Actividad2_6

Ricardo Kaleb Flores Alfonso

2024-10-09

0) Se importan librerias

```
library(ISLR)
library(dplyr)
```

```
##
## Adjuntando el paquete: '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)
```

1) Se carga la base de datos y se dividen

```
data <- Weekly
data <- data[,c("Year","Lag2","Direction")]
data$Direction <- ifelse(data$Direction=="Up", 1, 0)
train <- data[data$Year < 2008,]
test <- data[data$Year >= 2008,]
```

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

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

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

#resumen del modelo
summary(model)
```

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

```
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.22658    0.06621   3.422 0.000621 ***
## Lag2         0.04716    0.03230   1.460 0.144293
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1280.6  on 932  degrees of freedom
## Residual deviance: 1278.5  on 931  degrees of freedom
## AIC: 1282.5
##
## Number of Fisher Scoring iterations: 4
```

$$log(\frac{P(Up)}{P(Down)}) = 0.227 + 0.047 * Lag2$$

$$p = \frac{e^{0.227+0.047*Lag2}}{1 + e^{0.227+0.047*Lag2}}$$

4) Interprete en el contexto del problema:

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

El valor obtenido de $p = 0.144293 > 0.05$, por lo que esta variable es significativa para predecir si el mercado va subir o bajar

¿Qué indica el valor: β_1?:

Por por cada unidad que se incrementa la variable Lag2, se espera que el logaritmo de odds de la variable Direction se incremente en promedio: 0.047 unidades.

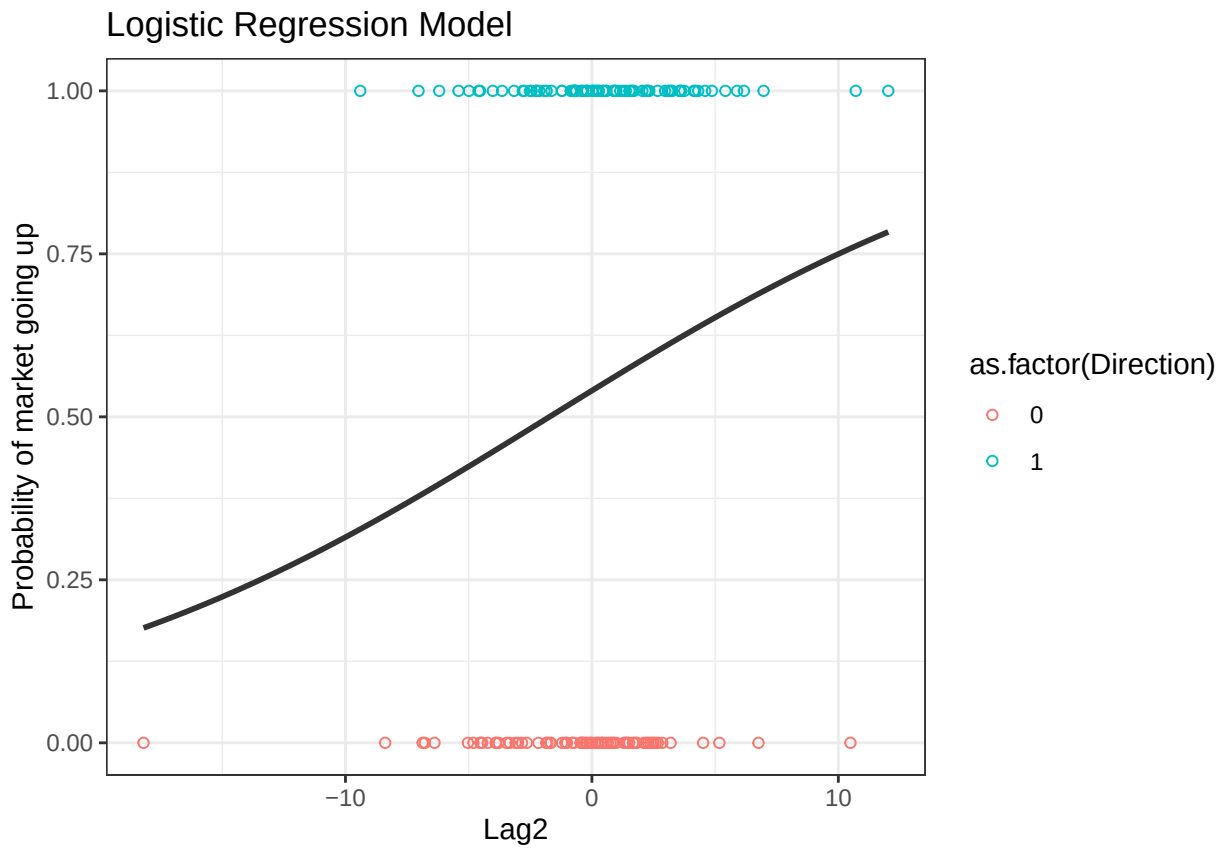
Es decir, por cada unidad que se incrementa la variable Lag2, los odds de que Direction sea "Up" se incrementan en promedio 1.048 unidades. Esto corresponde a una probabilidad de que el mercado tenga un valor positivo en el día de hoy de p, suponiendo que el lag2 tenga un valor de 1.5, entonces $p = \frac{e^{0.005}}{1+e^{0.005}} = 0.502$, por lo que se podria decir que es más probable que sea un día con rendimiento positivo

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

