

# Regresión Avanzada

## Universidad Austral

**PhD. Débora Chan**  
Junio - Julio de 2023

Facultad de Ingeniería

# Organización

- 1 Modelos GAM
- 2 Modelos GAMLSS
- 3 Modelos Lineales Generalizados
- 4 Regresión Logística
- 5 Modelo de Regresión Logística
- 6 Significado de los Coeficientes del Modelo de Regresión logística Simple
- 7 Tablas de Clasificación y Curvas ROC

# Motivación

- ☞ Predecir el valor esperado de la variable  $Y$  en función de uno o varios predictores, condicional a éstos, es un problema que puede enfocarse con múltiples métodos de estadística y aprendizaje automático.
- ☞ Algunos modelos sólo consideran una relación lineal entre las predictoras (LM, GLM) y la respuesta, otros permiten incorporar relaciones no lineales, incluso interacciones (SVM, Random Forest, Boosting).
- ☞ La mayoría de los modelos de regresión sólo modelan la relación de la media de la respuesta, asumiendo que las restantes características como la dispersión y la simetría son constantes, pero esto constituye una limitación importante.

# Modelos GAM (Generalizado Aditivo)

## Objetivo

Los modelos GAM nos permiten ajustar una función no lineal a cada  $X_j$ , por lo que no es necesario incorporar transformaciones a cada variable manualmente. Los ajustes no lineales que nos permite el modelo GAM pueden llegar a conseguir predicciones más precisas para la variable respuesta.

## Los Datos

El set de datos `rent` del paquete `gamlss.data` contiene información sobre la renta de 1969 viviendas situadas en Munich en el año 1993. El objetivo es obtener un modelo capaz de predecir el valor del alquiler.

## Ejemplo Alquiler: Las variables

**R:** precio del alquiler

**FI:** metros cuadrados de la vivienda.

**A:** año de construcción.

**Sp:** calidad del barrio de la vivienda superior la media (1) o no (0).

**Sm:** si la calidad del barrio donde está situada la vivienda es inferior la media (1) o no (0).

**B:** si tiene cuarto de baño (1) o no (0).

**H:** si tiene calefacción central (1) o no (0).

**L:** equipamiento de la cocina por encima de la media (1) o no (0).

**loc:** combinación de Sp y Sm indicando si la calidad del barrio donde está situada la vivienda es inferior (1), igual (2) o superior (3) a la media

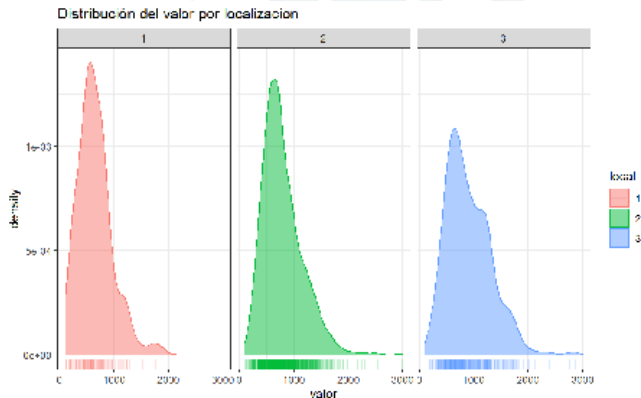
```
library(gamlss)  
library(tidyverse)  
library(ggpubr)  
library(skimr)
```

```
data('rent')
```

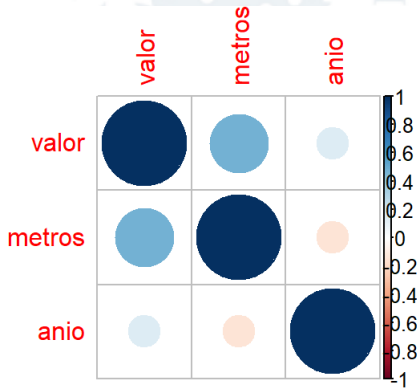
```
datos <- rent %>% select(R, FI, A, H, loc)  
datos %>% setNames(c('valor', 'metros', 'anio', 'calef', 'local'))
```

Facultad de Ingeniería

# Exploración de los Datos



# Exploración de los Datos





## Modelo OLS

```
mod_OLS <- gamlss( formula = valor  metros + anio + calef +  
local,  
family = NO, data = datos, trace = FALSE)  
summary(mod_OLS)
```

Call: gamlss(formula = valor metros + anio + calef + local, family  
= NO, data = datos, trace = FALSE)  
Fitting method: RS()  
Mu link function: identity

## Modelo OLS

Mu Coefficients:

|              | Estimate | Std. Error | t value | Pr(> t ) |
|--------------|----------|------------|---------|----------|
| X.Intercept. | -2775.04 | 470.14     | -5.90   | 0.00     |
| metros       | 8.84     | 0.34       | 26.23   | 0.00     |
| anio         | 1.48     | 0.24       | 6.21    | 0.00     |
| calef1       | -204.76  | 18.99      | -10.78  | 0.00     |
| local2       | 134.05   | 25.14      | 5.33    | 0.00     |
| local3       | 209.58   | 27.13      | 7.73    | 0.00     |

Facultad de Ingeniería

# Modelo OLS

Sigma link function: log

Sigma Coefficients:

Estimate Std. Error t value Pr(> |t|)

(Intercept) 5.73165 0.01594 359.7 < 2e - 16 \*\*\*

—

No. of observations in the fit: 1969

Degrees of Freedom for the fit: 7

Residual Deg. of Freedom: 1962

at cycle: 2

Global Deviance: 28159

AIC: 28173

SBC: 28212.1

## Observaciones

La función link para estimar la varianza es log

Recordemos que la varianza es no negativa.

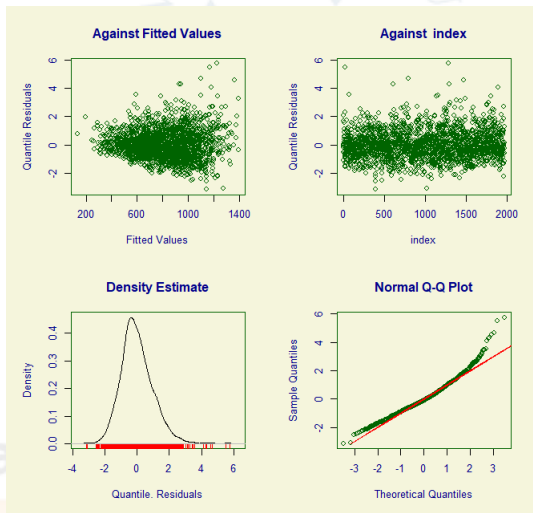
El valor estimado de la varianza se obtiene aplicando exponencial al resultado del modelo, es decir antitransformando.

$$\exp(5.73165) = 308.5$$

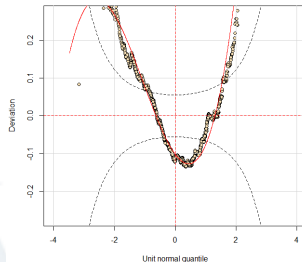
### Los gráficos worm

Son una forma visual para evaluar la calidad de ajuste del modelo. Se parecen al qqplot. Idealmente los residuos deberían ser nulos, cercanos a la línea horizontal. También el gráfico provee dos líneas punteadas que son el IC del 95% de confianza para los residuos. Por lo cual si más del 5% de los residuos caen fuera de este intervalo el modelo no ajusta correctamente.

# Diagnóstico (`plot(mod_OLS)`)



## Gráficos worm (wp(mod\_OLS))



La varianza de los errores aumenta con la media, la normalidad no se satisface, por lo cual el OLS no parece un modelo adecuado para este caso.

## Modelo GAM

Estos modelos permiten aplicar funciones (lineales o no lineales) a cada uno de los predictores. El paquete `gamlss` provee varias alternativas de suavizado que suelen dar buenos resultados. Vamos a usar para este caso P-splines (`pb()`). Sólo tiene sentido aplicar este suavizado para variables continuas.

```
mod_GAM <- gamlss( formula = valor  metros + anio + calef +  
local,  
family = GA, data = datos, trace = FALSE)  
summary(mod_GAM)
```

## Salida GAM

```
gamlss(formula = valor ~ pb(metros) + pb(anio) + calef +
local, family = GA, data = datos, trace = FALSE)
```

Fitting method: RS()

|              | Estimate | Std. Error | t value | Pr(> t ) |
|--------------|----------|------------|---------|----------|
| X.Intercept. | 3.09     | 0.57       | 5.42    | 0.00     |
| pb.metros.   | 0.01     | 0.00       | 25.57   | 0.00     |
| pb.anio.     | 0.00     | 0.00       | 4.86    | 0.00     |
| calef1       | -0.30    | 0.02       | -13.32  | 0.00     |
| local2       | 0.19     | 0.03       | 6.30    | 0.00     |
| local3       | 0.27     | 0.03       | 8.42    | 0.00     |



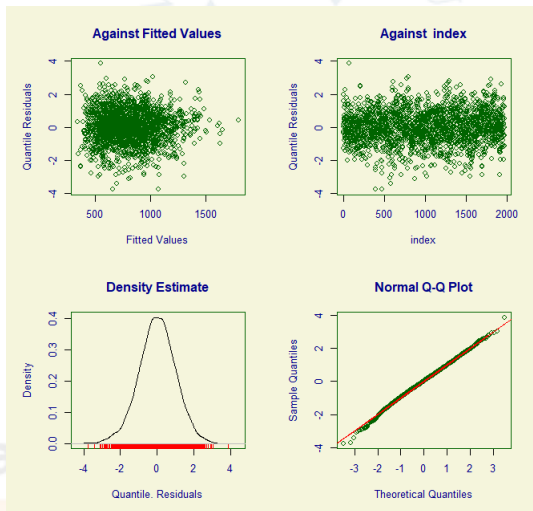
## Salida GAM

No. of observations in the fit: 1969  
Degrees of Freedom for the fit: 11.21547  
Residual Deg. of Freedom: 1957.785  
at cycle: 3

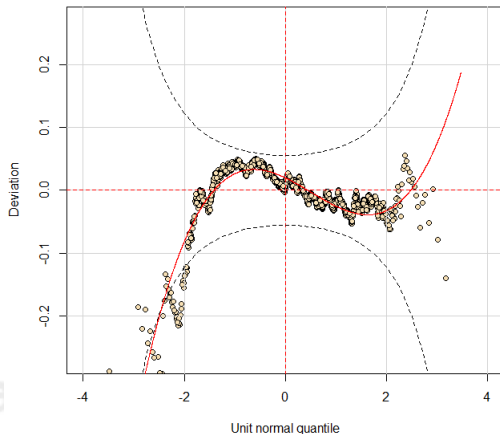
Global Deviance: 27683.22  
AIC: 27705.65  
SBC: 27768.29

Facultad de Ingeniería

# Análisis Diagnóstico GAM



# Gráfico worm GAM



# Interpretación del Modelo GAM

## Cuidado!

Los errores de este modelo sólo contemplan la parte lineal, no las funciones suavizadas(smooth: metros y anio); es decir que los errores asumen que las funciones smooth son fijas y no consideran la variabilidad introducida por ellas.

