M5_A7

Sofia Cantu

2024-11-05

Regresión logística

```
# Cargar librerías
if (!require(ISLR)) install.packages("ISLR")
```

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

```
library(ISLR)
if (!require(tidyverse)) install.packages("tidyverse")
```

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

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse
2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ ggplot2   3.5.1     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.1
## ✔ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force
all conflicts to become errors
```

```
library(tidyverse)

data("Weekly")
```

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

```
#Estadísticas descriptivas
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
##
##
##
##
```

```
cat("\n\n\n")
```

```
# Calcular matriz de correlación excluyendo la variable categórica
'Direction'
cor_matrix <- cor(Weekly[, -9])

# Visualizar la matriz de correlación
print(cor_matrix)
```

```
##               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
```

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

```r
# Ajustar el modelo logístico
glm_full <- glm(Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 + Volume,
                data = Weekly, family = binomial)

# Resumen del modelo
summary(glm_full)
```

```
##
## Call:
## glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
##     Volume, family = binomial, data = Weekly)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.26686    0.08593   3.106   0.0019 **
## Lag1        -0.04127    0.02641  -1.563   0.1181
## Lag2         0.05844    0.02686   2.175   0.0296 *
## Lag3        -0.01606    0.02666  -0.602   0.5469
## Lag4        -0.02779    0.02646  -1.050   0.2937
## Lag5        -0.01447    0.02638  -0.549   0.5833
## Volume      -0.02274    0.03690  -0.616   0.5377
## ---
## 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: 1486.4  on 1082  degrees of freedom
## AIC: 1500.4
##
## Number of Fisher Scoring iterations: 4
```

```r
# Calcular intervalos de confianza al 95%
confint(glm_full)
```

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

```
##                    2.5 %     97.5 %
## (Intercept)  0.098808746 0.43580101
## Lag1        -0.093477110 0.01029269
## Lag2         0.006197597 0.11169774
## Lag3        -0.068653910 0.03604309
## Lag4        -0.079952378 0.02401603
```

```
## Lag5        -0.066495108 0.03711989
## Volume      -0.095051949 0.04979338
```

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.

```r
# Crear indicador para el conjunto de entrenamiento
train <- (Weekly$Year >= 1990) & (Weekly$Year <= 2008)

# Conjunto de entrenamiento y prueba
Weekly_train <- Weekly[train, ]
Weekly_test <- Weekly[!train, ]

# Ajustar modelo al conjunto de entrenamiento
glm_train <- glm(Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 + Volume,
                 data = Weekly_train, family = binomial)

# Resumen del modelo
summary(glm_train)
```

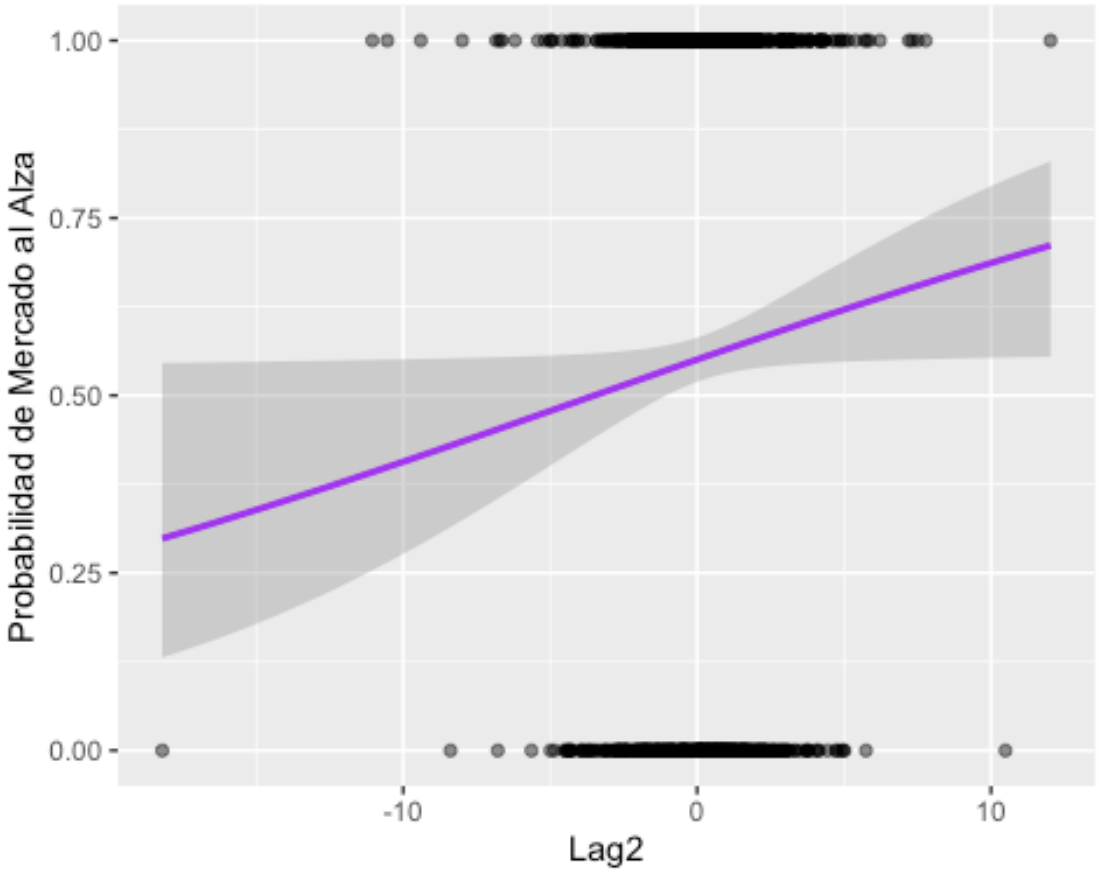
```
##
## Call:
## glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
##     Volume, family = binomial, data = Weekly_train)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.33258    0.09421   3.530 0.000415 ***
## Lag1        -0.06231    0.02935  -2.123 0.033762 *
## Lag2         0.04468    0.02982   1.499 0.134002
## Lag3        -0.01546    0.02948  -0.524 0.599933
## Lag4        -0.03111    0.02924  -1.064 0.287241
## Lag5        -0.03775    0.02924  -1.291 0.196774
## Volume      -0.08972    0.05410  -1.658 0.097240 .
## ---
## 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: 1342.3  on 978  degrees of freedom
## AIC: 1356.3
##
## Number of Fisher Scoring iterations: 4
```

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

