

# Regresión logística y regresión de Poisson

Sesión 6. Curso: 'Análisis estadístico aplicado con R Commander'

Paqui Corpas Burgos  
correo: `corpas_fra@gva.es`

- 1 Regresión logística
- 2 Regresión de Poisson
- 3 Ejercicio práctico

# Regresión logística

# Regresión logística

Los modelos de regresión tienen como objetivo estudiar el efecto de una o más variables explicativas (independientes) sobre una variable respuesta (dependiente). En muchas situaciones, la **variable respuesta** que se quiere explicar es binaria (**sólo toma dos valores posibles**).

## Ejemplo

- Un estudio en el que se quiere estudiar **si una persona contrae o no una enfermedad coronaria** según su edad, sexo, historial fumador, nivel de colesterol, peso corporal y presión sanguínea.
- Un estudio en el que se quiere estudiar **si un insecto sigue vivo o no** después de la aplicación de un insecticida en base a una serie de covariables como la dosis de insecticida, las condiciones ambientales, etc.
- Un estudio donde se quiere estudiar **si padecer cáncer de pulmón** depende del número de cigarrillos consumidos y otros rasgos de los individuos como la edad, el sexo, la alimentación, etc.

La regresión logística es la técnica más usual para el análisis de datos de respuesta binaria.

# Interpretación de una variable respuesta binaria

Supongamos que la **variable respuesta** que queremos estudiar  $Y$  **representa la ocurrencia o no de un suceso**.

Se dice que una variable sigue una **distribución de Bernoulli de parámetro  $p$  cuando sólo tiene dos posibles resultados**: “éxito” (ocurre el suceso) y “fracaso” (no ocurre el suceso), siendo  $p$  la probabilidad de éxito.

Si  $Y_i \sim \text{Bernoulli}(p_i)$ , su distribución de probabilidad es:

$Y_i$	Probabilidad
1 (“éxito”)	$P(Y_i = 1) = p_i$
0 (“fracaso”)	$P(Y_i = 0) = 1 - p_i$

- $E(Y_i) = p_i$
- $\text{Var}(Y_i) = p_i(1 - p_i)$

En los modelos de regresión logística, se pretende estudiar si la probabilidad de que ocurra un suceso ( $p_i$ ) depende o no de otra u otras variables.

# Problemática con una regresión lineal

Supongamos que queremos ajustar un modelo de regresión lineal simple:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad Y_i = 0, 1, \quad \epsilon_i \sim N(0, \sigma^2)$$

Como los errores tienen media 0:

$$E(Y_i) = \beta_0 + \beta_1 X_i = p_i = P(Y_i = 1)$$

Si utilizamos una variable respuesta binaria con un predictor lineal como el anterior aparecen varios **problemas**:

- **Errores no normales:** como la variable respuesta sólo toma dos valores, los errores  $\epsilon_i$  no son normales.
- **Varianza del error no constante:**  $\sigma^2(Y_i) = p_i(1 - p_i) = (\beta_0 + \beta_1 X_i)(1 - \beta_0 - \beta_1 X_i)$ .
- **Restricciones en el predictor lineal:** como  $p_i$  es una probabilidad, el predictor lineal debe estar entre 0 y 1:

$$0 \leq E(Y_i) = (\beta_0 + \beta_1 X_i) \leq 1$$

Es probable que algunos valores de las variables explicativas conlleven a valores erróneos de  $p_i$  (fuera del intervalo  $[0, 1]$ ).

- La **relación lineal** entre  $p_i$  y la variable explicativa  $X_i$  **no suele verificarse**.

# Problemática con una regresión lineal

**Para resolver este problema** y poder relacionar  $p_i$  con el predictor lineal **es necesario considerar una transformación sobre  $p_i$  que evite los problemas anteriores**. Una de las opciones más utilizadas es la **transformación logística**:

$$\text{logit}(p_i) = \log \left( \frac{p_i}{1 - p_i} \right) = \beta_0 + \beta_1 X_i$$

El cociente  $\left( \frac{p_i}{1 - p_i} \right)$  se denomina “**odds**” e indica cuánto más probable es el éxito que el fracaso.