Existe sin embargo un cálculo de AIC y significacion estadística de los predictores (todos!) que se plantea cuánto se pierde al eliminar secuencialmente a cada una de las variables del modelo. (drop1())

```
drop1(mod_GAM, parallel = 'multicore', ncpus = 4)
```

## Significación Estadística de los predictores (GAM)

Single term deletions for mu

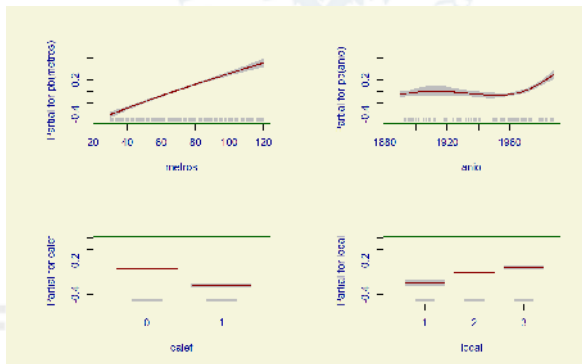
Model: valor  $\sim$  pb(metros) + pb(anio) + calef + local

|            | Df | AIC      | LRT    | Pr(Chi) |
|------------|----|----------|--------|---------|
| <none>     |    | 27705.65 |        |         |
| pb(metros) | 1  | 28261.31 | 558.59 | 0.00    |
| pb(anio)   | 4  | 27798.15 | 101.14 | 0.00    |
| calef      | 2  | 27862.35 | 160.39 | 0.00    |
| local      | 2  | 27769.60 | 68.02  | 0.00    |

Al añadir términos no lineales en un modelo, se complejiza su interpretabilidad, sin embargo puede ilustrarse el efecto de cada predictor. (`term.plot()`)

## Efectos individuales de los predictores (GAM)

```
term.plot(mod_GAM, , pages = 1, ask = FALSE, rug = TRUE)
```



# Interpretación de los Efectos

## Los efectos a partir de las gráficas




Estas gráficas representan la relación entre  $\eta = \log(\mu)$  con cada uno de los predictores:

- 😊 El efecto de la superficie es lineal con el valor de la vivienda, al menos en el rango estudiado.
- 😊 El efecto del año de construcción de la misma no se aprecia lineal, aumenta a partir de 1960 aproximadamente; antes de ello parece constante.
- 😊 La localización en un buen vecindario eleva el valor del alquiler.
- 😊 La calefacción(central?) disminuye el valor del alquiler.

# De todo lo observado surge...

## La Necesidad

De modelos que:

-  Sean capaces de modelar relaciones complejas entre varios predictores, incluyendo relaciones no lineales.
-  Sean capaces de modelar explícitamente la varianza en función de los predictores, ya que la misma no siempre es constante.
-  Sean capaces de modelar distribuciones con una marcada asimetría.

Facultad de Ingeniería



# Organización

- 1 Modelos GAM
- 2 Modelos GAMLSS
- 3 Modelos Lineales Generalizados
- 4 Regresión Logística
- 5 Modelo de Regresión Logística
- 6 Significado de los Coeficientes del Modelo de Regresión logística Simple
- 7 Tablas de Clasificación y Curvas ROC

# Modelos GAMLSS: Motivación

## Generalizados Aditivos de Posición Escala y Forma

### Qué modelábamos?

Hasta el momento, se ha modelado únicamente el valor esperado de la respuesta  $\mu$  en función de los predictores, asumiendo que la escala es constante y puede ser estimada a partir de la media.

### Qué podríamos modelar?

Una forma de mejorar el modelo es modelando también su parámetro de escala en función de los predictores. Esto resulta particularmente útil cuando los datos muestran heterocedasticidad (su varianza no es constante).

# Ajuste GAMLSS

Se repite el ajuste, pero esta vez modelando sus dos parámetros en función de los predictores.  $Y \sim GA(\mu, \sigma)$



$$\eta_1 = g_1(\mu) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p)$$



$$\eta_2 = g_2(\sigma) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p)$$

## El Ajuste (GAMLSS)

```
mod_GAMLSS <- gamlss(formula = valor ~ pb(metros)+
pb(anio)+calef+local, sigma.formula = ~ pb(metros)+pb(anio)+
calef+local, family = GA, data = datos, trace = FALSE )
summary(mod_GAMLSS)
```

### Mu Coefficients:

|              | Estimate | Std. Error | t value | Pr(> t ) |
|--------------|----------|------------|---------|----------|
| X.Intercept. | 2.88     | 0.58       | 4.95    | 0.00     |
| pb.metros.   | 0.01     | 0.00       | 29.12   | 0.00     |
| pb.anio.     | 0.00     | 0.00       | 5.07    | 0.00     |
| calef1       | -0.29    | 0.02       | -12.02  | 0.00     |
| local2       | 0.19     | 0.03       | 6.00    | 0.00     |
| local3       | 0.27     | 0.03       | 8.06    | 0.00     |

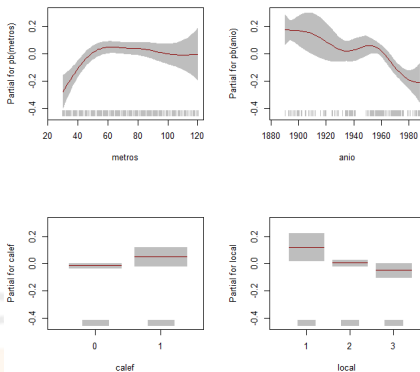
## El Ajuste (GAMLSS)

|                | Estimate | Std. Error | t value | Pr(> t ) |
|----------------|----------|------------|---------|----------|
| X.Intercept..1 | 5.92     | 0.86       | 6.85    | 0.00     |
| pb.metros..1   | 0.00     | 0.00       | 2.49    | 0.01     |
| pb.anio..1     | -0.00    | 0.00       | -8.19   | 0.00     |
| calef1.1       | 0.07     | 0.04       | 1.59    | 0.11     |
| local2.1       | -0.12    | 0.06       | -2.06   | 0.04     |
| local3.1       | -0.17    | 0.06       | -2.80   | 0.01     |

Facultad de Ingeniería

# Efecto individuales de los Predictores (GAMLSS)

```
term.plot(mod_GAMLSS, parameter = 'sigma', ask = FALSE, rug = TRUE)
```



## Contribución de cada uno de los predictores (GAMLSS)

```
drop1(mod_GAMLSS, parameter = 'sigma', parallel = 'multicore',
ncpus = 4)
```

Single term deletions for sigma

|            | Df | AIC      | LRT   | Pr(Chi) |
|------------|----|----------|-------|---------|
| <none>     |    | 27614.78 |       |         |
| pb(metros) | 4  | 27631.41 | 24.68 | 0.00    |
| pb(anio)   | 4  | 27659.19 | 52.17 | 0.00    |
| calef      | 1  | 27614.88 | 1.87  | 0.15    |
| local      | 2  | 27618.74 | 8.04  | 0.02    |

## Reformulación modelos GAMLSS(1)

Del gráfico y la salida del drop surge que debería eliminarse el predictor calefacción de la estimación del parámetro  $\sigma$ , dado que no es significativo.

```
mod_GAMLSS1 <- gamlss( formula = valor ~
  pb(metros)+pb(anio)+calef+local, sigma.formula = ~
  pb(metros)+pb(anio)+local, family = GA, data = datos, trace =
  FALSE ) summary(mod_GAMLSS1)
```

Facultad de Ingeniería



## Estimación de los efectos para $\mu$

| Mu           | link | function: |            | log<br>t value | Mu<br>Pr(> t ) | Coefficients: |
|--------------|------|-----------|------------|----------------|----------------|---------------|
|              |      | Estimate  | Std. Error |                |                |               |
| X.Intercept. |      | 2.83      | 0.58       | 4.83           | 0.00           |               |
| pb.metros.   |      | 0.01      | 0.00       | 29.15          | 0.00           |               |
| pb.anio.     |      | 0.00      | 0.00       | 5.14           | 0.00           |               |
| calef1       |      | -0.29     | 0.02       | -12.35         | 0.00           |               |
| local2       |      | 0.20      | 0.03       | 6.07           | 0.00           |               |
| local3       |      | 0.28      | 0.03       | 8.13           | 0.00           |               |

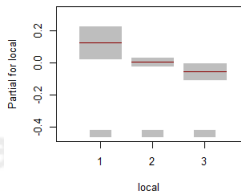
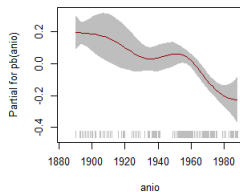
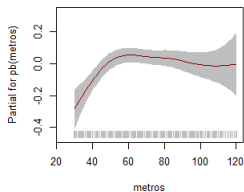
Facultad de Ingeniería

