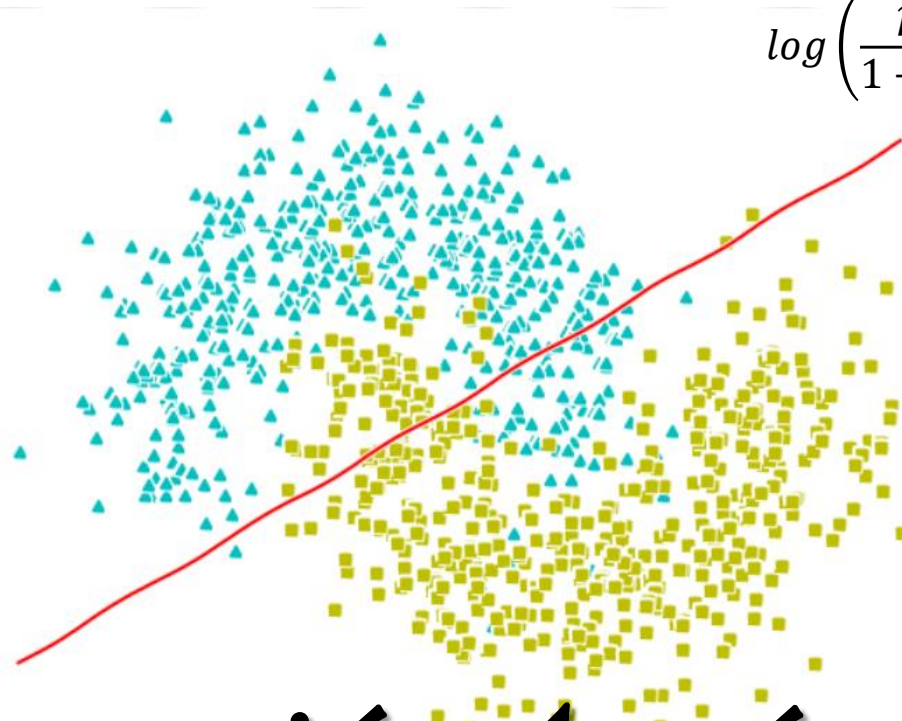


$$\log \left( \frac{p(x)}{1 - p(x)} \right) = \beta_0 + \beta_1 x$$



# Regresión logística

# ¿Por qué regresión logística?

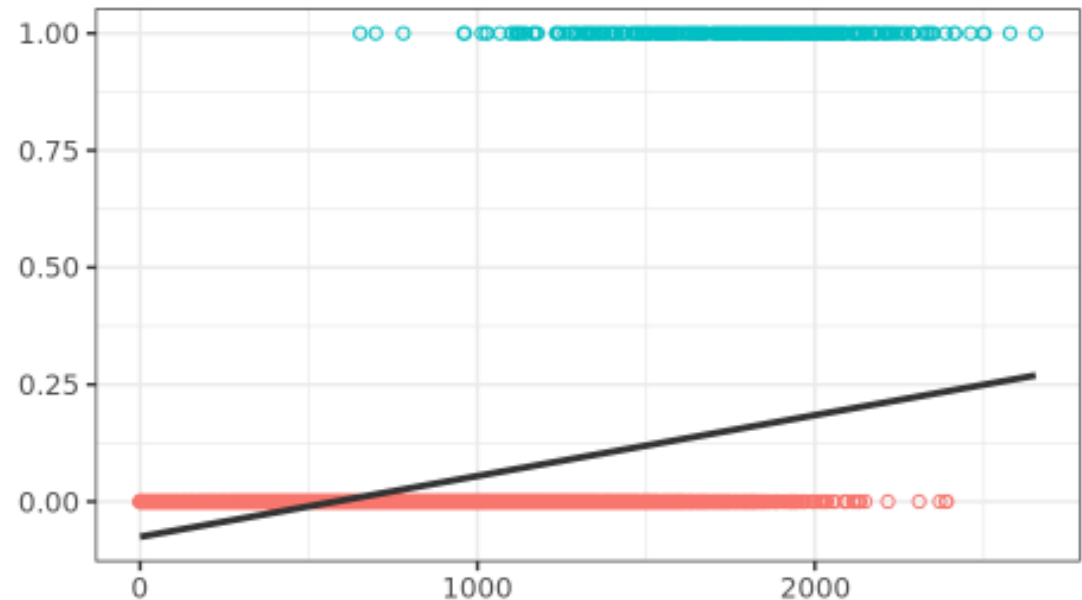
**Variable independiente**

**Variable cuantitativa:** Saldo en tarjeta de crédito

**Variable dependiente**

**Variable cualitativa (dummy):** Pago antes de fecha límite

Regresión lineal



La regresión lineal clasificaría todos los datos como 0.

