Actividad 2.8 Regresión Logística

Franco Mendoza Muraira A01383399

2023-11-25

```
library(ISLR)
```

```
## Warning: package 'ISLR' was built under R version 4.2.3
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.2.3
```

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

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

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

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.2.3
```

```
library(vcd)
```

```
## Warning: package 'vcd' was built under R version 4.2.3
```

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

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

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

```
library(lmtest)
```

```
## Warning: package 'lmtest' was built under R version 4.2.3
```

```
## Loading required package: zoo
```

```
## Warning: package 'zoo' was built under R version 4.2.3
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
#cargando la base de datos
data=Weekly

#resumen de dataset
summary(data)
```

```
##       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(data)
```

```
##   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. Divida el conjunto de datos en un conjunto de:

```
train_indices <- data$Year < 2009
test_indices <- data$Year >= 2009

train <- data[train_indices, ]
test <- data[test_indices, ]


table(train$Direction)
```

```
##
## Down   Up
##  441  544
```

```
table(test$Direction)
```

```
##
## Down   Up
##   43   61
```

2. Formule un modelo de regresión logística con el cual predecir el rendimiento actual del índice bursátil.

Mediante el uso de la función glm(modelo lineal) ajustamos el modelo de regresión logística para nuestra base de entrenamiento.

```
#Ajuste del modelo
model =glm(Direction~Lag2, family="binomial", data=train)

#para la notación científica en el resumen
options(scipen=999)

#resumen del modelo
summary(model)
```

```
##
## Call:
## glm(formula = Direction ~ Lag2, family = "binomial", data = train)
##
## 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
```

3. Escriba el modelo de regresion logistica:

$logit(Direction) = 0.20326 + 0.05810 * Lag_2$

$p(Direction) = \frac{e^{0.20326+0.05810*Lag_2}}{1+e^{0.20326+0.05810*Lag_2}}$

4. Interprete en el contexto del problema

-¿Es estadísticamente significativo el predictor (Lag2) ? ¿Cuál es su p-value?.

Dado a que el p-value de el predictor Lag2 es de 0.04298, Lag2 es estadísticamente significativo a un nivel de significancia del 0.05. En resumen, hay evidencia para sugerir que el coeficiente de Lag2 es distinto de cero, lo que significa que Lag2 tiene cierta influencia en la dirección del mercado.

-¿Qué indica el valor β_1?:

Por cada unidad que se incrementa la variable Lag2, se espera que el logaritmo de odds de la variable Direction se incremente en promedio: 0.05810 unidades.

Es decir, por cada unidad que se incrementa la variable Lag2, los odds de que Direction sea "Up" se incrementan en promedio 1.06 unidades. Esto corresponde a una probabilidad de que el mercado tenga un valor positivo en el día de hoy de p= 0.51

5. Represente gráficamente el modelo, grafique la curva de regresión logarítmica.

Predicciones de probabilidades

```
#Predicción de probabilidad de incumplimiento de cada individuo del set de prueba
prob_test=predict(model, test, type="response")
head(prob_test)
```

```
##       986       987       988       989       990       991
## 0.5261291 0.6447364 0.4862159 0.4852001 0.5197667 0.5401255
```