# Modelo de regresión logística

- Variable respuesta  $Y$  que toma dos valores 0 y 1 indicando fracaso o éxito  
 $Y \sim \text{Bernoulli}(p)$  ( $p$  probabilidad de éxito).
- Variables explicativas  $X_1, \dots, X_p$ .

**El modelo logístico establece la siguiente relación** entre la media de la variable respuesta (probabilidad de que ocurra el suceso  $p_i$ ) y las variables explicativas del estudio:

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip}, \quad i = 1, \dots, n$$

- Si  $\beta_j \approx 0$ , la probabilidad de que ocurra el suceso es independiente de la variable explicativa  $X_j$ .
- Si  $\beta_j > 0$ , la probabilidad de que ocurra el suceso aumenta conforme aumentan los valores de la variable explicativa  $X_j$ .
- Si  $\beta_j < 0$ , la probabilidad de que ocurra el suceso disminuye conforme aumentan los valores de la variable explicativa  $X_j$ .



# Interpretación de los parámetros del modelo

## Modelo logístico:

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{i1} + \dots, \beta_p X_{ip}, \quad i = 1, \dots, n$$

**La interpretación de los parámetros del modelo suele realizarse a partir de su exponencial:**

- $\exp(\beta_0)$  representa el valor del "odds" ( $\frac{p_i}{1-p_i}$ ) de que ocurra el suceso cuando las variables explicativas valen 0 (si éstas son cuantitativas) ó en sus categorías de referencia (si éstas son categóricas).
- $\exp(\beta_j)$ ,  $j = 1, \dots, p$  es una "odds ratio" y representa el cambio que se produce en la "odds" de que ocurra el suceso cuando la variable explicativa  $X_j$  se incrementa en una unidad (si ésta es cuantitativa) o al pasar a la categoría correspondiente de la variable (cuando ésta es categórica).
  - Si  $\exp(\beta_j) \approx 1$ , la probabilidad de que ocurra el suceso es independiente de la variable explicativa  $X_j$ .
  - Si  $\exp(\beta_j) > 1$ , la probabilidad de que ocurra el suceso aumenta conforme aumentan los valores de la variable explicativa  $X_j$ .
  - Si  $\exp(\beta_j) < 1$ , la probabilidad de que ocurra el suceso disminuye conforme aumentan los valores de la variable explicativa  $X_j$ .

En ocasiones, nos encontramos con situaciones en las que los datos no son exactamente binarios sino que vienen acumulados como el **número de éxitos sobre un determinado número de pruebas**.

## Ejemplo

Un estudio en el que se aplican diferentes dosis de un insecticida a varios insectos y se observa como variable respuesta cuántos siguen vivos y cuántos no tras la aplicación de cada dosis.

**Tanto los datos binarios como los binomiales se pueden explicar en base a otras covariables utilizando un modelo de regresión logística.**

# Modelo de regresión logística con datos binomiales

- Variable respuesta  $Y_j$  es el número de éxitos observado en el grupo  $j$  sobre un total de  $m_j$  pruebas (indicando  $p_j$  la probabilidad de éxito):

$$Y_j \quad (j = 1, \dots, g) \sim \text{Binomial}(m_j, p_j)$$

por lo que  $E(Y_j) = m_j p_j$ .

- Variables explicativas  $X_1, \dots, X_p$ .

**El modelo logístico establece la siguiente relación** entre la probabilidad de que ocurra el suceso en el grupo  $j$  y las variables explicativas del estudio:

$$\text{logit}(p_j) = \log \left( \frac{p_j}{1 - p_j} \right) = \beta_0 + \beta_1 X_{j1} + \dots + \beta_p X_{jp}, \quad j = 1, \dots, g$$

La interpretación de los coeficientes es similar al modelo anterior.