# Un modelo más apropiado...

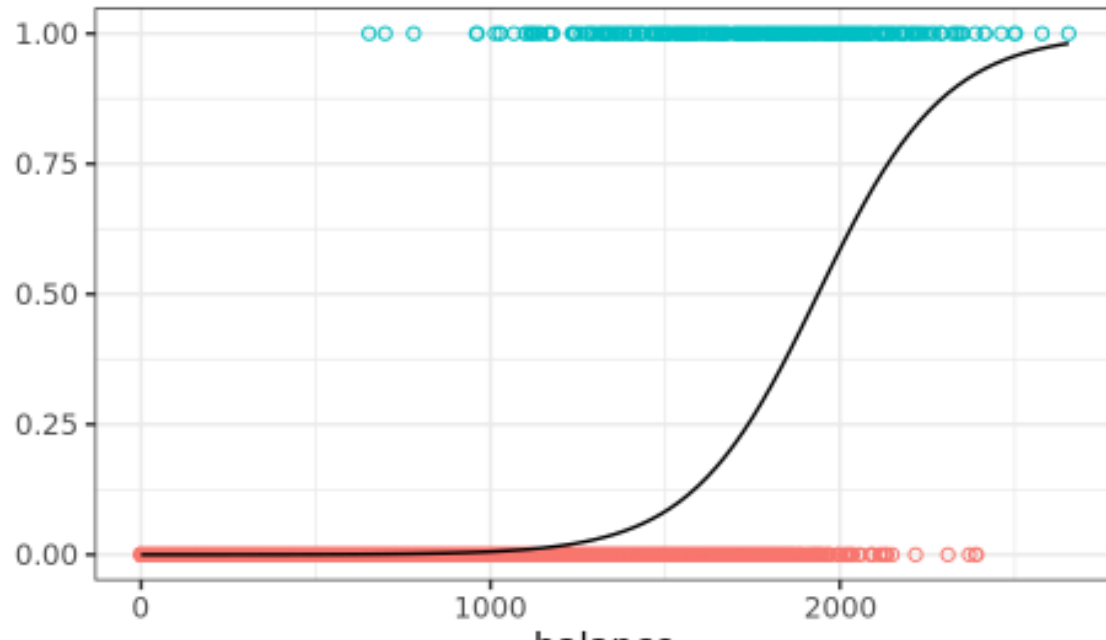
**Variable independiente**

**Variable cuantitativa:** Saldo en tarjeta de crédito

**Variable dependiente**

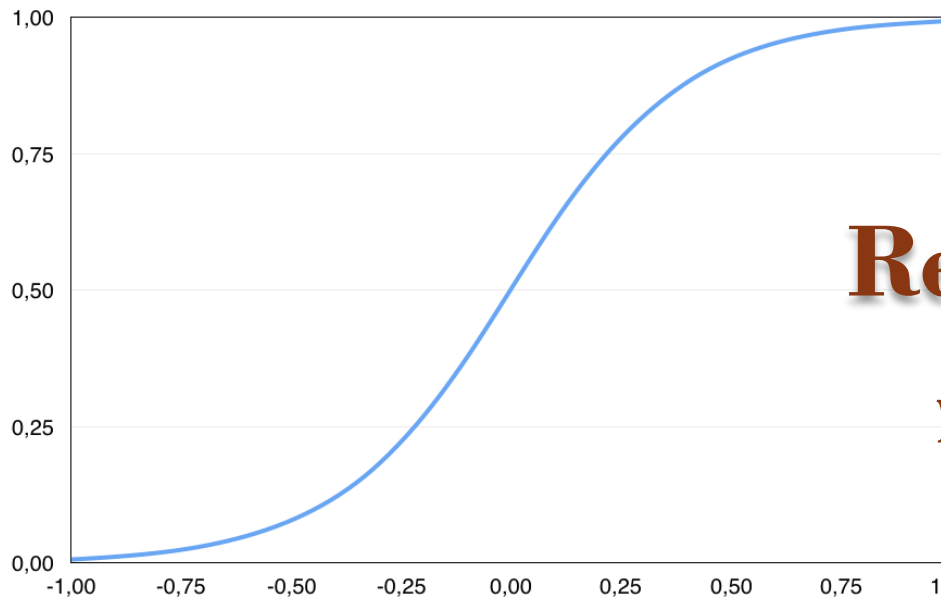
**Variable cualitativa (dummy):**  
Pago antes de fecha límite