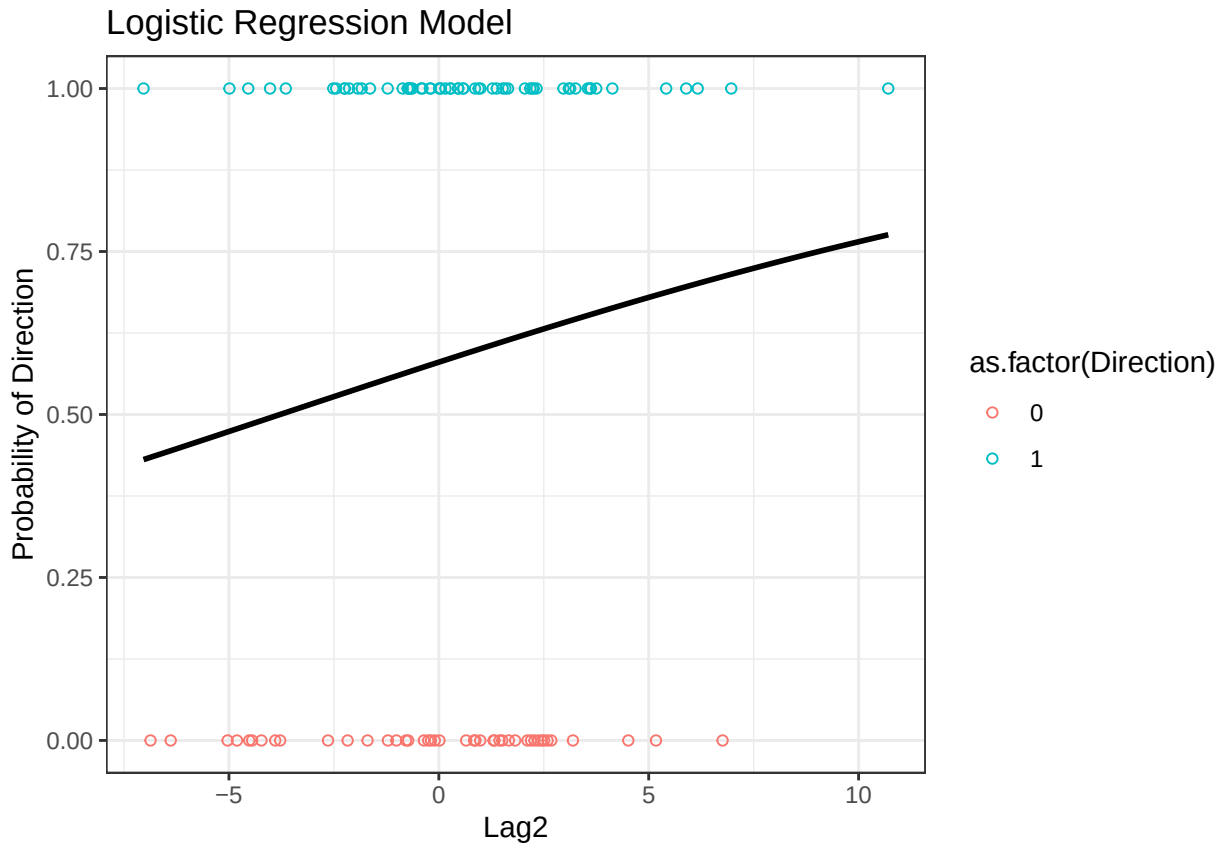
```
#Convierte la variable Direction: "Up" and "Down" a 1's y 0's
test$Direction= ifelse(test$Direction=="Up", 1, 0)
head(test$Direction)
```

```
## [1] 0 0 0 0 1 0
```

Grafica

```
test %>%
  ggplot(aes(Lag2, test$Direction)) +
  geom_point(aes(color = as.factor(Direction)), shape = 1) +
  geom_smooth(method = "glm", method.args = list(family = "binomial"),color = "black",se = FALSE) +
   theme_bw() +
  labs(
    title = "Logistic Regression Model",
    x = "Lag2",
    y = "Probability of Direction"
    )
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



6. Evalúe el modelo.

```
anova(model)
```

```
## Analysis of Deviance Table
##
```

```
## Model: binomial, link: logit
##
## Response: Direction
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev
## NULL                  984     1354.7
## Lag2  1   4.1666       983     1350.5
```

El modelo nulo (sin predictor) tiene 984 grados de libertad (Resid. Df) y una devianza de 1354.7. El modelo con el predictor Lag2 tiene 983 grados de libertad (Resid. Df) y una devianza residual de 1350.5. Aquí, la devianza del modelo con Lag2 es menor que la devianza del modelo nulo. Esto sugiere que el modelo que incluye Lag2 como predictor tiene una mejor capacidad para explicar la variabilidad en los datos en comparación con el modelo sin ningún predictor.

La diferencia en devianza indica que al incluir Lag2, se ha reducido la devianza en 4.2 unidades, lo que sugiere una mejora en la explicación del modelo.

En términos de significancia, esta reducción en la devianza nos dice que el modelo con Lag2 es estadísticamente significativo y proporciona una mejor explicación de los datos en comparación con un modelo sin ningún predictor.

```
library(MKclass)
```

```
## Warning: package 'MKclass' was built under R version 4.2.3
```

```
p_opt=optCutoff(prob_test, truth=test$Direction, namePos = 1)[1]
p_opt
```

```
## Optimal Cut-off
##       0.5143121
```

```
predicted.classes=ifelse(prob_test > p_opt, 1, 0)
head(predicted.classes)
```

```
## 986 987 988 989 990 991
##   1   1   0   0   1   1
```

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.2.3
```

```
## Loading required package: lattice
```

```
## Warning: package 'lattice' was built under R version 4.2.3
```

```
conf.table=table(pred=predicted.classes, true=test$Direction)
conf.table
```