# Estimación de los efectos para $\sigma$ (GAMLSS1)

| Sigma          | link | function: |            | log   | Sigma | Coefficients: |
|----------------|------|-----------|------------|-------|-------|---------------|
|                |      | Estimate  | Std. Error |       |       | Pr(> t )      |
| X.Intercept..1 |      | 6.63      | 0.81       | 8.22  |       | 0.00          |
| pb.metros..1   |      | 0.00      | 0.00       | 2.29  |       | 0.02          |
| pb.anio..1     |      | -0.00     | 0.00       | -9.59 |       | 0.00          |
| local2.1       |      | -0.12     | 0.06       | -2.12 |       | 0.03          |
| local3.1       |      | -0.18     | 0.06       | -2.94 |       | 0.00          |

Facultad de Ingeniería

# Efecto individual de los predictores (GAMLSS1)



## Contribución individual de las variables

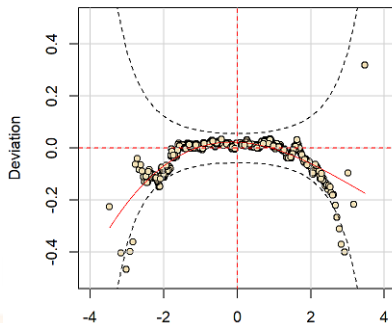
```
drop1(mod_GAMLSS1, parameter = 'sigma', parallel = 'multicore',
ncpus = 4)
```

Single term deletions for sigma

|            | Df | AIC      | LRT   | Pr(Chi) |
|------------|----|----------|-------|---------|
| <none>     |    | 27614.88 |       |         |
| pb(metros) | 4  | 27631.53 | 24.76 | 0.00    |
| pb(anio)   | 3  | 27671.84 | 63.78 | 0.00    |
| local      | 2  | 27619.63 | 8.76  | 0.01    |

## Worm plot de los residuos

```
wp(mod_GAMLSS1, ylim.all = 0.5)
```



## Observaciones

La inspección visual del wormplot indica que este modelo tiene los residuos dentro del rango de variación aceptable, sin embargo se aprecia una estructura de U invertida. Esto puede deberse a que el modelo no logre captar bien la asimetría de la distribución de los valores.

Comparamos los modelos ajustados:

`GAIC(mod_OLS, mod_GAMLSS1, mod_GAM)`

Facultad de Ingeniería

## Comparación de los Modelos

|             | df    | AIC      |
|-------------|-------|----------|
| mod_GAMLSS1 | 21.37 | 27614.88 |
| mod_GAM     | 11.22 | 27705.65 |
| mod_OLS     | 7.00  | 28173.00 |

De acuerdo con el criterio GAIC, el modelo GAMLSS es el que mejor explica la relación con el valor utilizando los mismos predictores.

Facultad de Ingeniería

# Organización

- 1 Modelos GAM
- 2 Modelos GAMLSS
- 3 Modelos Lineales Generalizados**
- 4 Regresión Logística
- 5 Modelo de Regresión Logística
- 6 Significado de los Coeficientes del Modelo de Regresión logística Simple
- 7 Tablas de Clasificación y Curvas ROC



# Componentes del GLM

| Componente Aleatoria                  | Componente de Enlace                                  | Componente Sistemática                       |
|---------------------------------------|---|--|
| Distribución de la variable respuesta | Aleatoria-Sistemática                                 | Función Lineal de las Variables Explicativas |
| normal, poisson, binomial, etc        | función logística, logaritmo, función identidad, etc. | Edad, peso, categoría, etc                   |

## Resumiendo

Los supuestos del GLM son:

$Y_i$  son las observaciones aleatorias independientes cuya distribución de probabilidad pertenece a una familia exponencial.

Las variables predictoras  $X_1, X_2, \dots, X_p$  proporcionan un conjunto de predictores lineales:

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi}$$

existe una función  $g()$  de enlace o link que establece que:

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi}$$

Es decir que modelamos una función de la esperanza o media de variable respuesta. En el caso del modelo lineal clásico esa

## Ejemplo binomial

### Enlace Logit

En este caso el objeto de modelización es una probabilidad de éxito.

$$g(\mu_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 x_i$$

donde  $\pi_i = \mu_i$ .

### Ejemplo

Se quiere estimar la probabilidad de insomnio en función de la edad de los pacientes ( $x_i$  es la edad del paciente)

## Ejemplo Poisson

### Enlace Log

En este caso el valor esperado de la respuesta es  $E(Y_i) = \mu_i$  donde  $\mu_i$  es la tasa de ocurrencia y la función de link más frecuente es el logaritmo.

$$g(\mu_i) = \log(\mu_i)$$

Es decir que el objetivo de modelización es el logaritmo del parámetro de interés.

### Ejemplo

El objetivo del estudio es relacionar el número esperado de pacientes con alergia en cada zona utilizando las siguientes variables explicativas: número de afiliados (en mil), ingreso promedio anual (en mil USD), edad media de los afiliados (en años), densidad de ciertos árboles de la región.

## Ejemplo Gamma

### Enlace Inversa

$$g(\mu_i) = 1/\mu_i$$

### Ejemplo

Por ejemplo si se realiza un ensayo clínico donde se registra el tiempo de supervivencia (en semanas) para pacientes de leucemia y su correspondiente conteo inicial de células blancas en la sangre (en escala log).

Se desea estimar el tiempo de supervivencia  $Y$  en función del conteo inicial de células blancas  $x_i$ .

Una distribución usual en la modelización de tiempos de supervivencia es la exponencial, que es un caso particular de la distribución Gamma.

## Estimación de los Coeficientes

### En pocas palabras

Los procedimientos de estimación de los parámetros utilizados para OLS no son adecuados ya que la estructura lineal se asume sobre la transformación de la media de la respuesta y no sobre los valores de la respuesta.

Se utiliza el método de máxima verosimilitud para la estimación de los parámetros que tiene en este caso estructura iterativa, que finalizan cuando se alcanzan las condiciones de convergencia del método.

En la mayoría de situaciones experimentales dichos métodos convergen en pocas iteraciones y proporcionan las estimaciones del parámetro del modelo.

También es posible obtener además intervalos de confianza para los parámetros del modelo.

## Ejemplo Poisson

### El problema

Se está conduciendo un experimento que se propone analizar el daño causado en las embarcaciones debido a la exposición al oleaje.

Como parte del proyecto se recolectaron los siguientes datos:

**type:** tipo de embarcación (codificado con las letra A - E)

**year:** año de construcción de la embarcación

**period:** periodo de operación (1960-74, 75-79)

**service:** acumulado de meses en servicio

**incidents:** número de incidentes por daño



## Ejemplo Poisson: Ajuste de prime modelo

```
# Utilizamos los datos ships de la biblioteca MASS
library(MASS)
data(ships)
datos_ships <- ships %>% filter(service != 0)
# eliminamos los registros que no tuvieron servicios
fit <- glm(incidents ~., datos_ships, family = poisson)
summary(fit)
```

Acá utilizamos todas las variables disponibles en la base.



## Ejemplo Poisson

|             | Estimate | Std. Error | z value | Pr(> z ) |
|-------------|----------|------------|---------|----------|
| (Intercept) | -4.3871  | 1.1512     | -3.81   | 0.0001   |
| typeB       | 0.9622   | 0.2058     | 4.67    | 0.0000   |
| typeC       | -1.2117  | 0.3274     | -3.70   | 0.0002   |
| typeD       | -0.8652  | 0.2875     | -3.01   | 0.0026   |
| typeE       | -0.1105  | 0.2350     | -0.47   | 0.6382   |
| year        | 0.0528   | 0.0138     | 3.83    | 0.0001   |
| period      | 0.0364   | 0.0092     | 3.94    | 0.0001   |
| service     | 0.0000   | 0.0000     | 6.79    | 0.0000   |

Null deviance: 614.54 on 33 degrees of freedom

Residual deviance: 120.56 on 26 degrees of freedom

AIC: 234.43

Number of Fisher Scoring iterations: 5

## Ejemplo Poisson: Bondad de Ajuste

```
coef(fit) # podemos pedir solamente los coeficientes del modelo
# comparamos la deviance del modelo con la del modelo nulo, para
# ver si es significativo
dev <- fit$deviance
nullDev <- fit$null.deviance
modelChi <- nullDev - dev
modelChi # calculamos la diferencia de deviances
chidf <- fit$df.null - fit$df.residual # calculamos los grados de lib-
# ertad de la diferencia
chisq.prob <- 1 - pchisq(modelChi, chidf)
chisq.prob
```

## Ejemplo Poisson

```
dev [1] 120.5642  
nullDev [1] 614.5393  
modelChi [1] 493.9752  
chisq.prob [1] 0
```

Nuestro modelo es estadísticamente significativo. Además en comparación con un barco tipo A, los tipo B sufren más incidentes dado que el coeficiente es positivo, y además también a comparación de los tipo A, los barcos tipo E, D y C son sufren menos incidentes, siendo 'más seguros' los del tipo C. También notamos que conforme aumentan el año de construcción, los barcos tienen menos incidentes. Finalmente vemos que durante el primer periodo de operación se sufren menos incidentes que en el segundo, es decir, si el periodo de operación es más reciente, más incidentes tendrá.

## Ejemplo Poisson: Bondad de Ajuste

### Observaciones

Sin embargo si analizamos un poco más veremos que la deviance obtenida es de 146.238 con 33 grados de libertad.

Si realizamos el cociente de ambas cantidades nos da 4.43 que está alejado de 1, lo que nos indica que el modelo no ajusta bien los datos.

Utilizando

$$R^2_{pseudo} = \frac{\text{nulldeviance} - \text{residualdeviance}}{\text{nulldeviance}} = 0.59$$

Se interpreta que el 59% de la variabilidad del número de accidentes es lo que el modelo logra explicar.