**Regresión logística**



$$Y = \frac{p(X)}{1 - p(X)} = e^{\beta_0 + \beta_1 x}$$

momio

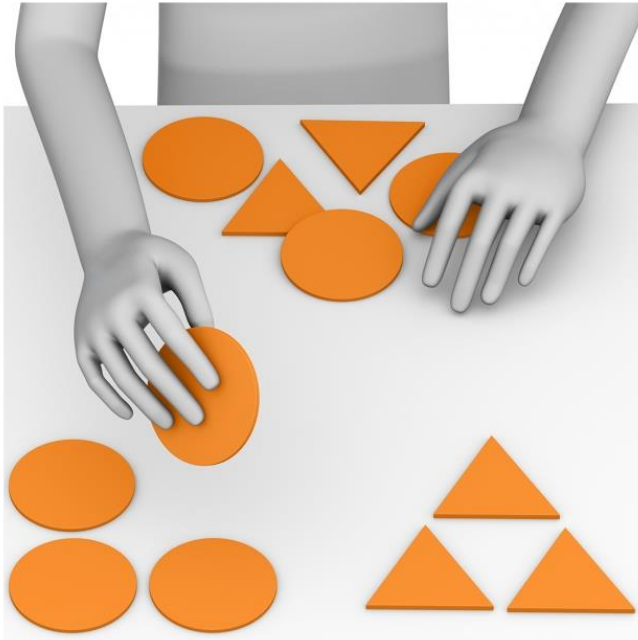


# Regresión logística

$$Y = \frac{p(X)}{1 - p(X)} = e^{\beta_0 + \beta_1 x}$$

- La variable dependiente asume dos valores o más valores: 0 y 1 (clasificación en un grupo u otro).
- Se usa para predecir cuál es el valor futuro de la variable dependiente basándose en variables independientes cuantitativas
- Modela el logaritmo del momio de pertenecer a cada grupo
- Es un método de clasificación muy usado por las pocas restricciones que tiene y la facilidad de su algoritmo
- Se puede usar con variable cualitativa con dos niveles (regresión logística simple) como con múltiples predictores (regresión logística múltiple)
- Se usa un conjunto de observaciones de entrenamiento para generar el clasificador y después se prueba en otras observaciones

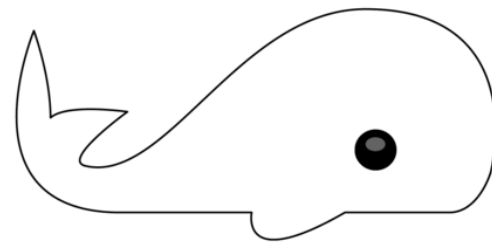
# ¿En qué consiste?



$$Y = \frac{p(X)}{1 - p(X)} = e^{\beta_0 + \beta_1 x}$$

- Se trata de estimar la posibilidad (momios o odds) de la clasificación, por lo que es un buen método apto para clasificar en uno o más grupos.
- Para una variable respuesta binaria (dos niveles) se crea una variable dummy con dos respuestas (0 y 1), donde, para clasificar a la variable codificada como 1,  $Y > 0.5$  (o dependiendo del interés).
- Si la variable cualitativa contara con más niveles, la asignación de números a los niveles no influiría en el modelo resultante cuando no hay un criterio que indique la forma en que se tienen que ordenar.

