# Práctica 5: Modelos Lineales Generalizados

## Modelos Lineales

Grados en Estadística y Matemáticas

# Índice

# Regresión Logística y regresión probit

# Regresión Logística y regresión probit

## Ejercicio 1

En el archivo `chdage.dat` aparecen los datos de edad (`age`) y de presencia o ausencia de una enfermedad coronaria (`chd`) de 100 sujetos seleccionados para participar en un estudio. Nuestro interés se centra en estudiar la relación entre la edad y la presencia de la enfermedad coronaria. Se pide:

1. Ajusta a estos datos un modelo de regresión logística, dando las estimaciones de los parámetros, así como intervalos de confianza al 95 % para los mismos.
2. ¿Qué podemos decir de la bondad del ajuste?
3. ¿Podemos decir que la edad es un factor de riesgo significativo en la presencia de la enfermedad coronaria?
4. ¿Qué probabilidad predice el modelo de presentar la enfermedad para una persona de 28 años? ¿Y para uno de 76?

## Ejercicio 1

**Apartado 1**

Variable respuesta: chd. Datos: $y_1, \ldots, y_n$
Variable predictora: age. Datos: $x_1, \ldots, x_n$.

La variable respuesta tiene distribución de Bernoulli: $y_i \sim B(1, p_i)$

Modelo de regresión logística:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i \quad i = 1, \ldots, n$$

Modelo de regresión probit:

$$F^{-1}(p_i) = \beta_0 + \beta_1 x_i \quad i = 1, \ldots, n$$

siendo $F(\cdot)$ la función de distribución de una $N(0, 1)$.

## Ejercicio 1

**Apartado 4**

$H_0$ : el modelo SÍ es adecuado para describir el conjunto de datos
$H_1$ : el modelo NO es adecuado para describir el conjunto de datos

Para realizar este contraste necesitamos introducir un concepto de gran importancia en Modelos Lineales Generalizados, el concepto de Deviance.

## Deviance

Para el contraste anterior, se compara la verosimilitud del modelo propuesto con la verosimilitud del modelo maximal o saturado.

**Modelo saturado**. Modelo con la misma distribución y función de enlace que el modelo de interés, donde el número de parámetros es igual al total de observaciones. Se considera que este modelo proporciona una interpretación completa de los datos.

El estadístico de la razón de verosimilitudes es

$$\Lambda = \frac{L(\hat{\boldsymbol{\beta}}_{max}; \mathbf{Y})}{L(\hat{\boldsymbol{\beta}}; \mathbf{Y})} \quad \text{o} \quad \log \Lambda = \ell(\hat{\boldsymbol{\beta}}_{max}; \mathbf{Y}) - \ell(\hat{\boldsymbol{\beta}}; \mathbf{Y})$$

siendo $L(\hat{\boldsymbol{\beta}}_{max}; \mathbf{Y})$, $L(\hat{\boldsymbol{\beta}}; \mathbf{Y})$ las funciones de verosimilitud del modelo saturado y del modelo ajustado, respectivamente.

## Deviance

A partir de $\log \Lambda$, se define la Deviance como

$$D = 2 \log \Lambda = 2(\ell(\hat{\beta}_{max}; \mathbf{Y}) - \ell(\hat{\beta}; \mathbf{Y}))$$

Se obtiene que si $H_0$ es cierta:

$$D \overset{aprox}{\sim} \chi^2(n-p)$$

En consecuencia, se rechaza $H_0$ al nivel $\alpha$ si $D \geq \chi^2_{n-p,\alpha}$.

La deviance está relacionada con otras medidas de la bondad de ajuste como AIC o BIC que ya vimos en el Modelo de Regresión Lineal.

## Ejercicio 1

**Apartado 2**

$H_0$ : el modelo SÍ es adecuado para describir el conjunto de datos
$H_1$ : el modelo NO es adecuado para describir el conjunto de datos

Deviance del modelo: $D$ =107.35
Grados de Libertad: 98

$$\chi^2_{98,\ 0.05} = 122.10$$

Como $D < \chi^2_{98,\ 0.05}$, podemos decir que el modelo ajusta bien a los datos.

## Ejercicio 1

### Apartado 3

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

El test para este contraste se puede dar utilizando el intervalo de confianza.