```
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 = "gray20",
              se = FALSE) +
  theme_bw() +
  labs(
    title = "Logistic Regression Model",
    x = "Lag2",
    y = "Probability of market going up"
    )
```

```
## Warning: Use of `test$Direction` is discouraged.
## i Use `Direction` instead.
## Use of `test$Direction` is discouraged.
## i Use `Direction` instead.
```

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




```
prob_test = predict(model,test,type="response")
predicted.classes = ifelse(prob_test > 0.5, 1, 0)
```

6) Evalúe el modelo.

```
anova(model, 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                  932     1280.7
## Lag2  1   2.1426       931     1278.5   0.1433
```

anova() -> Analiza la significancia de la diferencia ("Deviance") de residuos entre ambos modelos ("Null deviance" y "Residual deviance").

```
library(caret)
```

```
## Cargando paquete requerido: lattice
```

```
#Opción 1
conf.table = table(pred=predicted.classes, true=test$Direction)
conf.table
```

```
##     true
## pred  0  1
##    0  7  5
##    1 65 79
```

Vemos que el modelo se desempeña muy bien para predecir cuando va a haber subida en el mercado, sin embargo falla en mostrar cuando el mercado va a bajar

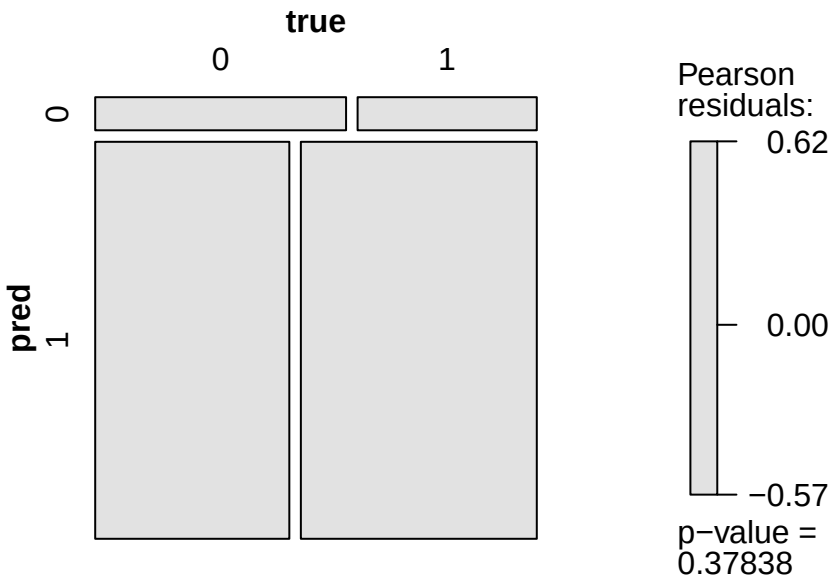
```
library(vcd)
```

```
## Cargando paquete requerido: grid
```

```
##
## Adjuntando el paquete: 'vcd'
```

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

```
# Gráfico de mosaico
mosaic(conf.table, shade = TRUE, legend = TRUE)
```



Esta distribución se ve más clara en el gráfico de mosaico.