# Regresión logística simple



Si **Z** es una variable que indica la pertenencia a un cierto grupo, la regresión logística modela la Esperanza de que **Z** pertenezca a la categoría **k**, dados los valores de una variable predictora **X** (**X=x**).

$$E(Z)$$

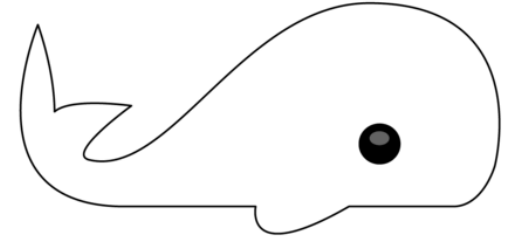
Para modelar la relación existente entre la  $E(Z)$  y **X** se usa:

$$E(Z) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

Como **E(Z)** sólo puede tomar dos valores, pertenece a la categoría **k** (1) o no pertenece (0), **E(Z)** se interpreta como una probabilidad para **X=x**:

$$p(X = x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

# Regresión logística simple



La expresión

$$p(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

A partir de esta fórmula se obtiene:

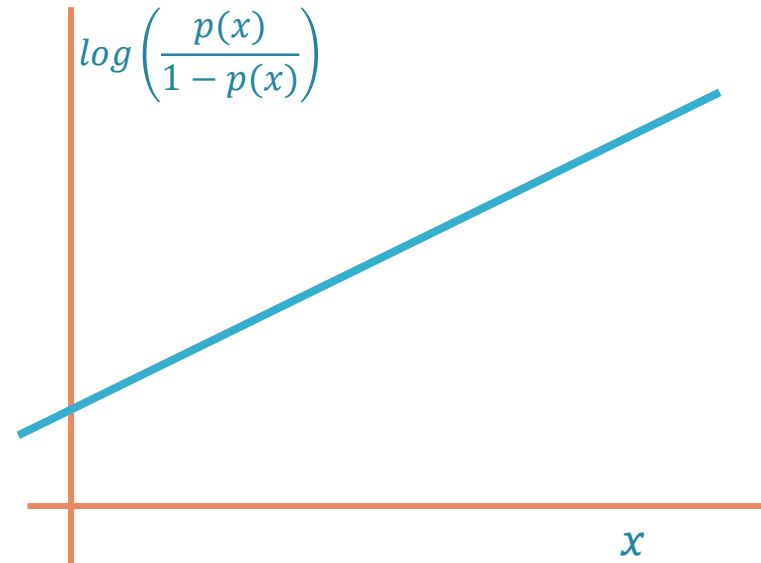
$$\text{momio} = \frac{p(x)}{1 - p(x)}$$

$$\frac{p(x)}{1 - p(x)} = e^{\beta_0 + \beta_1 x}$$

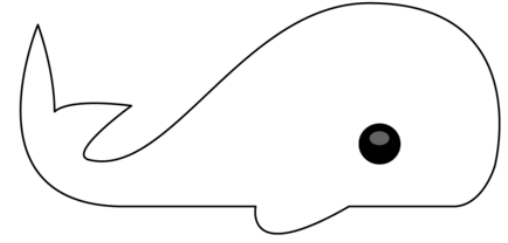
Introduciendo el logaritmo, queda:

$$\log \left( \frac{p(x)}{1 - p(x)} \right) = \beta_0 + \beta_1 x$$

Que se puede modelar con una regresión no lineal.