## Ejemplo Poisson: Selección de Modelos por AIC

stepAIC(fit)

|             | Estimate | Std. Error | z value | Pr(> z ) |
|-------------|----------|------------|---------|----------|
| (Intercept) | -4.3871  | 1.1512     | -3.81   | 0.0001   |
| typeB       | 0.9622   | 0.2058     | 4.67    | 0.0000   |
| typeC       | -1.2117  | 0.3274     | -3.70   | 0.0002   |
| typeD       | -0.8652  | 0.2875     | -3.01   | 0.0026   |
| typeE       | -0.1105  | 0.2350     | -0.47   | 0.6382   |
| year        | 0.0528   | 0.0138     | 3.83    | 0.0001   |
| period      | 0.0364   | 0.0092     | 3.94    | 0.0001   |
| service     | 0.0000   | 0.0000     | 6.79    | 0.0000   |

## Ejemplo Poisson

Por último, ajustamos un modelo sugerido:

```
fit_hint <- glm(incidents ~ offset(log(service)) + type + year + period,  
data = datos_ships, family = poisson)  
  
summary(fit_hint)
```

El uso de `offset()` es buena idea pues nos interesa más la tasa ya que no es lo mismo tener 10 incidentes en 3 meses de servicio que tener 10 incidentes en 300 meses de servicio. Estamos modelando la tasa de incidentes por service:

$$\lambda_i = \frac{\text{incidentes}_i}{\text{service}_i}$$

## Ejemplo Poisson

|             | Estimate | Std. Error | z value | Pr(> z ) |
|-------------|----------|------------|---------|----------|
| (Intercept) | -10.0791 | 0.8761     | -11.50  | 0.0000   |
| typeB       | -0.5461  | 0.1784     | -3.06   | 0.0022   |
| typeC       | -0.6326  | 0.3295     | -1.92   | 0.0549   |
| typeD       | -0.2323  | 0.2880     | -0.81   | 0.4200   |
| typeE       | 0.4060   | 0.2349     | 1.73    | 0.0840   |
| year        | 0.0422   | 0.0128     | 3.29    | 0.0010   |
| period      | 0.0237   | 0.0081     | 2.93    | 0.0034   |

Facultad de Ingeniería

# Organización

- 1 Modelos GAM
- 2 Modelos GAMLSS
- 3 Modelos Lineales Generalizados
- 4 Regresión Logística**
- 5 Modelo de Regresión Logística
- 6 Significado de los Coeficientes del Modelo de Regresión logística Simple
- 7 Tablas de Clasificación y Curvas ROC



## Cuándo se Aplica?

### Pertinencia

El modelo de regresión logística es muy pertinente para trabajos biológicos, epidemiológicos, de análisis de mercado, de clasificación de clientes, etc. Se utiliza para modelar repuestas dicotómicas (presencia o ausencia de una condición) en función de un conjunto de variables (covariables) que posiblemente afectan la respuesta.

|                            | Lineal          | Logística                       |
|----------------------------|-----------------|---------------------------------|
| <b>Variable Respuesta</b>  | Y: continua     | Y: binaria ( 0 / 1 )            |
| <b>Valor ajustado</b>      | Nivel de Y      | Probabilidad ( $Y = 1$ )        |
| <b>Interp de los parám</b> | Diferencia de Y | Cociente de odds que<br>$Y = 1$ |

## Objetivos Fundamentales

La Regresión Logística Simple, desarrollada por David Cox en 1958, es un método de regresión que permite estimar la probabilidad de una variable cualitativa binaria en función de una variable cuantitativa. Una de las principales aplicaciones de la regresión logística es la de clasificación binaria, en el que las observaciones se clasifican en un grupo u otro dependiendo del valor que tome la variable empleada como predictor. Por ejemplo, clasificar a un individuo desconocido como hombre o mujer en función del tamaño de la mandíbula.

Aunque la regresión logística permite clasificar, se trata de un modelo de regresión que modela el logaritmo de la probabilidad de pertenecer a cada grupo. La asignación final es función de las probabilidades predichas.

# Logística vs Lineal

## Similitudes y Diferencias entre Logística y Lineal

- (a) Ambas son flexibles ya que admiten variables cuantitativas, cualitativas o categóricas como predictoras.
- (b) La interpretación de ambas es directa.
- (c) La Regresión logística pertenece a la familia de Modelos Lineales Generalizados.
- (d) Ambas tienen muchas similitudes, pero los procedimientos matemáticos subyacentes son diferentes.
- (e) En regresión lineal la media de la variable respuesta ( $\mu$ ) se modela mediante una combinación lineal de variables explicativas ( $\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k$ ).
- (f) En los modelos lineales generalizados se modela una transformación de la media de la variable respuesta ( $g(\mu)$ ) como una combinación

# Transformación Logística

## NO SE TRANSFORMAN LOS DATOS

En el modelo de regresión logística, la media ( $p$ ) de una variable respuesta con distribución Binomial( $1, p$ ) se transforma mediante la “transformación logística”:

$$g(p) = \frac{p}{1 - p}$$

La esperanza del Modelo Logístico es:

$$E(Y) = P(Y = 1) = p$$

Es decir:

$$\ln \left( \frac{P(Y = 1/x_1, x_2, \dots, x_k)}{1 - P(Y = 1/x_1, x_2, \dots, x_k)} \right) = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k$$

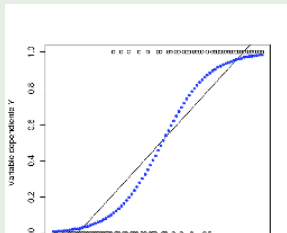
# Observaciones

## Debemos apreciar

Por qué en algunos casos el modelo logístico es adecuado y el lineal no.

Qué ventaja ofrece la curva logística?

Las ventajas del modelo logístico que incluyen una interpretación sencilla, una estimación de los efectos y un modelo de clasificación simultáneamente.



## Ejemplo: Diabetes

Consideremos los datos del archivo diabetes.xlsx

Corresponden a 145 adultos no obesos que participaron en un estudio para investigar factores asociados a diabetes. Utilizaremos inicialmente la variable SSPG “steady state plasma glucose” (una medida de la resistencia a la insulina) como variable explicativa y como variable respuesta DIABET (DIABET =1 indica que el paciente es diabético). En la tabla de la siguiente diapositiva se aprecia que a medida la proporción de diabéticos aumenta cuando aumenta el valor de la variable explicativa (SSPG).

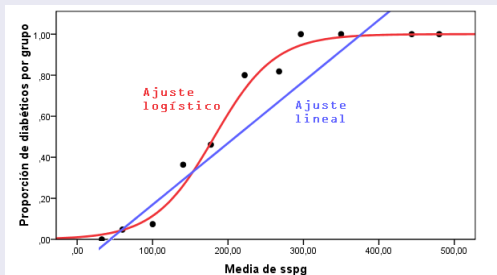


## SSPG y Proporción de Diabéticos

| grupos  | media SSPG | propor DIABET |
|---------|------------|---------------|
| 0-40    | 32.67      | 0             |
| 40-80   | 60,10      | 0.05          |
| 80-120  | 100.15     | 0.077         |
| 120-160 | 140.50     | 0.36          |
| 160-200 | 177.23     | 0.46          |
| 200-240 | 222.07     | 0.80          |
| 240-280 | 267.73     | 0.82          |
| 280-320 | 296.60     | 1.00          |
| 320-400 | 349.91     | 1.00          |
| 400-480 | 443.60     | 1.00          |

# Observaciones

## Ajuste de la Variable Proporción



Si ajustamos en forma errónea un modelo de regresión lineal a este conjunto de datos obtendremos la recta ajustada que muestra el diagrama de dispersión de la figura ( azul).



# Organización

- 1 Modelos GAM
- 2 Modelos GAMLSS
- 3 Modelos Lineales Generalizados
- 4 Regresión Logística
- 5 **Modelo de Regresión Logística**
  - Una variable continua
- 6 Significado de los Coeficientes del Modelo de Regresión logística Simple

## Variable Respuesta

### Naturaleza dicotómica

La variable respuesta de este modelo es dicotómica (dos categorías), indica presencia de una condición ( $Y = 1$ ) o ausencia ( $Y = 0$ ). Ejemplos: respuesta a un tratamiento, presencia de rechazo, presentación de efectos adversos, etc.

### Buscamos un modelo de regresión adecuado

que estime la proporción de individuos en la población con la característica de interés, o bien la probabilidad de que un individuo tenga dicha característica, para nivel de la variable explicativa. Indicamos con  $p$  la probabilidad de que un individuo tenga diabetes. La variable  $Y$ , que indica la presencia o ausencia del evento, vale 1 si el individuo tiene diabetes y 0 si no. Su distribución es  $Bi(1, p)$  y su valor esperado ó valor medio es  $E(Y) = p$ .

## Ejemplo Diabetes

### Objetivo

Estudiar la relación entre la glucosa-SSPG y la presencia o ausencia de diabetes en la población en estudio. ¿Qué significaría en este caso ajustar un modelo de regresión lineal simple para la variable respuesta ( $Y_i$  = presencia o ausencia de diabetes) con edad ( $X_i$ ) como variable explicativa?

Una vez conocido el valor de  $X$ , la media de  $Y$  es la probabilidad  $p$  de que  $Y = 1$ . Esto es, para cada valor de  $x$

$$0 \leq E(Y/X = x) = P(Y = 1/X = x) \leq 1$$

Facultad de Ingeniería

## Por qué el modelo lineal no resulta adecuado?

### Propuesta del Modelo Lineal

Por lo tanto el modelo lineal propone que la probabilidad de que un individuo elegido al azar (en la población en cuestión) entre los que tienen  $x$  glucosa tenga diabetes es una función lineal de la glucosa:

$$p(x) = \alpha + \beta x$$

### Inconveniente

Cuando ajustemos la regresión, podría ocurrir que el valor estimado de  $Y$  para valores de  $X$  dentro del rango de valores observados cayera fuera de los límites establecidos por la restricción y así el modelo no tendría sentido.

## Transformación de la Variable Respuesta

### Objetivo

Estimar la probabilidad de que un individuo elegido al azar (en la población en cuestión) entre los que tienen  $x$  glucosa tenga diabetes es una función lineal de la edad:

$$\ln \left( \frac{p(x)}{1 - p(x)} \right) = \alpha + \beta x$$

Así se logra que los valores estimados de  $p(x)$  se encuentren en el intervalo  $[0, 1]$ . Despejando de esta expresión:

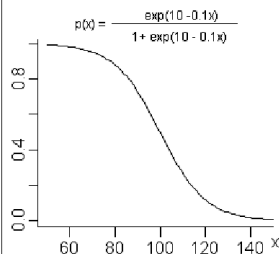
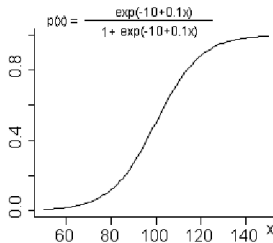
$$p(x) = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}}$$

## Cómo es la expresión que estima la probabilidad?

La probabilidad de que un individuo elegido al azar

(en la población en cuestión), entre los que tienen  $x$  glucosa, tenga diabetes se relaciona en forma curvilínea con la variable explicativa de acuerdo con la expresión (1). La curva dada en (1) toma valores dentro del intervalo  $[0, 1]$ .

