actividad 5

Diego Elian Rodriguez Cantú

2023-10-17

1. Análisis de datos: Estadísticas descriptivas y coeficiente de correlación entre las variables.

```
# Instala e importa las librerías necesarias
#install.packages("ISLR")
library(ISLR)

# Análisis descriptivo
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
##
##
##
##
```

```
# Coeficiente de correlación
cor(Weekly[, -9])  # Excluimos la columna de Direction que es categórica
```

```
##                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
```

1. Resumen estadístico:

- **Year**: El año de los datos varía entre 1990 y 2010.
- **Lag1** a **Lag5**: Representan los valores de mercado de semanas anteriores. Todos ellos tienen valores mínimos de -18.1950 y máximos de 12.0260. Su media oscila alrededor de 0.15, lo que significa que, en promedio, no hay grandes cambios semana a semana. La mediana es positiva, alrededor de 0.24, lo que sugiere que hay más semanas con ligeros aumentos que con descensos.
- **Volume**: Representa el volumen de acciones. Su valor medio es 1.575, con un rango que va desde 0.087 a 9.328. El volumen ha aumentado con los años, como lo indica su correlación positiva con el año.
- **Today**: Representa el valor del día actual. Sus estadísticas son similares a las de los valores de retraso, pero su correlación con las otras variables es diferente.
- **Direction**: De todas las observaciones, 484 tuvieron una dirección "Down" y 605 tuvieron una dirección "Up".

2. Coeficientes de correlación:

- **Year y Volume** tienen una fuerte correlación positiva de 0.842, lo que indica que el volumen de acciones ha aumentado con el tiempo.
- Las variables **Lag1** a **Lag5** tienen correlaciones bajas con las otras variables, lo que es esperado dado que representan valores pasados del mercado. No hay evidencia de multicolinealidad entre estas variables.
- **Volume y Today** tienen una correlación negativa (-0.033), aunque es débil.
- Las correlaciones entre **Lag1 a Lag5 y Today** son bajas, indicando que los valores de las semanas anteriores no tienen una correlación fuerte con el valor del día actual.

Interpretación General:

1. A lo largo del periodo 1990-2010, el volumen de acciones ha ido en aumento, como lo indica la correlación entre el año y el volumen.
2. En términos generales, el mercado ha tenido leves aumentos semanales, con más semanas "Up" que "Down".
3. No parece haber una relación fuerte entre los rendimientos de las semanas anteriores y el rendimiento de la semana actual, ya que las correlaciones entre las variables **Lag** y **Today** son bajas.
4. Es relevante considerar que, a pesar de que las correlaciones entre las variables **Lag** y **Today** son bajas, esto no necesariamente implica que no haya una relación predictiva. Es por ello que se debe explorar un modelo logístico.