# Construcción del mejor modelo

Una vez planteado un modelo de regresión con todas las variables explicativas del estudio, el siguiente paso es el desarrollo de estrategias para **seleccionar las variables que mejor explican a la variable respuesta**. Adoptaremos el principio de parsimonia que consiste en **seleccionar el modelo que con menor número de parámetros se ajuste bien a los datos y lleve a una interpretación sencilla**.

El procedimiento que utilizaremos en este curso para la inclusión o eliminación de variables es el de **selección paso a paso** tanto hacia adelante (**forward**) como hacia atrás (**backward**).

# Construcción del mejor modelo

- **Forward:** Se parte de un modelo inicial sin ninguna variable explicativa. En el primer paso, se ajustan todos los modelos que resultan de la inclusión en el modelo inicial de cada una de las variables explicativas. Entrará en el modelo aquella variable para la que se obtiene un mejor ajuste en base a algún criterio. Se repite este procedimiento hasta llegar a un paso en el que ninguna variable mejore el ajuste del modelo seleccionado en el paso anterior.
- **Backward:** Se parte de un modelo inicial con todas las variables explicativas del estudio. En el primer paso, se ajustan todos los modelos que resultan de la eliminación del modelo inicial de cada una de las variables explicativas. Se eliminará del modelo aquella variable para la que se obtiene un mejor ajuste en base a algún criterio. Se repite este procedimiento hasta llegar a un paso en el que eliminar alguna variable no mejore el ajuste del modelo seleccionado en el paso anterior.

Los criterios más usuales para decidir la inclusión o eliminación de una variable en el modelo son el *Criterio de Información de Akaike (AIC)* y el *Criterio de Información Bayesiano (BIC)*. **El AIC y el BIC son medidas que permiten valorar la bondad del ajuste de un modelo y penalizan su complejidad**, de forma que se preferirán aquellos modelos que tengan un valor de AIC o BIC menor.

# Valoración del ajuste del modelo

Una medida que permite valorar la calidad del ajuste de un modelo de regresión logística es la tasa de clasificaciones correctas.

La **tasa de clasificaciones correctas** es la **proporción de individuos clasificados correctamente por el modelo** (cociente entre el nº de individuos clasificados correctamente y el número total de individuos).

Un individuo es clasificado correctamente por el modelo cuando su valor observado de respuesta (1 ó 0) coincide con su valor estimado por el modelo en base a las variables explicativas. Para asignar respuesta  $Y = 1$  ó  $Y = 0$  bajo el modelo a los datos, se elige un punto de corte (*cut-point*)  $p \in (0, 1)$ :

- A un individuo  $i$  se le estima respuesta  $Y_i = 1$  si  $\hat{p}_i \geq p$ .
- A un individuo  $i$  se le estima respuesta  $Y_i = 0$  si  $\hat{p}_i < p$ .

Como punto de corte para clasificar las observaciones se suele elegir 0.5 ó aquel que proporciona la máxima tasa de clasificaciones correctas.

# Curva ROC

La **curva ROC** es una gráfica que **permite evaluar la capacidad de un modelo para discriminar**. El área bajo la curva ROC representa la probabilidad de que un individuo que presenta el suceso (elegido al azar) tenga mayor probabilidad estimada de presentarlo que un individuo que no lo presenta (elegido también al azar).

Para construir la curva ROC se dispone de un test diagnóstico (en nuestro caso, el modelo logístico) que se usa para detectar si los individuos de una población tienen cierta característica ( $Y = 1$ ).

Cuando el test diagnóstico es el modelo de regresión logística ajustado, a un individuo se le predice  $\hat{Y} = 1$  si su probabilidad predicha  $\hat{p}$  es mayor o igual que el *cut-point* elegido para discriminar.

- **Sensibilidad:**  $P(\hat{Y} = 1 | Y = 1)$ .
- **Especificidad:**  $P(\hat{Y} = 0 | Y = 0)$ .