### Funciones de Respuesta Logística



## Por qué esta función de Enlace(link)?

La transformación usada se denomina logit :

$$\ln \left( \frac{p}{1-p} \right) = \ln(odds)$$

donde  $p$  es la proporción poblacional de individuos con la característica y el cociente  $\left( \frac{p}{1-p} \right)$  se llama odds (oportunidad).

Dos razones fundamentales para elegirla:

- ① Desde el punto de vista matemático, es una función muy flexible y muy fácil de usar.
- ② Desde el punto de vista biológico, los coeficientes admiten una interpretación simple

# Organización

- 1 Modelos GAM
- 2 Modelos GAMLSS
- 3 Modelos Lineales Generalizados
- 4 Regresión Logística
- 5 Modelo de Regresión Logística
- 6 Significado de los Coeficientes del Modelo de Regresión logística Simple
  - Evaluación del Modelo



## El coeficiente importante es $\beta$

El coeficiente  $\beta$  es el cambio en el logit ( $\ln\left(\frac{p}{1-p}\right)$ )

cuando la variable  $X$  aumenta en 1 unidad o equivalentemente  $\beta$  es el cambio en el  $\ln(\text{OR})$  entre los grupos definidos por  $X = x + 1$  y por  $X = x$ .

El OR asociado a un cambio en una unidad de la variable  $X$  como

$$\text{OR} = e^{\beta}$$

En ocasiones un cambio unitario no es de interés biológico

Un cambio en 10 años puede ser más útil para interpretar que un cambio de un año. El  $\ln(\text{OR})$  entre el grupo definido por  $X = x + c$  y el grupo definido por  $X = x$  es:

$$\text{OR} = e^{\beta \cdot c}$$

## Ejemplo Diabetes

### Estimación de los Coeficientes del Modelo

La variable respuesta es la presencia ( $Y=1$ ) ó ausencia ( $Y=0$ ) de diabetes y la única variable explicativa es la SSPG.

### El Código

```
library(glmnet)
diabetes <- read_excel("../diabetes.xls")
attach(diabetes)
mod_1=glm(DIABET ~ SSPG,family="binomial")
summary(mod_1)
```

## La Salida

|             | Estimate | Std.Err  | z value | p value  |     |
|-------------|----------|----------|---------|----------|-----|
| (Intercept) | -4.54816 | 0.711406 | -6.393  | 1.62E-10 | *** |
| SSPG        | 0.02528  | 0.003939 | 6.418   | 1.38E-10 | *** |

Null deviance: 200.67 on 144 degrees of freedom

Residual deviance: 106.69 on 143 degrees of freedom

AIC: 110.69

Number of Fisher Scoring iterations: 6

### Deviance

El análisis de la deviance es una generalización del de la varianza para los GLM obtenido para una secuencia de modelos anidados (cada uno incluyendo más términos que los anteriores). Dada una secuencia de modelos anidados usamos la deviance como una medida de discrepancia y podemos formar una tabla de

## Coeficientes y Significación

- 😊  $\hat{\beta}$ : los coeficientes de cada variable en el modelo logístico, en este caso solamente el coeficiente de SSPG y la ordenada al origen (Constante).
- 😊 **Std.Err.:** son los errores típicos, errores estándar asociados a los coeficientes estimados.
- 😊 Estos errores estimados permiten realizar tests basados en la distribución Normal para decidir si el coeficiente es estadísticamente significativo (mayor a cero, menor a cero o distinto de cero).

## Significación de los Coeficientes y del Modelo

### Estadísticos y $p$ - valores

**estadístico del test de Wald:**  $(\hat{\beta}/Std.Err)^2$  el test para probar si el coeficiente es significativamente distinto de cero. Está basado en una distribución Chi-cuadrado con tantos grados de libertad como aparecen en la columna g.l. (cada variable continua aporta 1 grado de libertad). Sus  $p$ -valores tienen validez aproximada.

**p valor** corresponde al test Wald para cada coeficiente y son validez aproximada.

El modelo estimado de acuerdo con la tabla anterior es :

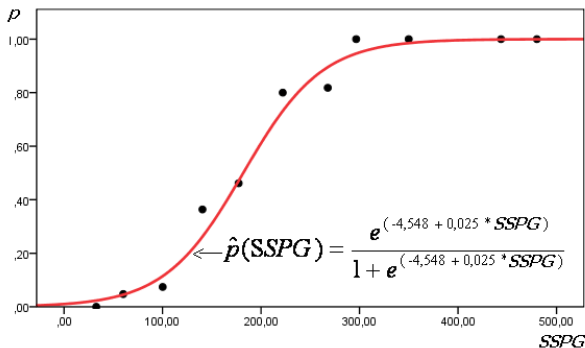
$$\text{logit estimado} = -4,548 + 0,025 * \text{SSPG}$$

$$\log(\text{odds est.}) = \log( p(\text{SSPG})/(1- p(\text{SSPG})) ) = -4,548 + 0,025 * \text{SSPG}$$

# Estimación de Probabilidades a partir del Modelo Ajustado

La expresión para estimar la probabilidad dado un valor de SSPG es:

$$\hat{p} = \frac{e^{-4.548+0.025*SSPG}}{1 + e^{-4.548+0.025*SSPG}}$$



## Ejemplo Diabetes

### Observaciones e Interpretación

- El coeficiente asociado a SSPG es positivo, lo que nos indica que la probabilidad de tener diabetes aumenta con el nivel de glucosa.
- El coeficiente asociado a SSPG es  $b = 0,025$ . Por lo tanto el odds ratio estimado para un aumento de una unidad de SSPG es (estadísticamente significativo) mayor a 1.

Podemos hallar un intervalo de confianza para los odds ratio estimados:

```
confint(object = mod_1, level = 0.95 )
```

|             | 2.50%    | 97.50%   |
|-------------|----------|----------|
| (Intercept) | -6.08666 | -3.27526 |
| SSPG        | 0.01828  | 0.03384  |

## Ejemplo Diabetes

### Intervalo de Confianza para el $OR_{10}$

El odds ratio estimado para un aumento de 10 unidades de SSPG es

$OR_{10} = e^{0.025 \cdot 10} = 1.29$  Para hallar el IC unitario:

```
exp(confint(object = mod_1, level = 0.95 ) [2,])
```

1.0184      1.0344

Para 10 unidades:

```
exp(10*(confint(object = mod_1, level = 0.95 ) [2,]))
```

### Cuidado!

La curva logística estimada, permite hallar una probabilidad estimada para cada valor posible de la variable SSPG, sin embargo, sólo tiene sentido la estimación dentro del rango de los valores observados de la variable explicativa. No se debe extrapolar.



## Ejemplo Diabetes

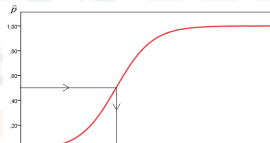
### Valor Límite

En general, una vez que se ha hallado la curva logística estimada es posible estimar el valor de  $X$  que corresponde a una cierta prevalencia. La prevalencia del 50% es la más fácil de calcular. Si el modelo estimado es cuando

$$p = 0.5 \Rightarrow \text{logit}(p) = 0$$

$$\text{y resulta } 0 = a + b \cdot x \longleftrightarrow x = -a/b = 179.91$$

Esto significa que se estima que en la población de la cual fue tomada la muestra, los pacientes con SSPG = 179,91 tienen una probabilidad estimada del 0.5 de tener diabetes.



## Test de Hosmer- Lemeshow

### Idea General

- ① Se ordenan los casos de acuerdo con la probabilidad predicha.
- ② Se dividen en 10 grupos con la misma cantidad de casos. El primer grupo (primer decil) consiste de los  $n/10$  casos con los valores mas bajos de probabilidad predicha (primer decil), etc. En la práctica, no es siempre posible formar grupos de exactamente el mismo tamaño.
- ③ Cada una de las categorías anteriores es subdividida en dos grupos en base al valor de la variable respuesta. Se calculan las frec. observadas y las frec. esperadas para cada una de las 20 celdas.
- ④ El grado de ajuste se obtiene calculando el estadístico de Chi-cuadrado de Pearson para una tabla de contingencia de  $2 \times 10$ .
- ⑤ El estadístico del test tiene una distribución aproximada chi-cuadrado con  $g-2$  g.l., donde  $g$  es la cantidad de grupos. Para la mayoría de los conjuntos de datos  $g = 10$  y los grados de libertad son 8.

## Test de Hosmer - Lemeshow: Objetivo

### Bondad de Ajuste del Modelo

Este test examina si las proporciones observadas de eventos son similares a las probabilidades predichas de ocurrencia en subgrupos del conjunto de datos, y realiza esta comparación mediante la aplicación de una prueba de chi cuadrado de Pearson.

¡Ojo! se trata de un test de bondad de ajuste, es decir que en este caso no quisiéramos rechazar la hipótesis nula  $H_0$ , dado que la hipótesis nula sostiene que el modelo se ajusta a los datos.