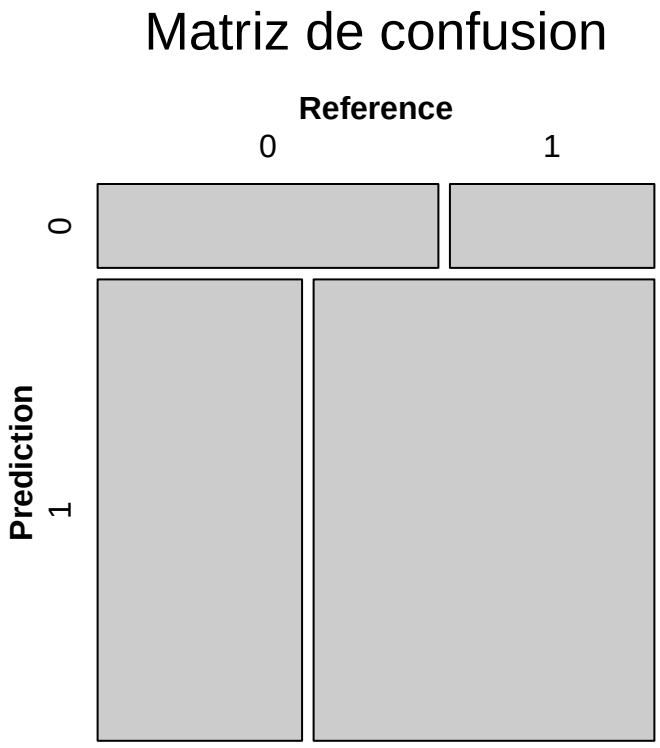
```
##     true
## pred  0  1
##    0 10  6
##    1 33 55
```

```
confusion =confusionMatrix(as.factor(predicted.classes), as.factor(test$Direction))
confusion
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 10  6
##          1 33 55
##
##                Accuracy : 0.625
##                  95% CI : (0.5247, 0.718)
##     No Information Rate : 0.5865
##     P-Value [Acc > NIR] : 0.2439
##
##                   Kappa : 0.1479
##
##  Mcnemar's Test P-Value : 0.00003136
##
##             Sensitivity : 0.23256
##             Specificity : 0.90164
##          Pos Pred Value : 0.62500
##          Neg Pred Value : 0.62500
##              Prevalence : 0.41346
##          Detection Rate : 0.09615
##    Detection Prevalence : 0.15385
##       Balanced Accuracy : 0.56710
##
##        'Positive' Class : 0
##
```

En resumen, el modelo tiene una precisión del 62.5%, pero muestra un desempeño asimétrico entre las dos clases (0 Down y 1 Up), con una sensibilidad baja para la clase 0 y una alta especificidad para la clase 1. Además, el coeficiente Kappa sugiere una concordancia ligera entre las predicciones y las clases reales.

```
mosaic(confusion$table,main="Matriz de confusion")
```



Aquí podemos ver un mosaico de la matriz de confusión, y se puede remarcar que la mayor parte de las predicciones son de que va a ir en una dirección para arriba.

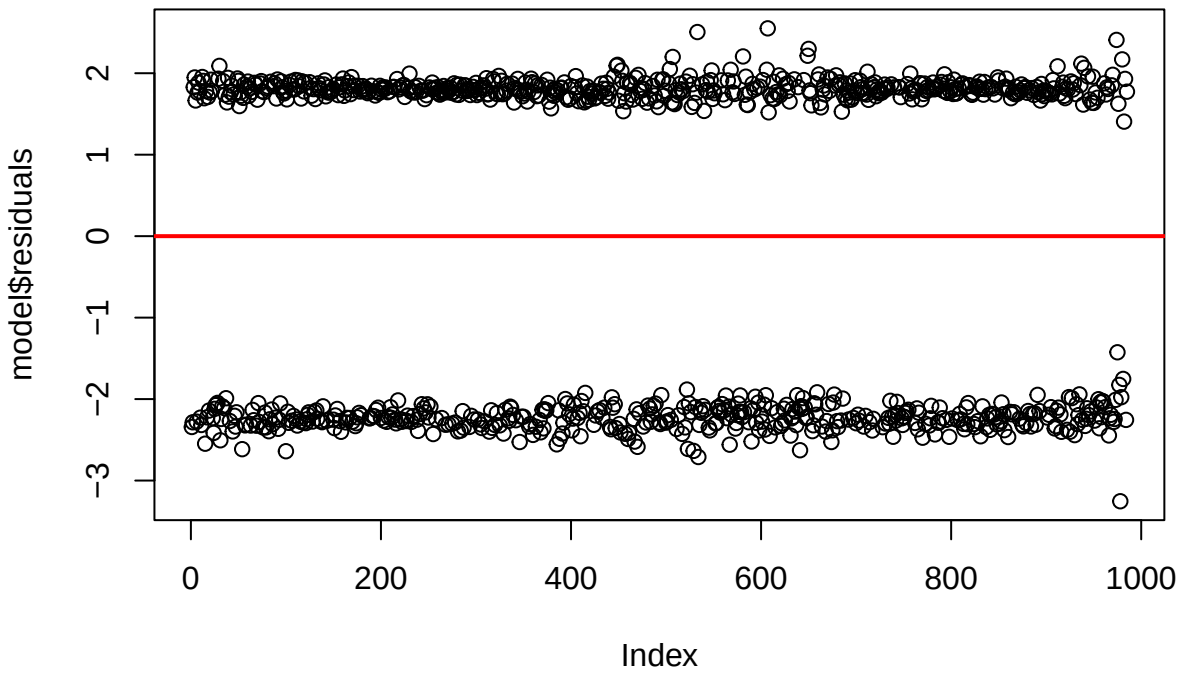
7. Valide los supuestos del modelo

Independencia

H_0 : Los residuos son independientes.