2. Modelo logístico con todas las variables menos la variable "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
```

Modelo:

El modelo logístico intenta predecir la dirección ("Up" o "Down") del índice bursátil S&P 500 basándose en las variables `Lag1`, `Lag2`, `Lag3`, `Lag4`, `Lag5`, y `Volume`.

Coeficientes:

- **(Intercept)**: La ordenada al origen, 0.26686, representa el logit de la probabilidad de que el mercado suba (es decir, Direction = "Up") cuando todas las otras variables son 0. Es significativo con un p-valor de 0.0019.
- **Lag1**: Por cada incremento de una unidad en `Lag1`, el logit de la probabilidad de que el mercado suba disminuye en 0.04127. No es estadísticamente significativo (p-valor = 0.1181).
- **Lag2**: Por cada incremento de una unidad en `Lag2`, el logit de la probabilidad de que el mercado suba aumenta en 0.05844. Es estadísticamente significativo con un p-valor de 0.0296.
- **Lag3**: Su coeficiente no es estadísticamente significativo (p-valor = 0.5469), lo que sugiere que `Lag3` no tiene un impacto significativo en predecir la dirección.
- **Lag4**: Al igual que `Lag1` y `Lag3`, `Lag4` tampoco es estadísticamente significativo (p-valor = 0.2937).
- **Lag5**: No es significativo (p-valor = 0.5833).
- **Volume**: El volumen de acciones no es un predictor significativo de la dirección, con un p-valor de 0.5377.

Interpretación general:

1. De las variables predictoras, solo `Lag2` resultó ser estadísticamente significativa para predecir la dirección del índice bursátil S&P 500.
2. Las demás variables (`Lag1`, `Lag3`, `Lag4`, `Lag5` y `Volume`) no parecen tener un impacto significativo en la predicción de la dirección.
3. Aunque `Lag2` es significativo, sería útil realizar más análisis y considerar otros posibles modelos para mejorar la predicción de la dirección.

Intervalos de confianza

Para calcular los intervalos de confianza para las estimaciones de los coeficientes B_i, podemos usar una aproximación normal. El intervalo de confianza para cada coeficiente se calcula como:

$$B_i \pm (z \times SE(B_i))$$

Donde B_i es la estimación del coeficiente y $SE(B_i)$ es el error estándar del coeficiente. Para un intervalo de confianza del 95%, z es aproximadamente 1.96.

Usando los datos proporcionados:

Variable	Estimación (B_i)	Error estándar (SE(B_i))
(Intercept)	0.26686	0.08593
Lag1	-0.04127	0.02641
Lag2	0.05844	0.02686
Lag3	-0.01606	0.02666
Lag4	-0.02779	0.02646
Lag5	-0.01447	0.02638
Volume	-0.02274	0.03690

Calculamos los intervalos de confianza para cada coeficiente:

(Intercept): $0.26686 \pm (1.96 \times 0.08593) = (0.09836, 0.43536)$ Lag1: $-0.04127 \pm (1.96 \times 0.02641) = (-0.09323, 0.01069)$ Lag2: $0.05844 \pm (1.96 \times 0.02686) = (0.00589, 0.11099)$ Lag3: $-0.01606 \pm (1.96 \times 0.02666) = (-0.06822, 0.03610)$ Lag4: $-0.02779 \pm (1.96 \times 0.02646) = (-0.07955, 0.02397)$ Lag5: $-0.01447 \pm (1.96 \times 0.02638) = (-0.06607, 0.03713)$ Volume: $-0.02274 \pm (1.96 \times 0.03690) = (-0.09514, 0.04966)$

Variables que influyen y no influyen en el modelo:

Las variables que no tienen 0 en sus intervalos de confianza y son significativas (basándonos en los p-valores) influyen en el modelo. De los resultados anteriores:

- **Influyen**: (Intercept) y Lag2.
- **No influyen**: Lag1, Lag3, Lag4, Lag5 y Volume.

Interpretación del efecto de las variables en los odds (momios):

- **Lag2**: Para un incremento de una unidad en Lag2, el odds de que el mercado suba (en comparación con que baje) aumenta aproximadamente por un factor de $e^{0.05844} = 1.060$. Esto significa que el mercado es aproximadamente un 6% más probable que suba por cada unidad de incremento en Lag2, manteniendo todas las demás variables constantes.

Las demás variables, al no ser significativas, no tienen un impacto claro en los odds, por lo que no se interpretan en este contexto.

3. Divide la base de datos en conjunto de entrenamiento y prueba.

1. **Dividir la Base de Datos**:

```
# Separar los datos
entrenamiento <- Weekly[Weekly$Year < 2009,]
prueba <- Weekly[Weekly$Year >= 2009,]
```

2. **Ajustar el Modelo**:

```r
modelo_entrenamiento <- glm(Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 + Volume, data=entrenamiento, f
summary(modelo_entrenamiento)
```