```
correct_predictions <- sum(diag(conf.table))
total_predictions <- sum(conf.table)
accuracy <- correct_predictions / total_predictions * 100
error_rate <- (1 - correct_predictions / total_predictions) * 100

print(paste("Porcentaje de predicciones correctas: ", round(accuracy, 2), "%"))
```

```
## [1] "Porcentaje de predicciones correctas:  55.13 %"
```

```
print(paste("Tasa de error: ", round(error_rate, 2), "%"))
```

```
## [1] "Tasa de error:  44.87 %"
```

```
#Sensibilidad(TP / (TP + FN))
conf.table[2,2]/(conf.table[1,2]+conf.table[2,2])
```

```
## [1] 0.9404762
```

```
#Especificidad (VN / (VN + FP))
conf.table[1,1]/(conf.table[1,1]+conf.table[2,1])
```

```
## [1] 0.09722222
```

Sensibilidad: Para aquellas semanas con un valor de mercado al alza, el modelo clasifica correctamente el:0.94 % de las observaciones Especificidad: Para las semanas con un valor de mercado a la baja, el modelo acierta en un: 0.1 % de las observaciones

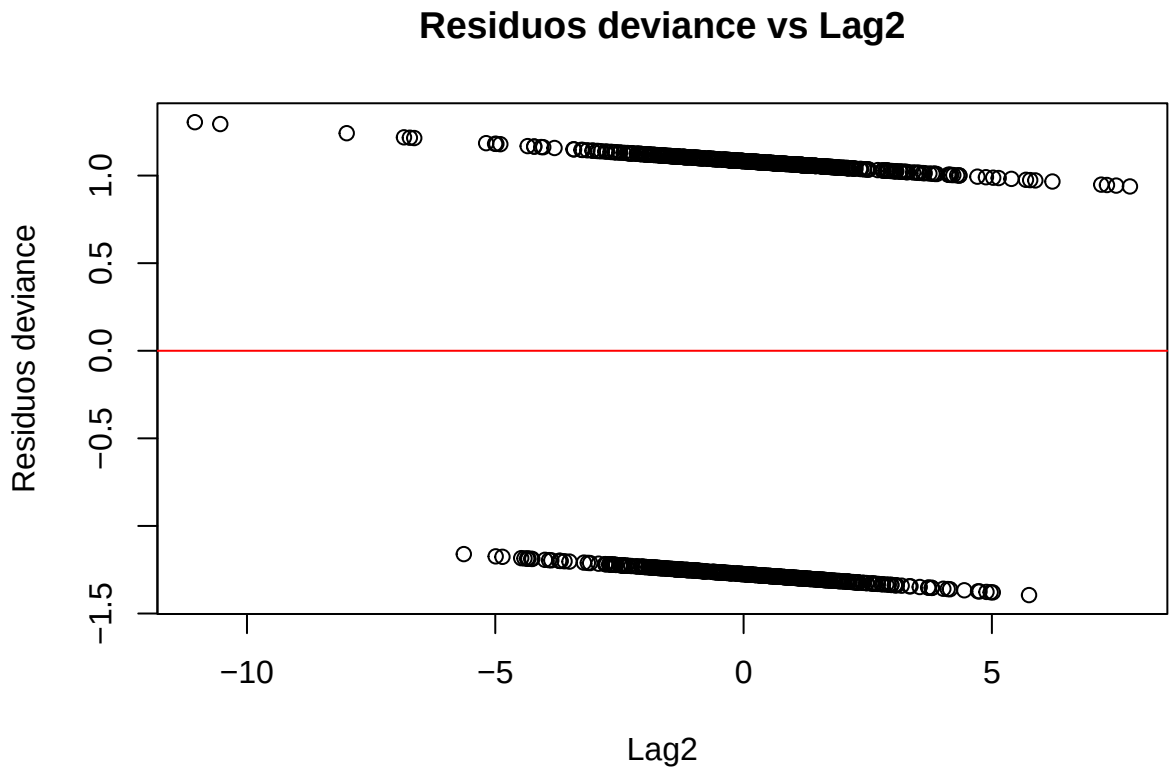
Esto nos muestra que el modelo tiene una alta sensibilidad, pero una baja especificidad

7) Valide los supuestos del modelo.

Independencia: las observaciones han de ser independientes.

```
# Calcular los residuos deviance
residuos <- residuals(model, type = "deviance")

# Crear gráfico de residuos contra el Lag2
plot(train$Lag2, residuos, xlab = "Lag2", ylab = "Residuos deviance", main = "Residuos deviance vs Lag2
abline(h = 0, col = "red")
```




```
library(lmtest)
```

```
## Cargando paquete requerido: zoo
```

```
##
## Adjuntando el paquete: 'zoo'
```

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

```
# Realizar la prueba de Durbin-Watson
dwtest(model)
```

```
##
##  Durbin-Watson test
##
## data:  model
```

```
## DW = 2.1582, p-value = 0.9923
## alternative hypothesis: true autocorrelation is greater than 0
```

El valor de DW sugiere una autocorrelación, de igual manera se observa en el gráfico que los valores se mantienen constantes. Sin embargo en el gráfico se observa como existe una relación lineal hacia abajo mientras el valor del lag2 aumenta.

Multicolinealidad : muy poca a ninguna relación lineal entre los predictores (para regresión logística múltiple).