### Apartado 4

Para un nuevo valor $x_0$ de la variable AGE, la probabilidad predicha es:

$$\hat{p}_0 = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_0}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_0}}$$

# Regresión Logística y regresión probit

## Ejercicio 2

En enero de 1986, el transbordador espacial Challenger explotó poco después del lanzamiento. Se inició una investigación sobre la causa del accidente y la atención se centró en las juntas de goma (orings) de los cohetes aceleradores. A temperaturas más bajas, la goma se vuelve más frágil y es un sellador menos eficaz. En el momento del lanzamiento, la temperatura era de 31°F. ¿Podría haberse predicho el fallo de las juntas? En las 23 misiones anteriores del transbordador de las que existen datos, se registraron algunas evidencias de daños por soplado y erosión en algunas juntas. Cada transbordador tenía dos aceleradores, cada uno con tres juntas. Para cada misión, conocemos el número de juntas (de un total de seis) que mostraban algún daño y la temperatura de lanzamiento. Los datos se encuentran en el archivo `orings.dat`

1. Ajustar modelos de regresión logística y regresión probit a estos datos Obtener los intervalos de confianza al 95 % para los parámetros de estos modelos.
2. ¿Podemos decir que los modelos ajusta suficientemente bien a los datos?
3. ¿Podemos decir que la temperatura es un factor de riesgo significativo en la probabilidad de rotura de las juntas de goma?
4. ¿Qué probabilidad de morir predice el modelo para una temperatura de 31ºF?

# Regresión Logística y regresión probit

## Ejercicio 3

El archivo `icu.dat` contiene los datos de 200 individuos que formaron parte de estudio mucho mayor sobre la supervivencia de los pacientes en una Unidad de Cuidados Intensivos (UCI). La descripción de cada una de las variables aparece en la hoja anexa. El objetivo del estudio era la predicción de la probabilidad de supervivencia de un paciente tras ser ingresado en la UCI.

1. Ajustar un modelo de regresión logística de la variable `sta` sobre las variables `age`, `can`, `cpr`(Cardiopulmonary Resuscitation), `inf`, y `race`. Obtener los intervalos de confianza al 95 % para dichos parámetros.
2. ¿Podemos decir que el modelo ajusta suficientemente bien?
3. ¿De entre las variables que hemos elegido nosotros cuáles resultan significativas? Si utilizamos un procedimiento *forward* o *backward*, ¿qué modelo sería el resultante?
4. ¿Qué probabilidad de morir predice el modelo para un paciente de 47 años, de raza blanca, que no tenga cáncer, pero que haya requerido de CPR y que tenga síntomas de infección?

# Regresión de Poisson

# Regresión de Poisson

## Ejercicio 4

El número de muertes por Sida en Australia en períodos de tres meses entre 1983 y 1986 aparecen en el archivo `sida.dat`.

1. Ajustar un modelo de regresión de Poisson a dichos datos: determinar los coeficientes e intervalos de confianza al 95 % para los mismos.
2. Determinar la bondad del ajuste.
3. ¿Resulta significativa la variable periodo en la mortalidad?
4. ¿Qué mortalidad predecía el modelo para el año 1987?

## Ejercicio 4

**Apartado 1**

Variable respuesta: MUERTOS. Datos: $Y_1, \ldots, Y_n$
Variable predictora: PERIODO. Datos: $x_1, \ldots, x_n$.

La variable respuesta tiene distribución de Poisson: $Y_i \sim Poisson(\lambda_i)$

Modelo de regresión de Poisson (link= log):

$$\log \lambda_i = \beta_0 + \beta_1 x_i \quad i = 1, \ldots, n$$

Modelo de regresión de Poisson (link= sqrt):

$$\sqrt{\lambda_i} = \beta_0 + \beta_1 x_i \quad i = 1, \ldots, n$$

## Ejercicio 4

**Apartado 2: link= log**

$H_0$ : el modelo SÍ es adecuado para describir el conjunto de datos
$H_1$ : el modelo NO es adecuado para describir el conjunto de datos

Deviance del modelo: $D$ =29.654
Grados de Libertad: 12

$$\chi^2_{12,\ 0.05} = 21.026$$

Como $D > \chi^2_{12,\ 0.05}$, podemos decir que el modelo no ajusta bien a los datos.

## Ejercicio 4

**Apartado 2: link= sqrt**

$H_0$ : el modelo SÍ es adecuado para describir el conjunto de datos
$H_1$ : el modelo NO es adecuado para describir el conjunto de datos

Deviance del modelo: $D$ =16.905
Grados de Libertad: 12

$$\chi^2_{12,\ 0.05} = 21.026$$

Como $D < \chi^2_{12,\ 0.05}$, podemos decir que este modelo sí ajusta bien a los datos.

