeados, levels = c(0, 1))
conjunto_entrenamiento$Direction <- factor(conjunto_entrenamiento$Direction, levels = c(0, 1))

# Crea la matriz de confusión
confusion_matrix <- confusionMatrix(valores_redondeados, conjunto_entrenamiento$Direction)
confusion_matrix
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0  23  20
##          1 418 524
##
##                Accuracy : 0.5553
##                  95% CI : (0.5237, 0.5867)
##     No Information Rate : 0.5523
##     P-Value [Acc > NIR] : 0.4368
##
##                   Kappa : 0.0168
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.05215
##             Specificity : 0.96324
##          Pos Pred Value : 0.53488
##          Neg Pred Value : 0.55626
##              Prevalence : 0.44772
##          Detection Rate : 0.02335
##    Detection Prevalence : 0.04365
##       Balanced Accuracy : 0.50769
##
##        'Positive' Class : 0
##
```

Observaciones:

- El modelo clasifica correctamente le 55.5% de las muestras, es prácticamente equivalente al azar.

- El kappa indica un pobre desempeño.
- La sensibilidad para la clase positiva es muy baja.

Interpretación: Las métricas muestran que el modelo tiene un pobre desempeño de clasificación, especialmente para detectar la clase positiva.

*Evaluación del modelo con la matriz de confusión: Prueba*

```
conjunto_prueba$Direction <- ifelse(conjunto_prueba$Direction == "Down", 1, 0)

predicciones <- predict(modelo.log.s, newdata = data.frame(Lag2 = conjunto_prueba$Lag2),se.fit = FALSE,
valores_redondeados <- ifelse(predicciones <= 0.5, 0, 1)

# Asegurémonos de que los dos vectores sean factores con los mismos niveles
valores_redondeados <- factor(valores_redondeados, levels = c(0, 1))
conjunto_prueba$Direction <- factor(conjunto_prueba$Direction, levels = c(0, 1))

# Crea la matriz de confusión
confusion_matrix <- confusionMatrix(valores_redondeados, conjunto_prueba$Direction)
confusion_matrix
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0  5  9
##          1 56 34
##
##                Accuracy : 0.375
##                  95% CI : (0.282, 0.4753)
##     No Information Rate : 0.5865
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : -0.1097
##
##  Mcnemar's Test P-Value : 1.159e-08
##
##             Sensitivity : 0.08197
##             Specificity : 0.79070
##          Pos Pred Value : 0.35714
##          Neg Pred Value : 0.37778
##              Prevalence : 0.58654
##          Detection Rate : 0.04808
##    Detection Prevalence : 0.13462
##       Balanced Accuracy : 0.43633
##
##        'Positive' Class : 0
##
```

Observaciones:

- El modelo clasifica correctamente le 13.5% de las muestras, es prácticamente peor que al azar.
- El kappa indica que el modelo no mejora sobre el azar.
- La sensibilidad para la calse positiva es muy baja.

Interpretación: Todas las métricas indican que el modelo clasifica muy mal la clase positiva, cun una tasa muy alta de falsos negativos.

**7. Escribe (ecuación), grafica el modelo significativo e interprétalo en el contexto del problema. Añade si es buen modelo, en qué no lo es, cuánto cambia.**

Ecuación del modelo significativo:

$$\ln(\frac{p}{1-p}) = 0.20326 + 0.05810 \cdot Lag2$$

Tomando como punto de referencia la evaluación completa del modelo se puede observar que este modelo no es el adecuado para determinar la dirección.