```
##
## Call:
## glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
##     Volume, family = binomial, data = entrenamiento)
##
## 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
```

Variables que son estadísticamente significativas al nivel de 0.05 (y su nivel de significancia): - `(Intercept)` con p-value = 0.000415 (**) - ***`Lag1`*** *con p-value = 0.033762 (*)*

`Volume` tiene un nivel de significancia ligeramente superior al 0.05 (p-value = 0.097240), por lo que podríamos considerarlo marginalmente significativo.

Por lo tanto, para formular el modelo logístico utilizando únicamente las variables que son estadísticamente significativas, debemos incluir `Lag1`:

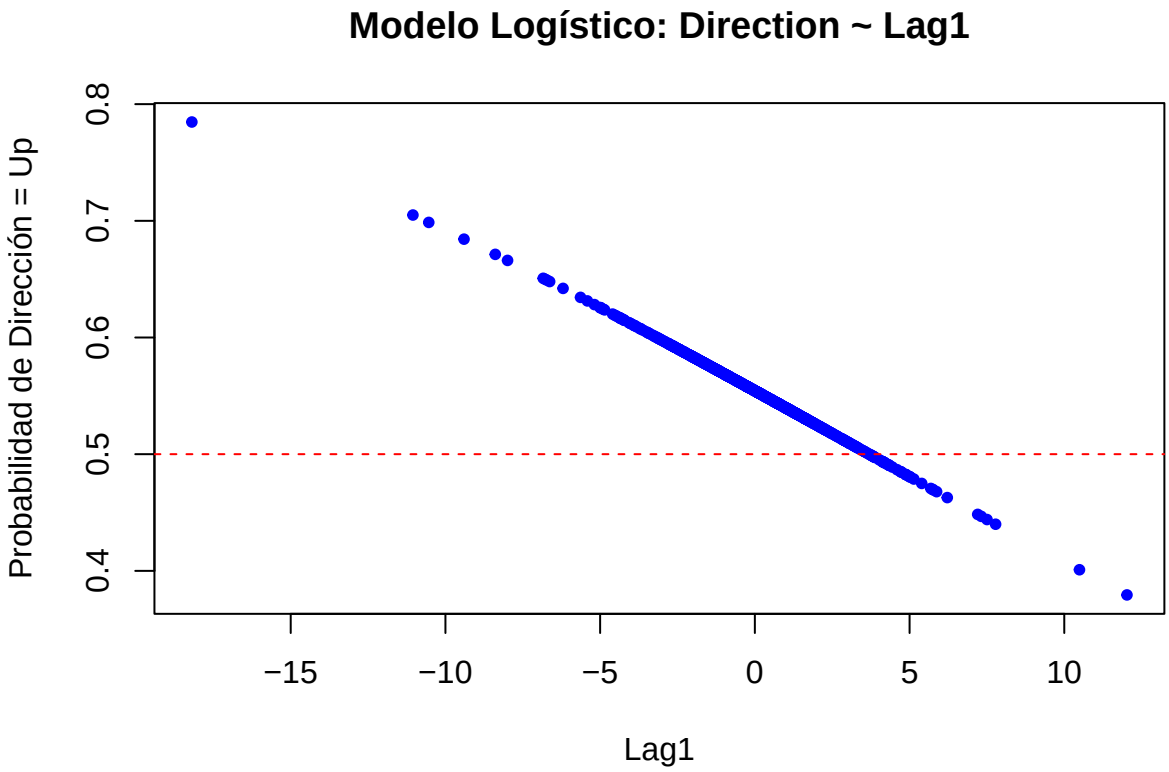
```r
modelo_significativo <- glm(Direction ~ Lag1, data=entrenamiento, family=binomial)
```

Representación Gráfica del Modelo:

Podemos representar gráficamente la relación entre la dirección y `Lag1` usando el modelo logístico:

```r
# Predicciones del modelo
entrenamiento$prediccion <- predict(modelo_significativo, type="response")

# Crear gráfico
plot(entrenamiento$Lag1, entrenamiento$prediccion, col="blue", pch=20, xlab="Lag1", ylab="Probabilidad d
abline(h=0.5, col="red", lty=2) # línea horizontal en y=0.5
```



Este gráfico muestra la probabilidad de que `Direction` sea "Up" en función del valor de `Lag1`. La línea roja punteada indica una probabilidad de 0.5, lo que puede ayudar a identificar qué valores de `Lag1` llevan a una probabilidad mayor o menor que 0.5.

```
modelo.log.s <- modelo_significativo

nuevos_puntos <- seq(from = min(Weekly$Lag1), to = max(Weekly$Lag1),
by = 0.5)