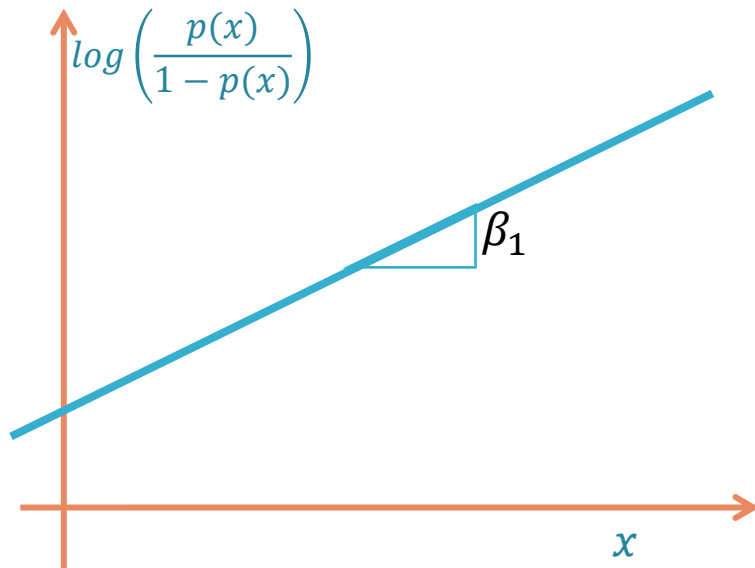


# Regresión logística simple



$$\log \left( \frac{p(x)}{1 - p(x)} \right) = \beta_0 + \beta_1 x$$

$$p(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$



**$\beta_1$**

- ✓  $\beta_1$  indica cuanto cambia el logaritmo de los momios cuando  $x$  se incrementa en una unidad.
- ✓ La cantidad con la que  $p(x)$  cambia debido a un cambio en  $x$  dependerá del valor actual de  $x$  (no es constante)
- ✓ Si  $\beta_1$  es positivo, un aumento en  $x$  provocará un aumento de  $p(x)$ .
- ✓ La intersección  $\beta_0$  corresponde con el resultado predicho para el nivel de referencia.



# Estimación de los coeficientes



$$\log \left( \frac{p(x)}{1 - p(x)} \right) = \beta_0 + \beta_1 x \quad p(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

Para estimar  $\beta_0$  y  $\beta_1$  se utiliza el método de máxima verosimilitud (máxima credibilidad de la estimación dados los datos de entrenamiento).

## *Prueba de hipótesis*

- Se realiza una prueba de hipótesis a partir del estadístico  $Z \sim N(0,1)$ .

- Hipótesis:

- $H_0: \beta_1 = 0$  (la probabilidad no depende de  $x$ )

- $H_1: \beta_1 \neq 0$  (la probabilidad depende de  $x$ )

- El estadístico  $Z$  asociado a  $\beta_1$  es:

$$z_0 = \frac{\hat{\beta}_1}{SE_{\hat{\beta}_1}}$$

Como  $x$  es categórica y se modela su probabilidad, el único requisito para la distribución muestral es un tamaño de muestra grande.

# Regresión logística múltiple



Cuando se tienen muchas variables:

$$p(x) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p}}$$

El modelo queda:

$$\log \left( \frac{p(x_1, x_2, \dots, x_p)}{1 - p(x_1, x_2, \dots, x_p)} \right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

- ✓ De nuevo usamos el método de máxima verosimilitud para estimar los coeficientes  $\beta_0, \beta_1, \dots, \beta_p$ . Cada coeficiente se interpreta manteniendo fijos al resto.
- ✓ También se elaboran pruebas de hipótesis para cada  $\beta_1$  para verificar la importancia de la variable  $x_i$  en el modelo.

# Condiciones del modelo



Las principales condiciones que este modelo requiere son:

- **Respuesta binaria:** La variable dependiente ha de ser binaria.
- **Independencia:** las observaciones han de ser independientes.
- **Multicolinealidad:** se requiere de muy poca a ninguna multicolinealidad entre los predictores (para regresión logística múltiple).
- **Linealidad** entre la variable independiente y el logaritmo natural de nomios.
- **Tamaño muestral:** como regla general, se requiere un mínimo de 10 casos con el resultado menos frecuente para cada variable independiente del modelo.

## Evaluación del modelo

Para evaluar si el modelo logístico es válido, se analiza tanto el modelo en su conjunto como los predictores que lo forman. El modelo se considerará útil si es capaz de mejorar la predicción de las observaciones respecto al modelo nulo sin predictores. Para ello se analiza la significancia de la diferencia (“Deviance”) de residuos entre ambos modelos (“Null deviance” y “Residual deviance”), con un estadístico que sigue la distribución chi-cuadrado con grados de libertad correspondientes a la diferencia de los grados de libertad de ambos modelos.


## Evaluación del modelo

Es importante también analizar el porcentaje de predicciones correctas además de los falsos positivos y falsos negativos que hace nuestro modelo para evaluar su potencial. Normalmente se usa un límite (threshold) de 0,5. Si la probabilidad predicha de que el valor del mercado sea positivo es mayor de 0.5, la observación se asignará al nivel 1 (“Up”), y si es menor se asignará al nivel 0 (“Down”).