La curva ROC se obtiene representando gráficamente la *sensibilidad* y *1-especificidad* ( $P(\hat{Y} = 1 | Y = 0)$ ) para distintos test diagnósticos (en nuestro caso, distintos *cut-point*).

Cuando el área bajo la curva ROC es al menos 0.7, el modelo logit ajustado se considera preciso con capacidad de discriminación alta.

# Test de Hosmer-Lemeshow

El test de Hosmer-Lemeshow permite evaluar la bondad de ajuste de un modelo de regresión logística comparando el número predicho de eventos ( $\hat{Y} = 1$ ) con el número observado ( $Y = 1$ ) en  $G$  grupos de individuos (habitualmente 10) establecidos en base a las probabilidades estimadas de que ocurra el evento (de menor a mayor).

**Estadístico Hosmer-Lemeshow:**

$$H_g = \sum_{g=1}^G \frac{(o_g - n_g \bar{p}_g)^2}{n_g \bar{p}_g (1 - \bar{p}_g)}$$

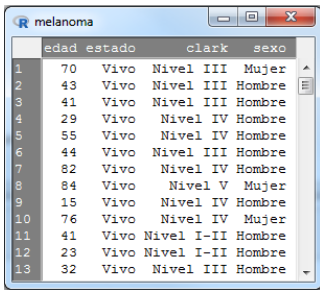
- $o_g$ : número de observaciones con  $Y = 1$  en el  $g$ -ésimo grupo.
- $n_g$ : número total de observaciones en el  $g$ -ésimo grupo.
- $\bar{p}_g$ : probabilidad estimada de respuesta  $Y = 1$  bajo el modelo para el  $g$ -ésimo grupo (media de las probabilidades  $\hat{p}$  en dicho grupo).

Si el ajuste del modelo es bueno, el estadístico Hosmer-Lemeshow sigue una distribución  $\chi^2$  con  $G - 2$  grados de libertad.



## Ejemplo: Base de datos *melanoma*

Una base de datos con 122 individuos (filas) y 4 variables (columnas).



	edad	estado	clark	sexo
1	70	Vivo	Nivel III	Mujer
2	43	Vivo	Nivel III	Hombre
3	41	Vivo	Nivel III	Hombre
4	29	Vivo	Nivel IV	Hombre
5	55	Vivo	Nivel IV	Hombre
6	44	Vivo	Nivel III	Hombre
7	82	Vivo	Nivel IV	Hombre
8	84	Vivo	Nivel V	Mujer
9	15	Vivo	Nivel IV	Hombre
10	76	Vivo	Nivel IV	Mujer
11	41	Vivo	Nivel I-II	Hombre
12	23	Vivo	Nivel I-II	Hombre
13	32	Vivo	Nivel III	Hombre

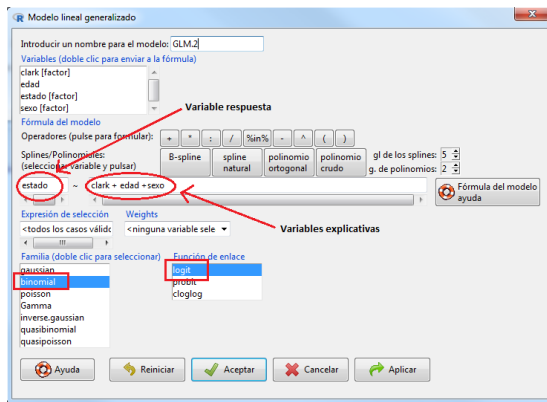
**Objetivo:** estudiar si la probabilidad de morir por cáncer de piel se relaciona con el nivel de Clark (profundidad a la que se encuentra el cáncer), la edad y el sexo de los individuos. Para estudiar esta relación emplearemos un modelo de regresión logística en el que:

- **Variable respuesta:** estado (Vivo/Muerto).
- **Variables explicativas:** clark, edad y sexo.