Esta prueba ha sido criticada <https://statisticalhorizons.com/hosmer-lemeshow>

## Test de Hosmer- Lemeshow: Idea

### Código y Biblioteca para el Test

```
library(ResourceSelection)  
hoslem.test(diabetes$DIABET, mod_1$fitted.value)
```

### Salida de R

Hosmer and Lemeshow goodness of fit (GOF) test

data: diabetes\$DIABET, mod\_1\$fitted.val X-squared = 3.5997, df  
= 8, p-value = 0.8913

**No se rechaza la hipótesis nula que sostiene que la distribución de observados y predichos es similar.**

## La Trastienda de la prueba

`cbind(HL$observed,HL$expected)`

| Intervalos   | Observados | Esperados |
|--------------|------------|-----------|
| 0.0216-0.046 | 15         | 15.427    |
| 0.046-0.0914 | 13         | 12.113    |
| 0.0914-0.14  | 15         | 14.1049   |
| 0.14-0.238   | 10         | 10.624    |
| 0.238-0.371  | 9          | 10.3502   |
| 0.371-0.652  | 7          | 7.0701    |
| 0.652-0.826  | 3          | 3.5498    |
| 0.826-0.91   | 3          | 1.8944    |
| 0.91-0.976   | 1          | 0.7440    |
| 0.976-0.999  | 0          | 0.1209    |
| 0.0216-0.046 | 1          | 0.572     |
| 0.046-0.0914 | 0          | 0.8863    |
| 0.0914-0.14  | 1          | 1.8950    |
| 0.14-0.238   | 3          | 2.3758    |
| 0.238-0.371  | 6          | 4.6497    |
| 0.371-0.652  | 7          | 6.929     |
| 0.652-0.826  | 11         | 10.4501   |
| 0.826-0.91   | 12         | 13.1055   |
| 0.91-0.976   | 13         | 13.255    |
| 0.976-0.999  | 15         | 14.879    |

## Test de Hosmer-Lemeshow

### Intepretación

Nos interesan modelos en los cuales no se rechace la hipótesis de igualdad entre los valores observados y los valores predichos por el modelo, lo que implicaría que el modelo ajusta a los datos bastante bien.

### **CUIDADO!**

Para que el estadístico de H-L se aproxime razonablemente a la distribución chi-cuadrado algunos sugieren que haya suficientes casos como para que el 95% de las celdas tengan una frec. esp. mayor a 5 y ninguna menor a 1.

Facultad de Ingeniería

# Organización

- 1 Modelos GAM
- 2 Modelos GAMLSS
- 3 Modelos Lineales Generalizados
- 4 Regresión Logística
- 5 Modelo de Regresión Logística
- 6 Significado de los Coeficientes del Modelo de Regresión logística Simple
- 7 Tablas de Clasificación y Curvas ROC

# Sensibilidad-Especificidad

## El Modelo pensado como Test Diagnóstico

Una forma de evaluar la calidad de la regresión es mediante una tabla de clasificación en la que los sujetos se clasifican de acuerdo a dos factores:

- 1 la presencia o ausencia observada de una condición, por ejemplo tener diabetes
- 2 la presencia o ausencia predicha por el modelo logístico utilizando como punto de corte 0.5, es decir que consideramos que un sujeto tiene la condición si la probabilidad predicha por el modelo  $\geq 0.5$  y no la tiene si es  $< 0.5$ .
- 3 El modelo se puede considerar como un test (diagnóstico o de screening) que puede ser usado para predecir si un sujeto tiene o no tiene una condición. En nuestro caso la condición es tener diabetes y el test es positivo si la probabilidad predicha por el modelo es  $\geq 0.5$ .



## Ejemplo Diabetes

### Tabla de Clasificación

La probabilidad estimada de tener diabetes ahora depende del valor de la variable explicativa. Cuando esa probabilidad estimada es menor a 0,5 el caso es clasificado como  $DIABET = 0$  y si es mayor a 0,5 como  $DIABET=1$ , obteniéndose así la tabla siguiente.

```
table(DIABET,1*(predict(mod_1)> 0.5))
```

**Tabla:** Tabla de Clasificación Cruzada

|        |    | Modelo |    |
|--------|----|--------|----|
|        |    | No     | Sí |
| Reales | No | 67     | 9  |
|        | Sí | 14     | 55 |

## Sensibilidad Especificidad

Las calculamos a partir de la Tabla de Clasificación

Estos conceptos conducen a la comprensión de las curvas ROC

$$\text{SENSIBILIDAD} = 55 / (14 + 55) = 55 / 69 = 0.797$$

$$\text{ESPECIFICIDAD} = 67 / (67 + 9) = 67 / 76 = 0.882$$

$$1\text{-ESPECIFICIDAD} = 11 / 76 = 1 - 0.882 = 0.118$$

$$\text{Valor Predictivo Positivo (VPP)} = 55 / (55 + 9) = 0.8593$$

$$\text{Valor Predictivo Negativo (VPN)} = 67 / (67 + 14) = 0.827$$

Facultad de Ingeniería

# Tablas de Clasificación con distintos Puntos de Corte

## Buena Clasificación

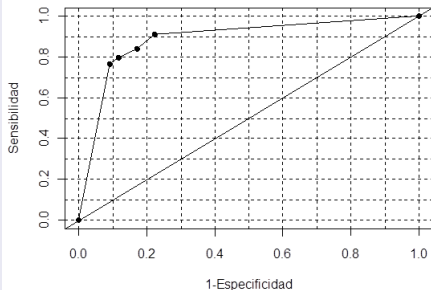
La sensibilidad, la especificidad y la proporción total de individuos correctamente clasificados dependen del punto de corte. ¿Qué ocurriría si eligiéramos otro punto de corte? Se podrían construir tablas de clasificación para distintos puntos de corte y evaluar para cada una de ellas la bondad de la clasificación. Uniendo los puntos cuyas coordenadas son (1-Especificidad, Sensibilidad), considerando todos los posibles puntos de corte  $\alpha$  entre 0 y 1, se construyen las Curvas, o Característica Operativa del Receptor.

Ver el RMarkdown para analizar cómo se modifican estos valores al mover el valor de corte.

## Construcción de la Curva ROC

### Utilizamos cuatro puntos de corte

Representaremos gráficamente la proporción de verdaderos positivos (sensibilidad) y la proporción de falsos positivos (1-especificidad) obtenidas con los cuatro criterios de clasificación utilizados.



## ROC (Receiver Operating Characteristic)

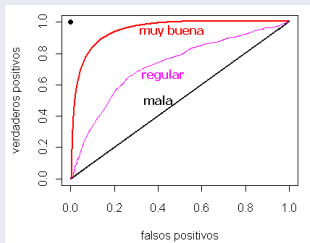
### Construcción de la Curva ROC

La curva que se obtiene uniendo los puntos (falsos positivos, verdaderos positivos) graficados en un diagrama de dispersión para los distintos puntos de corte posibles en una prueba diagnóstica se denomina curva ROC.

- i) Obtuvimos una curva ROC con cuatro puntos de corte.
- ii) Una prueba ideal tendría un único punto: (falsos positivos = 0, verdaderos positivos = 1).
- iii) Si la prueba no sirviera para nada (falsos positivos = verdaderos positivos) la curva sería la recta que une el (0,0) con el (1,1).
- iv) En general las pruebas tienen curvas intermedias.
- v) El área bajo la curva ROC oscila entre 0.5 y 1 siendo 1 si el test es perfecto y 0.5 si es equivalente a arrojar una moneda.

## Bondad a partir de la Curva ROC

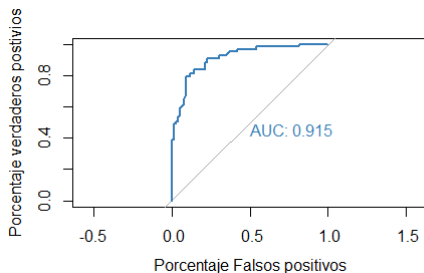
Cuándo un modelo es bueno?



Que significa el área bajo la curva?

Hanley y McNeil (1982) probaron que es la probabilidad de que dos individuos elegidos al azar, uno con la condición (enfermo) y el otro sin la condición (sano) sean clasificados correctamente mediante esa prueba. Si esa probabilidad es  $1/2$  significa que es tan probable clasificarlos

# Estimación del Area bajo la Curva ROC

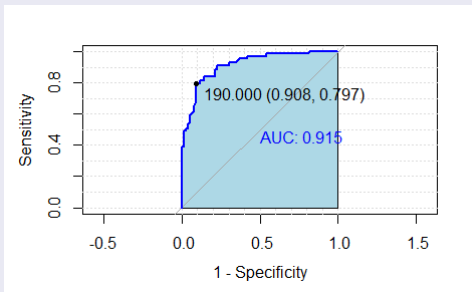


```
library(pROC)
```

```
roc(DIABET=="1", mod_1$fitted.values, plot = TRUE, legacy.axes =  
TRUE, add=TRUE, print.auc = TRUE,  
percent = TRUE, xlab = "Porcentaje Falsos positivos",  
ylab = "Porcentaje verdaderos positivos", col = "#377eb8", lwd =  
2, xlim=c(100,0))
```

## Estimación del Area bajo la Curva ROC

Se puede pedir también un intervalo de confianza para el área bajo la curva ROC



El punto señalado en la curva es el punto de corte cuya consideración conduce al máximo valor de suma de sensibilidad y especificidad. Suele ser una sugerencia de punto de corte.



# Estimación del Area bajo la Curva ROC

## Comparación de Áreas bajo la Curva ROC

Las curvas ROC pueden ser usadas para comparar dos modelos de regresión logística. Si un modelo logístico produce una curva ROC siempre encima y a izquierda de la otra, este tendrá mayor área bajo la curva. Pero aún cuando las curvas se cortan es posible calcular las áreas bajo las curvas y testearlas mediante la prueba de Hanley y McNeil

Facultad de Ingeniería

## Comparación de Modelos

### Test de Hanley y McNeil (1983)

Si dos curvas han sido estimadas con diferentes sujetos, de manera que las estimaciones sean independientes, es sencillo decidir si la diferencia entre sus áreas es estadísticamente significativa conociendo el error estándar de cada una de las áreas utilizando el siguiente estadístico que tiene una distribución aproximadamente Normal.

$$Z_{obs} = \frac{A_1 - A_2}{\sqrt{EE_1^2 + EE_2^2}} \approx N(0, 1)$$

## Ejemplo Comparación Curvas ROC

### Ejemplo

Consideremos curva ROC de nuestro ajuste logístico tiene un área  $A_1 = 0.915$  y un error estándar  $EE_1 = 0.023$ . Queremos compararla con un modelo ajustado por otro grupo de investigación cuya curva ROC tiene un área  $A_2 = 0,875$  y un error estándar  $EE_2 = 0.04$ :

$$Z_{obs} = \frac{0.915 - 0.875}{\sqrt{0.023^2 + 0.04^2}} = 0.866$$

El p valor ( como el test es bilateral) se calcula:

$$P(|z| > |Z_{obs}|) = P(|z| > 0.866) = 0.386$$

**Conclusión:** No tenemos suficiente evidencia para decidir que nuestro ajuste logístico provee una prueba diagnóstica significativamente mejor.