predicciones <- predict(modelo.log.s, newdata = data.frame(Lag1 =
nuevos_puntos),se.fit = TRUE, type = "response")
```

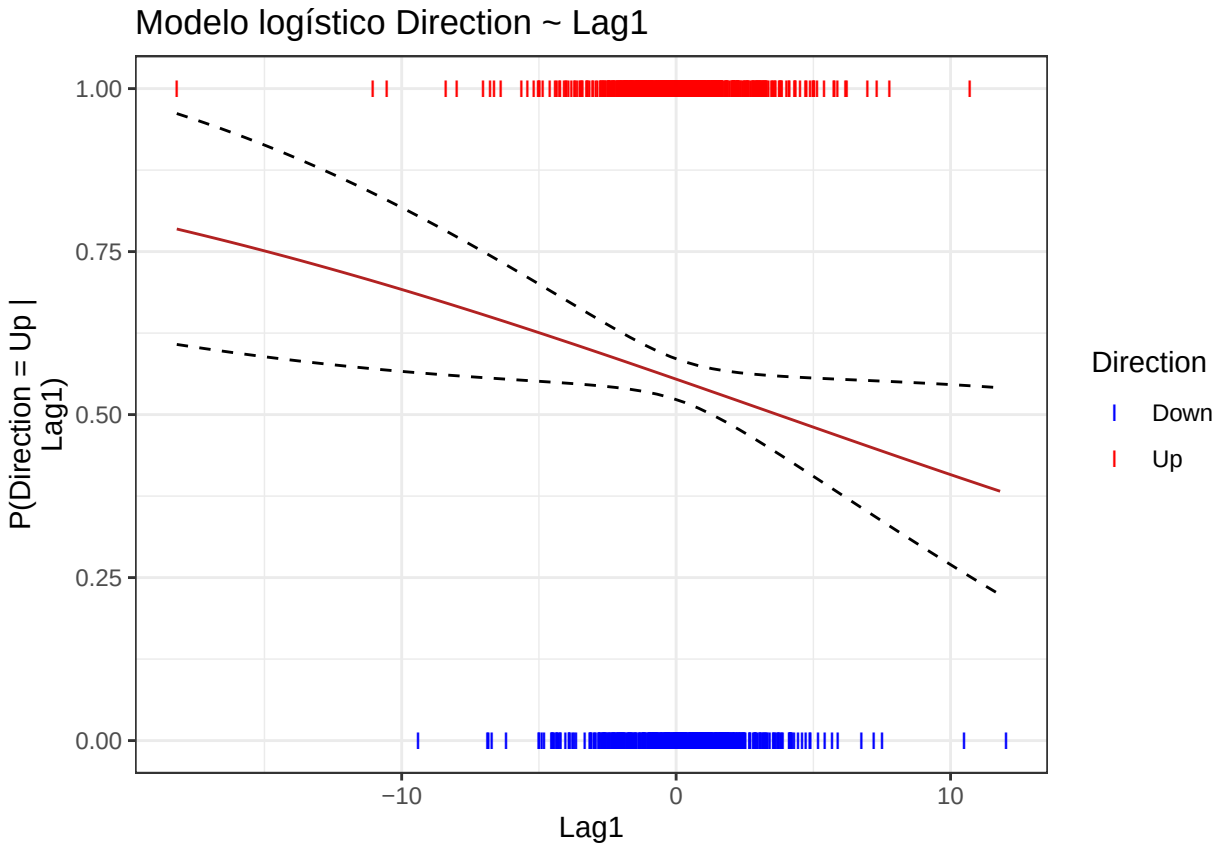
Límites de los intervalos de confianza:

```
# 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

datos_curva <- data.frame(Lag1 = nuevos_puntos, probabilidad =
predicciones$fit, CI.inferior = CI_inferior, CI.superior = CI_superior)
```

```
library(ggplot2)
Weekly$Direction <- ifelse(Weekly$Direction == "Down", yes = 0, no = 1)
ggplot(Weekly, aes(x = Lag1, 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 ~ Lag1", y = "P(Direction = Up |
Lag1)", x = "Lag1") +
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()
```



4. Evalua el modelo

```
anova(modelo_significativo, 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
## Lag1  1   4.2634       983     1350.5  0.03894 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El p-valor para Lag1 es 0.03894 (marcado con un *), lo que indica que Lag1 es estadísticamente significativo al nivel de 0.05. Dado que el p-valor es menor que 0.05, podemos rechazar la hipótesis nula de que Lag1 no tiene ningún efecto sobre la variable de respuesta “Direction”. Esto sugiere que hay una relación significativa entre Lag1 y “Direction”. En resumen, Lag1 es una variable significativa en el modelo y tiene un efecto en la

variable de respuesta “Direction”.

5. Prueba modelo

```
prob.modelo <- predict(modelo_significativo, newdata = prueba, type = "response")
```

```
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  = prueba$Direction
# Matriz de confusión
matriz.confusion <- table(pred.modelo, Direction.0910)
matriz.confusion
```

```
##            Direction.0910
## pred.modelo Down Up
##        Down    4  6
##        Up     39 55
```

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