# Modelo de regresión logística con R Commander

Para ajustar un modelo de regresión logística con **R Commander** seleccionaremos:

Estadísticos > Ajuste de modelos > Modelo lineal generalizado



Una vez especificada la variable respuesta y las variables explicativas, se pulsa **Aceptar**.

# Resultados

```
Call:
glm(formula = estado ~ sexo + clark + edad, family = binomial(logit),
    data = melanoma)
```

Deviance Residuals:

	Min	1Q	Median	3Q	Max
	-1.408	-0.877	-0.585	1.041	2.005

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-2.761186	1.001912	-2.756	0.00585 **
sexo[T.Mujer]	0.601454	0.426494	1.410	0.15847
clark[T.Nivel III]	0.705286	0.858233	0.822	0.41120
clark[T.Nivel IV]	1.509601	0.871332	1.733	0.08318 .
clark[T.Nivel V]	1.869025	0.831159	2.249	0.02453 *
edad	0.009752	0.013073	0.746	0.45565

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 157.09 on 121 degrees of freedom  
Residual deviance: 140.46 on 116 degrees of freedom  
AIC: 152.46

Number of Fisher Scoring iterations: 4

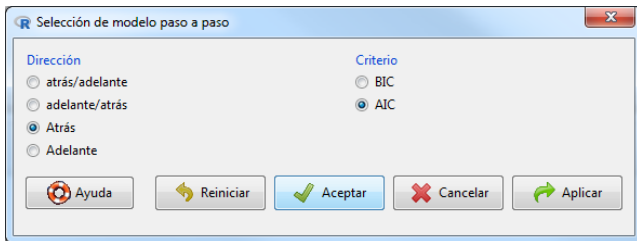
- En la columna  $\text{Pr}(>|z|)$  se muestran los p-valores correspondientes a los contrastes de significación de los coeficientes del modelo. Si un p-valor es inferior a 0.05 (se rechaza  $H_0 : \beta_j = 0$ ), concluiremos que existe una relación significativa entre la variable explicativa y la probabilidad de morir por cáncer de piel.

- Como podemos observar no todas las variables explicativas muestran una relación significativa con la probabilidad de morir por cáncer de piel y, por tanto, podríamos plantearnos la eliminación de algunas variables del modelo.

# Construcción del mejor modelo

Tras haber ajustado el modelo con todas la variables explicativas, podemos aplicar un procedimiento automático de selección de variables paso a paso en **R Commander** seleccionando:

Modelo > Selección de modelo paso a paso



# Selección del mejor modelo

```
Step: AIC=151.02  
estado ~ sexo + clark
```

	Df	Deviance	AIC
<none>		141.02	151.02
- sexo	1	143.36	151.36
- clark	3	152.29	156.29

```
Call: glm(formula = estado ~ sexo + clark, family = binomial(logit),  
data = melanoma)
```

Coefficients:

(Intercept)	sexo[T.Mujer]	clark[T.Nivel III]	clark[T.Nivel IV]	clark[T.Nivel V]
-2.2917	0.6431	0.7248	1.5842	1.9743

Degrees of Freedom: 121 Total (i.e. Null); 117 Residual

Null Deviance: 157.1

Residual Deviance: 141 AIC: 151

El modelo que proporciona un mejor ajuste en base al estadístico AIC es aquel que tiene las variables explicativas: sexo y clark.

# Resultados del modelo final

Ajustando de nuevo el modelo con las variables explicativas seleccionadas, se obtiene:

```
Call:
glm(formula = estado ~ sexo + clark, family = binomial(logit),
    data = melanoma)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3184  -0.8952  -0.6157   1.0426   1.9103

Coefficients:
(Intercept)          -2.2917    0.7709   -2.973    0.00295 **
sexo[T.Mujer]         0.6431    0.4221   1.524    0.12763
clark[T.Nivel III]    0.7248    0.8577   0.845    0.39805
clark[T.Nivel IV]     1.5842    0.8662   1.829    0.06742 .
clark[T.Nivel V]      1.9743    0.8201   2.407    0.01607 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 157.09  on 121  degrees of freedom
Residual deviance: 141.02  on 117  degrees of freedom
AIC: 151.02

Number of Fisher Scoring iterations: 4
```