## Consideraciones sobre las Curvas ROC

Las Curvas ROC permiten:

- ☹ Determinar una calidad global de la prueba mediante el área.
- ☹ Comparar dos modelos.
- ☹ Comparar dos puntos de corte, dentro de la misma prueba o entre dos pruebas.
- ☹ Elegir el punto de corte que mejor se adecue a los requerimientos de sensibilidad y especificidad.



## Modelo Logístico con Variables Categóricas

Consideramos ahora dos Variables Categóricas como Predictoras para estimar la probabilidad de bajo peso al nacer (Base BajoPeso.xlsx)

| Variable |   |
|----------|---|
| LOW      | Bajo Peso al Nacimiento (0: $\geq 2500g$ 1: $< 2500g$ )       |
| AGE      | Edad de la Madre en Años                                      |
| LWT      | Peso en Libras en el Último Período Menstrual                 |
| RACE     | Raza (1 = Blanca, 2 = Negra, 3 = Otra)                        |
| SMOKE    | Fumó durante el embarazo (0 = No, 1 = Sí)                     |
| PTL      | Historia de trabajo de parto prematuro (0 = Ninguno, 1 = Uno) |
| HT       | Historia de Hipertensión (0 = No, 1 = Sí)                     |
| UI       | Presencia de Irritabilidad Uterina (0 = No, 1 = Sí)           |
| FTV      | Cant. de consultas médicas durante el Primer Trimestre        |

## Estimamos un modelo Logístico con la Variable SMOKE

El modelo General es

$$\text{logit}(p) = \beta_0 + \beta_1 * \text{SMOKE} \quad (\text{Modelo 1})$$

El Modelo estimado con nuestros datos es:

$$\text{logit}(\hat{p}) = -1.087 + 0.704 * \text{SMOKE}$$

Cómo se Interpreta?

$$\text{logit}(\hat{p}, \text{SMOKE} = 1) - \text{logit}(\hat{p}, \text{SMOKE} = 0) = 0.704 \quad (\text{Modelo 1})$$

$$\text{logit}(\text{OR}, \text{SMOKE} = 1) / (\text{OR}, \text{SMOKE} = 0) = 0.704$$

$$\frac{\text{OR}(\text{SMOKE} = 1)}{\text{OR}(\text{SMOKE} = 0)} = e^{0.704} = 2.022$$

## Interpretación de los Coeficientes del Modelo

La oportunidad de bajo peso al nacer del hijo de una madre fumadora es 2.022 veces esa oportunidad para una madre no fumadora. Se estima un aumento del 102.2% del odds de bajo peso al nacer entre las madres que fuman en comparación con las que no fuman.

### Estudiamos a continuación

Si el peso de la madre en el último período menstrual (LWT) puede explicar, aunque sea en parte, el bajo peso al nacer. Ajustamos el modelo:

$$\text{logit}(p) = \beta_0 + \beta_1 * LWT \quad (\text{Modelo 2})$$

El Modelo estimado con nuestros datos es:

$$\text{logit}(\hat{p}) = 0.998 - 0.014 * LWT$$

## Interpretación de los Coeficientes

¿Cómo podemos interpretar ahora el coeficiente de LWT ?

El coeficiente estimado es  $\hat{\beta}_1 = -0.014$  Se estima que un aumento de una libra en el peso de la madre resulta en una reducción del logit de tener bajo peso al nacer en 0.014:

$$\log(ODDS, LWT = x + 1) - \log(ODDS, LWT = x) = -0.014$$

$$\frac{ODDS(LWT = x + 1)}{ODDS(LWT = x)} = e^{-0.014} = 0.986$$

Es decir que la oportunidad estimada para  $LWT = x + 1$  es 0.986 veces la estimada para  $LWT = x$



## Modelo Logístico: una Categórica y una Continua

### Sin Interacción

$$\text{logit}(p) = \beta_0 + \beta_1 * \text{SMOKE} + \beta_2 * \text{LWT} \quad (\text{Modelo 3})$$

Se trata de un modelo de 2 rectas paralelas en la escala logit:

(a) si  $\text{SMOKE} = 0$  resulta

$$\text{logit}(p) = \beta_0 + \beta_2 * \text{LWT}$$

modelo 3.0

(b) si  $\text{SMOKE} = 1$  resulta

$$\text{logit}(p) = \beta_0 + \beta_1 * \text{SMOKE} + \beta_2 * \text{LWT}$$

modelo 3.1

## Interpretación de las Diferencias

### Observaciones

Las rectas propuestas en los modelos 3.0 y 3.1 difieren en la ordenada al origen y tienen la misma pendiente.

El coeficiente de SMOKE ( $\beta_1$ ) indica el cambio en el  $\text{logit}(p)$  entre las categorías  $SMOKE = 1$  y  $SMOKE = 0$ , para cada valor fijo de la variable LWT, es decir para cada peso fijo, es decir controlando por la variable peso.

$e^{\beta_1} = \text{OR}$  (bajo peso al nacer, entre las madres que fuman y las que no fuman) controlando por la variable peso.

El coeficiente de LWT es el cambio en el  $\text{logit}(p)$  entre los grupos que tienen peso  $LWT = x + 1$  y  $LWT = x$ , para cada valor fijo de la variable SMOKE, es decir dentro de cada grupo (el grupo de las embarazadas que fuman y las embarazadas que no fuman), es decir controlando por la condición de fumar.

## Interpretación de los Coeficientes del Modelo

Esto significa que:

- 1 se estima que el odds (la oportunidad) de bajo peso al nacer entre madres que fuman es 1,967 veces el odds (la oportunidad) de bajo peso al nacer entre madres que no fuman , para cada valor fijo de la variable LWT.
- 2 Se estima un aumento del 96,7% del odds de bajo peso al nacer entre las madres que fuman en comparación con las que no fuman para cada valor fijo de la variable LWT.
- 3 Como el cambio de odds entre los grupos de madres que fuman y no fuman no depende del valor de la otra variable (LWT), decimos también “independientemente” del valor de la variable LWT

## Interpretación de los Coeficientes del Modelo

¿Qué significa el coeficiente de LWT estimado  $\beta_2 = -0.013$  ?

1 Se estima que un aumento de una libra en el peso de la madre resulta en una reducción del logit de tener bajo peso al nacer en 0.013, para cada valor fijo de la variable FUMA:

$$\text{logit}(\hat{p}, LWT = x + 1) - \text{logit}(\hat{p}, LWT = x) = -0.013$$

$$\log(\hat{p}/(1 - \hat{p}), LWT = x + 1) - \log(\hat{p}/(1 - \hat{p}), LWT = x) = -0.013$$

$$\log(OR_{\text{entre los grupos } LWT = 1 \text{ y } LWT = 0}) = -0.013$$

$$OR(\text{entre los grupos } LWT = x + 1 \text{ y } LWT = x) = 0.987$$

para cada valor fijo de la variable FUMA

## Interpretación de los Coeficientes del Modelo

Se interpreta que:

- 1 se estima que el odds (la oportunidad) de bajo peso al nacer entre madres con  $LWT = x + 1$  es 0.987 veces el odds (la oportunidad) de bajo peso al nacer entre madres con peso = LWT, para cada valor fijo de la variable FUMA.
- 2 Se estima una reducción del 1.3% del odds de bajo peso al nacer entre las madres con peso = LWT+1 en comparación con las de peso  $LWT = x$ , para cada valor fijo de la variable FUMA.
- 3  $100 * (0,987 - 1) = -1.3$
- 4 Como el cambio de odds de bajo peso al nacer entre los grupos de madres referidos no depende del valor de la otra variable (FUMA), decimos también: OR (entre los grupos  $LLWT=x+1$  y  $LWT=x$ ) = 1.967 “independientemente” de la condición de fumar.

## Evaluando un modelo con una variable

Hay dos opciones para evaluar la contribución de una variable a un modelo. Si es una sola variable, se puede comparar con el modelo nulo (en relación con la deviance).

```
anova(MOD_2, test = 'Chisq')
```

|      | Df | Deviance | Resid. Df | Resid. Dev |
|------|----|----------|-----------|------------|
| NULL |    |          | 188       | 234.67     |
| LWT  | 1  | 5.98     | 187       | 228.69     |

Equivale a hacer la prueba de diferencia de deviances:

```
1-pchisq(234.67-228.69,1)
```

## Comparación de Coeficientes de SMOKE

### Modelo 1 vs Modelo 3

Coeficiente de SMOKE (MODELO 1) = 0.704

Coeficiente de SMOKE (MODELO 3) = 0.677

¿ LWT es factor de confusión entre fumar y bajo peso?

DIFERENCIA:  $0,704 - 0,677 = 0.027$

La reducción del coeficiente de SMOKE es  $0.024 / 0.704 * 100 = 3,409$  (menor al 10 %) y sigue siendo estadísticamente significativo. No existen razones para considerar que el peso de la madre en el último período menstrual es un factor de confusión en la relación entre fumar y bajo peso.





## Eliminando las Variables de a una

Ajustamos un modelo con las dos variables y evaluamos la significación

```
MOD_3=glm(LOW ~ LWT+SMOKE,family='binomial',data=bajo_peso)
anova(MOD_3)
```

Comparamos este modelo con los univariados

|       | Df | Deviance | Resid. Df | Resid. Dev | Pr(>Chi) |
|-------|----|----------|-----------|------------|----------|
| NULL  |    |          | 188       | 234.67     |          |
| LWT   | 1  | 5.98     | 187       | 228.69     | 0.0145   |
| SMOKE | 1  | 4.35     | 186       | 224.34     | 0.0370   |

De acuerdo con la tabla anterior eliminar cada variable por separado en el modelo 3 produce un cambio estadísticamente significativo distinto de cero. Por este motivo no deberíamos eliminar ninguna de las dos.