## Ejercicio 4

### Apartado 3

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

El test para este contraste se puede dar utilizando el intervalo de confianza.

### Apartado 4

Para un nuevo valor $x_0$ de la variable PERIODO, la media de mortalidad sería:

$$\hat{\lambda}_0 = e^{\hat{\beta}_0 + \hat{\beta}_1 x_0}$$

## Ejercicio 5

Para 30 islas de las Galápagos, disponemos de un recuento del número de especies de plantas encontradas en cada isla y del número de especies que son endémicas de dicha isla. También contamos con cinco variables geográficas para cada isla y dos climáticas. Se trata de determinar qué variables geográficas influyen en el número de tortugas. Se pide:

1. Ajustar un modelo de regresión de Poisson a dichos datos: determinar los coeficientes e intervalos de confianza al 95 % para los mismos.
2. Determinar la bondad del ajuste.
3. Determinar qué variables resultan significativas y realiza un procedimiento de selección de variables.
4. ¿Resulta adecuado el modelo para realizar predicciones? Razona la respuesta.

## Ejercicio 6

Los datos del archivo cancer.dat muestran el número de supervivientes en un estudio sobre cáncer de pulmón. En los datos aparecen los intervalos en que se hizo el seguimiento, el tipo de histología, el estadío de la enfermedad y el período de seguimiento (en meses). La variable de interés el el recuento del número de muertes. Estos recuentos son tratados como variables de Poisson. Se pide:

1. Ajustar un modelo de regresión de Poisson a dichos datos: determinar la significación de los distintos factores en la mortalidad.
2. Determinar la bondad del ajuste.
3. Determinar qué elmentos del modelo resultan significativos y realiza un procedimiento de selección de variables.

# Regresión Gamma

# Regresión Gamma

## Ejercicio 7

En el archivo `clot.dat` aparecen los datos del tiempo de coagulación (`time`, en segundos), para plasma normal diluido con diferentes concentraciones de plasma sin protombina (conc). La coagulación fue inducida por dos lotes de tromboplastina (lot)

1. Ajustar un modelo de regresión de Gamma, usando la inversa como función de enlace y el logaritmo de la concentración como variable predictora: determinar los coeficientes del modelo, especificando si hay diferencia significativa para ambos lotes.
2. ¿Resulta significativa la influencia de la variable concentración y del lote en el tiempo de coagulación?
3. Determinar la bondad del ajuste del modelo elegido.

## Ejercicio 7

**Apartado 1**

Variable respuesta: TIME. Datos: $Y_1, \ldots, Y_n$
Variables predictoras: log(COAG). Datos: $x_1, \ldots, x_n$.
LOT. Datos: "one" o "two"

La variable respuesta tiene distribución Gamma: $Y_i \sim \Gamma(\phi, \theta_i)$

$$\mu_i = E[Y_i] = \phi\theta_i \ , \quad Var[Y_i] = \phi\theta_i^2 = \frac{\mu_i^2}{\phi}$$

Modelo de regresión Gamma con función de enlace inversa :

$$\frac{1}{\mu_i} = \beta_0 + \beta_1 x_i + \tau lot.two + \delta(x_i * lot.two) \quad i = 1, \ldots, n$$

siendo *lot.two* la variable dummy que vale 0 si `lot=1` y 1 si `lot=2`.

## Ejercicio 7

**Apartado 2**

$$H_0 : \delta = 0$$
$$H_1 : \delta \neq 0$$

El test para este contraste se puede dar utilizando el intervalo de confianza.

$$H_0 : \tau = \delta = 0$$
$$H_1 : \tau = \delta \neq 0$$

Comparamos la deviance del modelo original y del reducido.

## Ejercicio 7

**Apartado 3**

$H_0$ : el modelo SÍ es adecuado para describir el conjunto de datos
$H_1$ : el modelo NO es adecuado para describir el conjunto de datos

Deviance del modelo: $D$ =0.029401
Parámetro de dispersión: $\hat{\phi}$ =0.002129707
Grados de Libertad: 14

$$\chi^2_{14,0,05} = 23{,}68479$$

Como $D/\hat{\phi} = 13{,}80519 < \chi^2_{14,0,05}$, podemos decir que el modelo ajusta bien a los datos.