- El p-valor asociado al contraste de significación de la variable "clark[T.Nivel V]" es  $0.016 < 0.05$ , por tanto, concluimos que **existe asociación entre el nivel de clark y la probabilidad de morir por cáncer de piel**. La estimación del coeficiente que acompaña a la variable "clark[T.Nivel V]" es positiva, lo que nos indica que la probabilidad de morir por cáncer de piel aumenta cuando el nivel de clark es V en comparación con el nivel I-II (categoría de referencia).
- No se observan diferencias significativas entre la probabilidad de morir por cáncer de piel de los pacientes con nivel de clark III y IV en comparación con los pacientes con nivel de clark I-II (categoría de referencia).
- La variable "sexo", aunque permanece en el modelo, no muestra relación significativa con la probabilidad de morir por cáncer de piel ( $p\text{-valor} = 0.1276 > 0.05$ ).

# Interpretación de los parámetros del modelo

La interpretación de los parámetros del modelo se realizará considerando la exponencial de los coeficientes estimados:

```
exp(coef(GLM.4)) # Exponentiated coefficients ('odds ratios')
(Intercept)      sexo[T.Mujer] clark[T.Nivel III] clark[T.Nivel IV] clark[T.Nivel V]
  0.1010919      1.9023363      2.0643479      4.8756160      7.2012276
```

- $\exp(\beta_0) = 0.101$ : indica cuánto más probable es morir por cáncer de piel frente a no morir cuando el sexo de los individuos es “hombre” y el nivel de clark es I-II (categorías de referencia de las v.explicativas). Puesto que  $0.101 < 1$ , concluimos que la probabilidad de morir por cáncer de piel es menor en los hombres con nivel de clark I-II.
- $\exp(\beta_4) = 7.201$ : indica que la “odds” de morir por cáncer de piel se multiplica aproximadamente por 7 en los pacientes con nivel de clark V en comparación con los pacientes con nivel de clark I-II. Esta asociación es significativa puesto que el p-valor (0.016) es inferior a 0.05.

# Intervalos de confianza

Podemos obtener los **intervalos de confianza** para los parámetros del modelo seleccionando en **R Commander**:

Modelos > Intervalos de confianza

```
> Confind(GLM.4, level = 0.95, type = "LR")
```

	Estimate	2.5 %	97.5 %
(Intercept)	-2.2917254	-4.1658681	-0.9903615
sexo[T.Mujer]	0.6430828	-0.1822876	1.4814492
clark[T.Nivel III]	0.7248144	-0.8209796	2.7039336
clark[T.Nivel IV]	1.5842464	0.0309844	3.5793264
clark[T.Nivel V]	1.9742515	0.5379218	3.9081522

También es posible determinar si existe asociación entre una variable explicativa y la probabilidad de morir por cáncer de piel a partir de los intervalos de confianza para los coeficientes del modelo ( $\beta_j$ ). **Si el intervalo no contiene al cero, concluiremos que existe una asociación significativa entre la variable explicativa y la probabilidad de morir por cáncer de piel.**



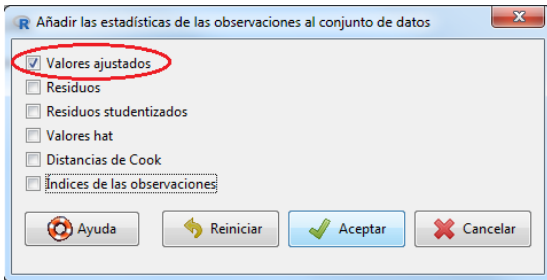
# Evaluación del modelo de regresión logística

Una forma de valorar si el ajuste de un modelo de regresión logística es bueno es ver cómo de bien predice, es decir, valorar el porcentaje de observaciones clasificadas correctamente.