```
# Ajustar modelo con variables significativas
glm_sig <- glm(Direction ~ Lag2, data = Weekly_train, family = binomial)

# Resumen del modelo
summary(glm_sig)

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

5. Representa gráficamente el modelo:

```
# Graficar modelo logístico
ggplot(Weekly_train, aes(x = Lag2, y = as.numeric(Direction == "Up"))) +
  geom_point(alpha = 0.5) +
  stat_smooth(method = "glm", method.args = list(family = "binomial"), color
= "purple") +
  labs(x = "Lag2", y = "Probabilidad de Mercado al Alza")

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



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

```
# Prueba de chi cuadrada
anova(glm_sig, 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

# Predicciones en el conjunto de prueba
pred_probs <- predict(glm_sig, Weekly_test, type = "response")
```

```r
pred_direction <- ifelse(pred_probs > 0.5, "Up", "Down")

# Matriz de confusión
conf_matrix <- table(Predicted = pred_direction, Actual =
Weekly_test$Direction)
print(conf_matrix)
```

```
##          Actual
## Predicted Down Up
##      Down    9  5
##      Up     34 56
```

```r
# Calcular precisión
accuracy <- mean(pred_direction == Weekly_test$Direction)
print(paste("Precisión del modelo:", round(accuracy * 100, 2), "%"))
```

```
## [1] "Precisión del modelo: 62.5 %"
```

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)

7.1 Ecuación del Modelo Significativo

logit(P(Direction=Up)) = B0+B1*Lag2* $logit(P(Direction=Up)) = 0.20326+0.505810$Lag2

Donde: - B0: Intercepto del modelo - B1: Coeficiente asociado a Lag2 - logit(P(Direction = Up)) es la función logit de la probabilidad de que la dirección sea “Up”. - Lag2 es el segundo desfase de la variable dependiente. Esto significa que por cada unidad de incremento en Lag2, el log-odds de que Direction sea “Up” aumenta en 0.05810.

7.2 Gráfica del Modelo Significativo

La gráfica generada del modelo muestra cómo la variable Lag2 influye en la probabilidad de que el mercado suba. A medida que Lag2 incrementa, la probabilidad de que Direction sea “Up” también aumenta, aunque de manera moderada, como se observa en la pendiente suave de la línea ajustada.

7.3 Interpretación en el Contexto del Problema

En este caso, Direction representa la dirección del mercado, donde “Up” implica un incremento y “Down” una caída. La variable Lag2, que representa el valor de Direction con un retraso de dos periodos, tiene una influencia significativa sobre la dirección actual. Esto podría interpretarse como una posible correlación entre los movimientos pasados y la dirección futura del mercado.

7.4 Evaluación del Modelo

Precisión: El modelo obtuvo una precisión de 62.5% en la matriz de confusión. Aunque es mejor que el azar, esta precisión es limitada y podría no ser suficiente para aplicaciones prácticas en el mercado financiero, donde se requiere mayor confiabilidad.

Ventajas: - Simplicidad: Al reducirse a una sola variable (Lag2), el modelo es fácil de interpretar y puede ser útil en contextos donde los datos o recursos son limitados. - Interpretabilidad: Es posible entender cómo Lag2 afecta la probabilidad de la dirección futura del mercado, lo que puede dar indicios sobre la inercia del mercado. Limitaciones: - Capacidad Predictiva: Con una precisión de 62.5%, el modelo no es muy preciso en la predicción de la dirección del mercado. Existen muchos factores que influyen en los movimientos del mercado, y es probable que este modelo sea demasiado simplista. - Posibles Omisiones: El modelo no incluye otras variables que podrían ser relevantes, como factores macroeconómicos o indicadores de sentimiento, lo cual limita su aplicabilidad en situaciones reales. - Falta de Robustez: Al basarse en una única variable de retraso, el modelo podría ser sensible a variaciones y ruido en los datos históricos, reduciendo su estabilidad en otros contextos.

7.5 Posibles Mejoras

Para mejorar el modelo, podrías considerar: - Incorporar más variables: Incluir otros lags o variables externas relacionadas con el mercado. - Métodos de Machine Learning: Explorar algoritmos más complejos como redes neuronales o árboles de decisión. - Validación cruzada: Aumentar la confiabilidad del modelo mediante validación cruzada para evaluar su desempeño en diferentes subconjuntos de datos.