```
library(car)
```

```
## Cargando paquete requerido: carData
```

```
##
## Adjuntando el paquete: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
# Calcular el VIF (esto es relevante si tienes múltiples predictores)
#vif(model)
```

No existe multicolinealidad pues solo exite una variable predictora que es Lag2

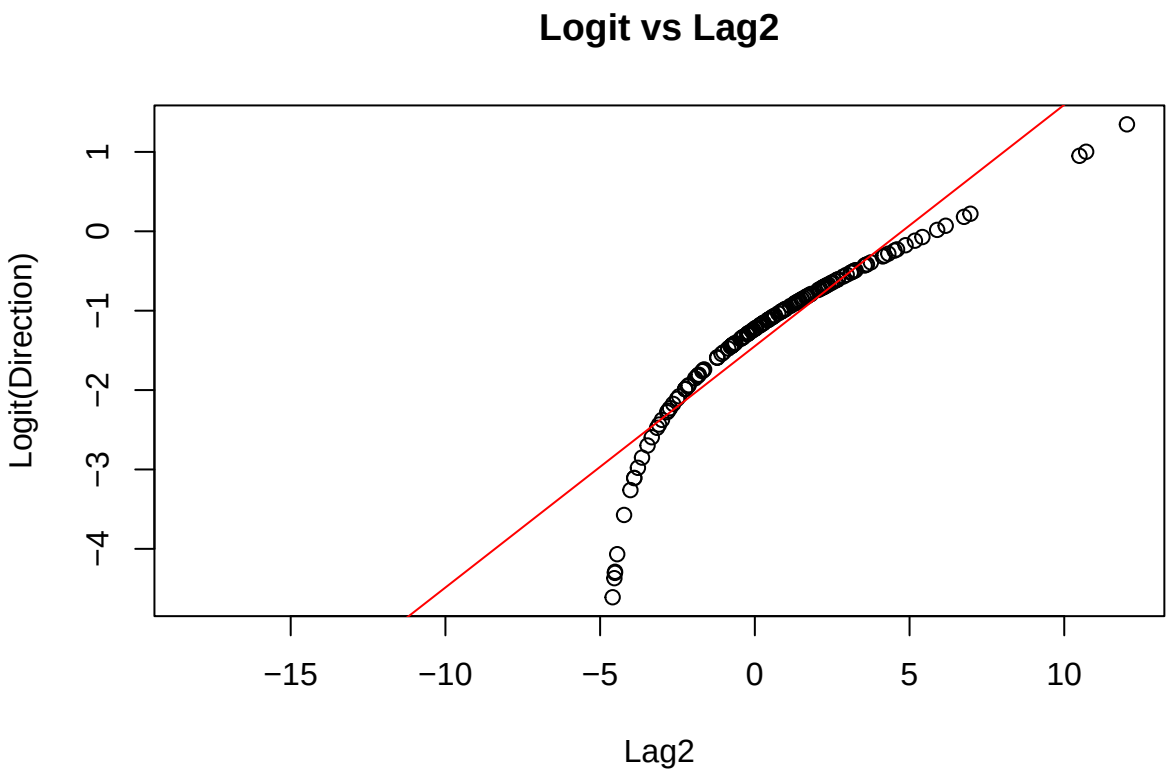
Linealidad: entre la variable independiente y el logaritmo natural de odds.

```
predicciones <- predict(model,test)
```

```
# Graficar Logit(Dirección) vs Lag2
logit <- log( predicciones / (1 - predicciones))
```

```
## Warning in log(predicciones/(1 - predicciones)): Se han producido NaNs
```

```
# Crear gráfico Logit vs Lag2
plot(test$Lag2, logit, xlab = "Lag2", ylab = "Logit(Direction)", main = "Logit vs Lag2")
abline(lm(logit ~ test$Lag2), col = "red")
```



No se observa linealidad en el caso de la relación entre logit y Lag2

```
# ANOVA para analizar la significancia del modelo
anova(model, 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                  932     1280.7
## Lag2  1   2.1426       931     1278.5   0.1433
```

La prueba de anova muestra que si existe significancia en el modelo obtenido para predecir las variables

Tamaño muestral:

```
# Ver la distribución de la variable respuesta
table(train$Direction)
```

```
##
##   0   1
```

```
## 412 521
```

Se observa que la variable respuesta se encuentra bien distribuida, incluso casi se tendria una distribución 1 a 1