En este ejemplo, consistiría en contar cuántas veces el modelo ajustado predice correctamente que la persona muere o no por cáncer de piel. Para ello:

- 1 Debemos añadir al conjunto de datos las probabilidades de morir estimadas por el modelo para cada uno de los individuos, seleccionando en **R Commander**:

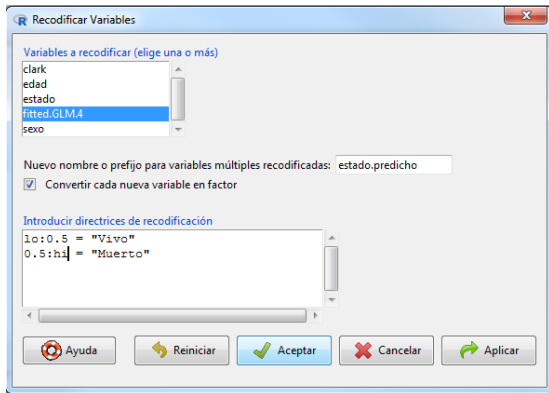
Modelos > Añadir las estadísticas de las observaciones junto a los datos



# Evaluación del modelo de regresión logística

- 2 A continuación, eligiendo como punto de corte  $p=0.5$  en la probabilidad de morir por cáncer de piel para predecir que si un individuo está vivo o no, recalculamos una nueva variable “estado.predicho” seleccionando en **R Commander**:

Datos > Modificar variables del conjunto de datos activo > Recodificar variables...



# Evaluación del modelo de regresión logística

- 3 Construimos una tabla de contingencia entre la variable respuesta y la predicha para obtener la tasa de clasificaciones correctas:

Frequency table:

	estado.predicho	
estado	Muerto	Vivo
Vivo	12	68
Muerto	16	26

Row percentages:

	estado.predicho		Total	Count
estado	Muerto	Vivo		
Vivo	15.0	85.0	100	80
Muerto	38.1	61.9	100	42

Eligiendo como punto de corte  $p=0.5$ :

- El modelo clasifica correctamente al 69 % de los individuos ( $\frac{16+68}{122}$ ).
- El modelo clasifica correctamente al 38.1 % de los individuos que están muertos (sensibilidad) y al 85 % de los individuos que están vivos (especificidad).

Para el *cut-point* seleccionado, la sensibilidad del modelo no es muy buena.

Para construir la curva ROC del modelo de regresión logística ajustado es necesario instalar y cargar el plugin `RcmdrPlugin.ROC`.

- Para instalarlo seleccionamos en la consola de R: Paquetes > Instalar paquetes(s)...
- Para cargarlo seleccionamos en R Commander: Herramientas > Cargar plugin(s) de Rcmdr...

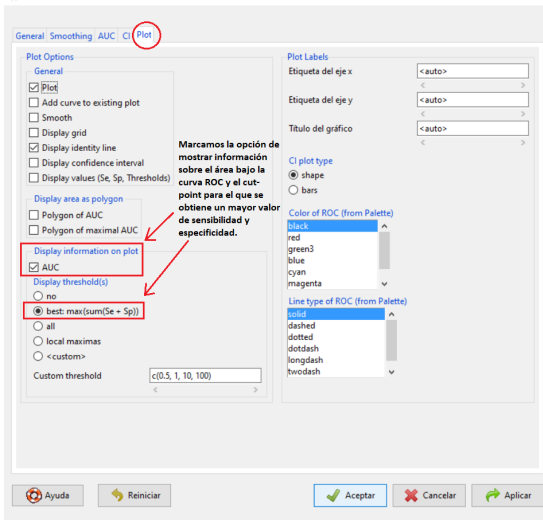
Una vez cargado el plugin, para construir la curva ROC seleccionamos en el menú de **R Commander**:

```
ROC > pROC > Plot ROC curve for logistic regression model...
```