# Ejemplo

En este ejemplo trabajaremos con el set de datos *Weekly*, que forma parte del paquete ISLR. Este set de datos contiene información sobre el rendimiento porcentual semanal del índice bursátil S&P 500 entre los años 1990 y 2010.



# En R

## Leyendo los datos:

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

## Análisis de los datos

```
head(Weekly)  
glimpse(Weekly)  
summary(Weekly)  
pairs(Weekly)  
cor(Weekly[, -9])
```

```
attach(Weekly)  
plot(Volume)
```

## Cálculo del modelo logístico

Modelo con todos los predictores, excluyendo  
“Today”

```
modelo.log.m <- glm(Direction ~ . -Today, data  
= Weekly, family = binomial)  
summary(modelo.log.m)  
contrasts(Direction)  
confint(object = modelo.log.m, level = 0.95)
```

## Modelo logístico con variables significativas

### # Gráfico de las variables significativas (boxplot), ejemplo: Lag2):

```
ggplot(data = Weekly, mapping = aes(x = Direction, y = Lag2)) +  
  geom_boxplot(aes(color = Direction)) +  
  geom_point(aes(color = Direction)) +  
  theme_bw() +  
  theme(legend.position = "null")
```

# Training: observaciones desde 1990 hasta 2008

```
datos.entrenamiento <- (Year < 2009)
```

# Test: observaciones de 2009 y 2010

```
datos.test <- Weekly[!datos.entrenamiento, ]
```

# Verifica:

```
nrow(datos.entrenamiento) + nrow(datos.test)
```

# Ajuste del modelo logístico con variables significativas

```
modelo.log.s <- glm(Direction ~ variables significativas, data = Weekly,  
  family = binomial, subset = datos.entrenamiento)  
summary(modelo.log.s)
```



## Representación gráfica del modelo

El modelo devuelve las predicciones del logaritmo de Odds. La predicción se debe convertir en probabilidad. Eso se logra con el comando 'predict' y el 'type="response"'.

```
# Vector con nuevos valores interpolados en el rango del predictor Lag2:  
nuevos_puntos <- seq(from = min(Weekly$Lag2), to = max(Weekly$Lag2),  
  by = 0.5)
```

```
# Predicción de los nuevos puntos según el modelo con el comando predict() se  
calcula la probabilidad de que la variable respuesta pertenezca al nivel de  
referencia (en este caso "Up")
```

```
predicciones <- predict(modelo.log.s, newdata = data.frame(Lag2 =  
  nuevos_puntos), se.fit = TRUE, type = "response")
```

# Representación gráfica del modelo

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
```

```
# Matriz de datos con los nuevos puntos y sus predicciones
```

```
  datos_curva <- data.frame(Lag2 = nuevos_puntos, probabilidad =  
    predicciones$fit, CI.inferior = CI_inferior, CI.superior = CI_superior)
```

## Representación gráfica del modelo

# Codificación 0,1 de la variable respuesta Direction

```
Weekly$Direction <- ifelse(Weekly$Direction == "Down", yes = 0, no = 1)
```

```
ggplot(Weekly, 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()
```

## Evaluación del modelo

# Chi cuadrada: Se evalúa la significancia del modelo con predictores con respecto al modelo nulo (“Residual deviance” vs “Null deviance”). Si valor p es menor que alfa será significativo.

```
anova(modelo.log.s, test ='Chisq')
```

## Evaluación del modelo

Cálculo de las predicciones correctas así como de los falsos negativos y positivos. Normalmente se usa un límite de 0.5.

```
# Cálculo de la probabilidad predicha por el modelo con los datos de test
prob.modelo <- predict(modelo.log.s, newdata = datos.test, type = "response")
```

```
# Vector de elementos "Down"
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 = Direction[!datos.entrenamiento]
```

```
# Matriz de confusión
matriz.confusion <- table(pred.modelo, Direction.0910)
matriz.confusión
library(vcd)
mosaic(matriz.confusion, shade = T, colorize = T,
       gp = gpar(fill = matrix(c("green3", "red2", "red2", "green3"), 2, 2)))
mean(pred.modelo == Direction.0910)
```