H_1 : Los residuos no son independientes.

```
plot(model$residuals)
abline(h=0, col = "red", lwd = 2)
```



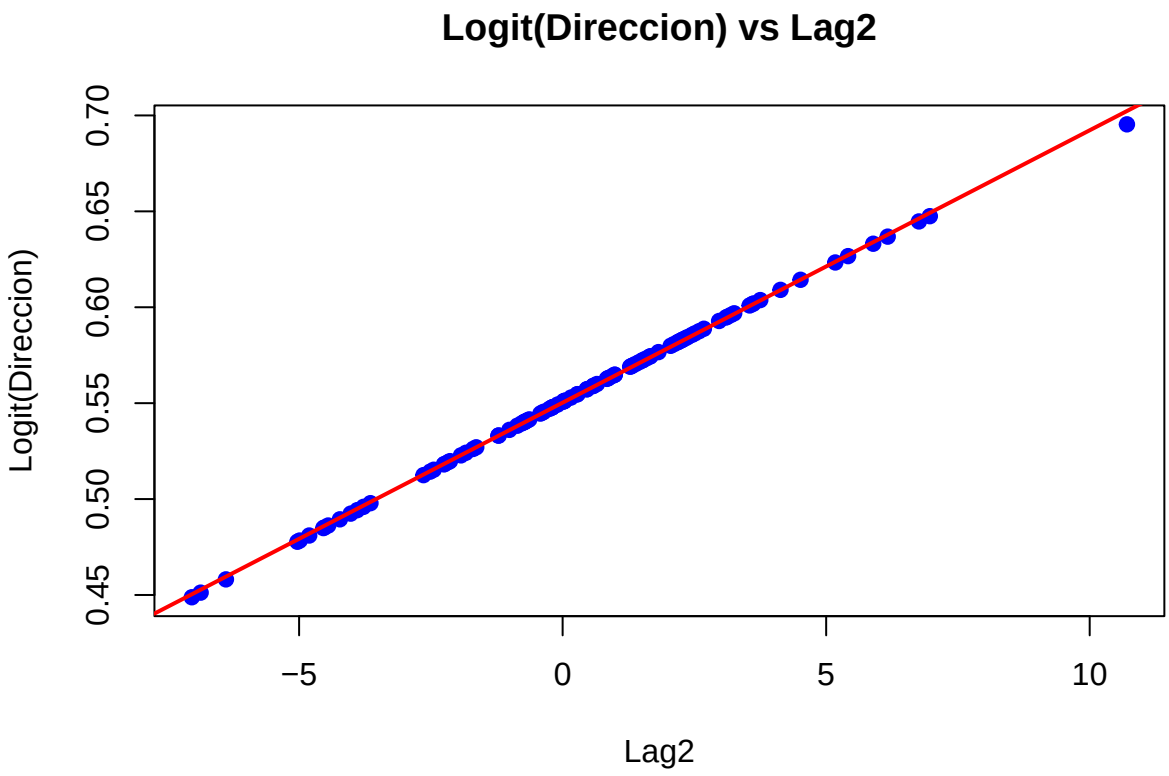

```
dwtest(model)
```

```
##
##  Durbin-Watson test
##
## data:  model
## DW = 2.154, p-value = 0.9923
## alternative hypothesis: true autocorrelation is greater than 0
```

Debido al valor p de 0.9923, y por cómo se ven los residuos en la gráfica, estando lejos de la línea roja todos, podemos ver que hay independencia, ya que no tenemos suficiente evidencia para rechazar la hipótesis nula. En conclusión, sí hay independencia en los residuos.

Linealidad

```
plot(test$Lag2,prob_test,pch=19,col="blue",xlab="Lag2",ylab="Logit(Direccion)",main="Logit(Direccion) vs
abline(lm(prob_test ~ test$Lag2), col="red", lwd=2)
```



Debido a que los valores de las predicciones están en la línea roja del modelo lineal, se acepta que hay linealidad en el modelo.

Tamaño muestral

```
table(data$Direction)
```

```
##
## Down   Up
##  484  605
```

El resultado menos frecuente es el de Down, y se tienen 484 resultados, siendo que es mucho mayor que 10, se cumple el tamaño muestral.