# Comparación de Modelos

## ¿Cómo se comparan modelos anidados?

|   |                        |
|---|------------------------|
| -2log (MOD 2 (sólo con LWT))= 229,05    | 229.05 - 224.34 = 5.46 |
| -2log (MOD 3 (con LWT y FUMA))= 224,341 |                        |
| -2log (MOD 2 (sólo con FUMA))= 228,691  | 228.69-224.34 = 4.35   |
| -2log (MOD 3 (con LWT y FUMA))= 224,341 |                        |



Facultad de Ingeniería

## Cómo hacerlo en R

Comparamos el primer modelo univariado con el modelo bivariado

```
anova(MOD_1,MOD_3 ,test ='Chisq')
```

|   | Resid. Df | Resid. Dev | Df | Deviance | Pr(>Chi) |
|---|-----------|------------|----|----------|----------|
| 1 | 187       | 229.80     |    |          |          |
| 2 | 186       | 224.34     | 1  | 5.46     | 0.0194   |

Se comparan las deviances de ambos modelos y se contrastan contra una distribución Chi cuadrado con tantos grados de libertad como diferencia de variables predictoras tengan ambos modelos.

## Comparemos con el test de Wald

### Variables en la ecuación

|         |           | B     | E.T. | Wald  | gl | Sig. | Exp(B) |
|---------|-----------|-------|------|-------|----|------|--------|
| Paso 1a | LWT       | -,013 | ,006 | 4,788 | 1  | ,029 | ,987   |
|         | SMOKE     | ,677  | ,325 | 4,343 | 1  | ,037 | 1,967  |
|         | Constante | ,622  | ,796 | ,611  | 1  | ,435 | 1,863  |

a. Variable(s) introducida(s) en el paso 1: LWT, SMOKE.

Deberían incluirse LWT? y SMOKE ? juntas?

Facultad de Ingeniería

## Modelos con Interacción y sin Interacción

Agregaremos ahora un término de interacción al modelo 3

$$\text{logit}(p) = \beta_0 + \beta_1 * \text{SMOKE} + \beta_2 * \text{LWT} + \beta_{12} \text{SMOKE} * \text{LWT}$$

(Modelo 4)

Son 2 rectas en la escala logit, que pueden no ser paralelas:

si  $\text{SMOKE} = 0$  resulta  $\text{logit}(p) = \beta_0 + \beta_2 \text{LWT}$  (mod 4.0)

si  $\text{SMOKE} = 1$  resulta  $\text{logit}(p) = (\beta_0 + \beta_1) + (\beta_2 + \beta_{12}) * \text{LWT}$   
( mod 4.1)

Las dos rectas difieren en ordenada al origen y pendiente.

## ¿Qué significan los coeficientes en un modelo con interacción?

### SMOKE

El coeficiente de SMOKE ( $\beta_1$ ) indica el cambio en el logit (p) entre las categorías SMOKE = 1 y SMOKE = 0, cuando la variable LWT es cero. No tiene sentido biológico.

$e^{\beta_1} = OR$  (bajo peso al nacer, entre las madres que fuman y las que no fuman) cuando LWT = 0. No tiene sentido biológico.

El OR (bajo peso al nacer, entre las madres que fuman y las que no fuman) depende del valor de LWT:

$$e^{\beta_1 + \beta_{12}LWT} = e^{\beta_1} e^{\beta_{12}LWT}$$

## ¿Qué significan los coeficientes en un modelo con interacción?

LWT

$$\text{logit}(p, \text{SMOKE}, \text{LWT} = x + 1) = \beta_0 + \beta_1 \text{SMOKE} + \beta_2(x + 1) + \beta_{12}(x + 1) -$$

$$\beta_0 - \beta_1 \text{SMOKE} - \beta_2(x) + \beta_{12}(x) = \beta_2 + \beta_{12} \text{SMOKE}$$

El coeficiente de LWT  $\beta_2$  es el cambio del  $\text{logit}(p)$  entre los grupos que tienen  $\text{LWT} = x + 1$  y  $\text{LWT} = x$ , cuando la variable SMOKE es 0, vale decir que la madre no es fumadora.

El OR (bajo peso al nacer, entre estos grupos ) depende de la condición de fumar:

$$e^{\beta_2 + \beta_{12} \text{SMOKE}} = e^{\beta_2} e^{\beta_{12} \text{SMOKE}}$$

El OR de bajo peso al nacer entre los grupos determinados por el aumento de una unidad en una variable depende del valor de la otra variable.

## Modelo Ajustado con Interacción

### Salida

Tabla: Primer Paso

|                     | B      | E.T.  | Wald  | gl | Sig. | Exp(B) |
|---------------------|--------|-------|-------|----|------|--------|
| <b>LWT</b>          | -,024  | ,010  | 5,284 | 1  | ,022 | ,976   |
| <b>SMOKE</b>        | -1,511 | 1,617 | ,873  | 1  | ,350 | ,221   |
| <b>LWT by SMOKE</b> | ,018   | ,013  | 1,885 | 1  | ,170 | 1,018  |
| <b>Constante</b>    | 1,932  | 1,292 | 2,236 | 1  | ,135 | 6,906  |

Es el término de interacción estadísticamente significativo?



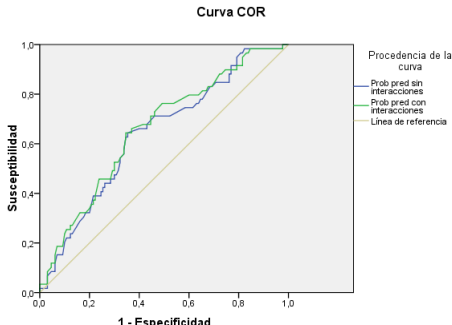
## Con las Verosimilitudes

### Comparación de los Modelos con y sin interacción

$-2\log (\text{MOD 3 (con LWT y FUMA)}) = 224,341$

$-2\log (\text{MOD 4 (con interacción)}) = 222,372$

$$224.34 - 222.379 = 1.96$$





## Comparación de Modelos con y sin Interacción

### Intervalos de Confianza para el área bajo la curva

#### Área bajo la curva

| Var. resultado de contraste | IC asintótico al 95% |              |
|-----------------------------|----------------------|--------------|
|                             | Límite inferior      | Lím superior |
| Prob pred sin interac       | ,557                 | ,724         |
| Prob pred con interac       | ,576                 | ,741         |

Para ambos modelos el límite inferior de los intervalos de confianza del área bajo la curva ROC son cercanos a 0.5. Ninguno de los dos modelos parece demasiado adecuado para predecir el bajo peso al nacer.

## Ejemplo: Default

### El Problema

En el siguiente ejemplo se modela la probabilidad de fraude por impago (default) en función del balance de la cuenta bancaria (balance). Los datos están disponibles en la biblioteca ISLR de R.



## Cargamos las librerías y los datos

```
library(tidyverse)
```

```
library(ISLR)
```

```
datos <- Default
```

```
# Se recodifican los niveles No, Yes a 1 y 0
```

```
datos <- datos %>% select(default, balance) %>%
```

```
mutate(default = recode(default, "No" = 0, "Yes" = 1))
```

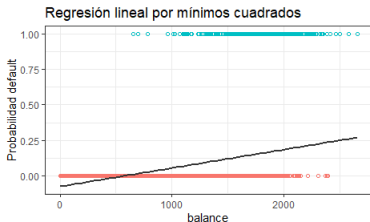
```
head(datos)
```

|   | default | balance |
|---|---------|---------|
| 1 | 0.00    | 729.53  |
| 2 | 0.00    | 817.18  |
| 3 | 0.00    | 1073.55 |
| 4 | 0.00    | 529.25  |
| 5 | 0.00    | 785.66  |
| 6 | 0.00    | 919.59  |

## Ajustamos un Modelo Lineal y Visualizamos el Ajuste

```
# Ajuste de un modelo lineal por mínimos cuadrados. modelo_lineal <-  
lm(default ~ balance, data = datos)
```

```
# Representación gráfica del modelo. ggplot(data = datos, aes(x =  
balance, y = default)) +  
geom_point(aes(color = as.factor(default)), shape = 1) +  
geom_smooth(method = "lm", color = "gray", se = FALSE) +  
theme_bw() + labs(title = "Reg. lineal por mín. cuadrados",  
y = "Probab. default") + theme(legend.position = "none")
```



## Cuál es el inconveniente con este modelo?

Al tratarse de una recta, si por ejemplo, se predice la probabilidad de default para alguien que tiene un balance de 10000, el valor obtenido es mayor que 1.

```
predict(object = modelo_lineal, newdata = data.frame(balance = 10000))  
1.22353
```

**No tiene sentido!**

Facultad de Ingeniería

## Ajuste Logístico

# Ajuste de un modelo logístico.

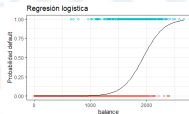
```
modelo_logistico <- glm(default ~ balance, data = datos, family = "binomial")
```

# Representación gráfica del modelo. `ggplot(data = datos, aes(x = balance, y = default)) +`

```
geom_point(aes(color = as.factor(default)), shape = 1) +
```

```
stat_function(fun = function(x)predict(modelo_logistico,  
newdata = data.frame(balance = x), type = "response")) +
```

```
theme_bw() + labs(title = "Regresión logística",  
y = "Probab. default") + theme(legend.position = "none")
```





## Convertir probabilidad en clasificación

- 1) Una de las principales aplicaciones de un modelo de regresión logística es clasificar la variable cualitativa en función de valor que tome el predictor.
- 2) Sin embargo una función más importante es la de estimación de efectos de las variables predictoras.
- 3) Constituye una de las alternativas de clasificación supervisada.
- 4) Para conseguir esta clasificación, es necesario establecer un threshold (umbral) de probabilidad a partir de la cual se considera que la variable pertenece a uno de los niveles. Por ejemplo, se puede asignar una observación al grupo 1 si  $\hat{p}(Y = 1|X) > 0.5$  y al grupo 0 si de lo contrario.