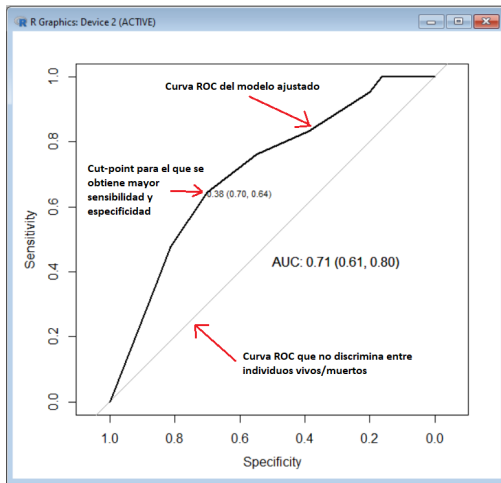
# Curva ROC

R Plot ROC curve

X



# Curva ROC



Para el cut-point  $p=0.38$ , la capacidad discriminatoria del modelo es buena (sensibilidad = 0.61, especificidad = 0.80).

# Test de Hosmer-Lemeshow

Para contrastar la bondad de ajuste del modelo de regresión logística ajustado mediante el test de Hosmer-Lemeshow seleccionamos en **R Commander**:

```
R0C > pROC > Hosmer-Lemeshow GOF test
```

```
Hosmer and Lemeshow goodness of fit (GOF) test  
data: .matrix[, 1], .matrix[, 2]  
X-squared = 4.306, df = 8, p-value = 0.8285
```

Puesto que el p-valor del test es 0.8285 (mayor que 0.05), no puede rechazarse que el número predicho de eventos por el modelo ( $\hat{Y} = 1$ ) es similar al número observado ( $Y = 1$ ) en los grupos establecidos. Concluimos por tanto que el ajuste del modelo de regresión logística es bueno.

# Regresión de Poisson



# Regresión de Poisson

En muchas situaciones **la variable que queremos explicar sólo toma valores enteros no negativos** (0, 1, 2, ...).

## Ejemplo

- Un estudio sobre el **número total de visitas al médico que una persona de edad avanzada ha hecho durante el año pasado** según su sexo, estado de salud, etc.
- Un estudio sobre el **número de quejas que recibe un profesional sanitario** según el número de horas que trabaja, sus ingresos, etc.
- Un estudio sobre el **número de muertes por cáncer en los municipios de una región** según sus condiciones ambientales, densidad de población, etc.

En todos estos casos, **la variable respuesta representa un recuento de sucesos o hechos** (poco frecuentes) y se quiere estudiar si ciertas variables explicativas influyen en la variable respuesta.

**La regresión de Poisson nos servirá para modelizar datos discretos que representan conteos de sucesos.**

# Interpretación de una variable respuesta Poisson

La distribución de Poisson se utiliza para modelizar **variables respuesta que sólo toman valores discretos que representan conteos** (0, 1, 2, ...).

Si  $Y \sim \text{Poisson}(\mu)$  entonces su distribución de probabilidad es:

$$P(Y = y) = \frac{\mu^y \exp(-\mu)}{y!}, \quad y = 0, 1, 2, \dots$$

donde:

- $\mu$  representa el número medio de ocurrencias.
- $E(Y) = \mu$  y  $\text{Var}(Y) = \mu$ .

# Problemática con una regresión lineal

Como la variable respuesta  $Y$  toma valores discretos no negativos, **no procede utilizar un modelo de regresión lineal directo para estudiar su relación con las variables explicativas:**

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad Y_i = 0, 1, 2, \dots$$

Ya que:

- La variable respuesta (y, por consiguiente, el error) no se distribuye según una Normal.
- La varianza de la variable respuesta no es constante:  $\text{Var}(Y_i) = \mu_i = \beta_0 + \beta_1 X_i$ .
- La relación entre la media de la variable respuesta no suele ser lineal con el predictor lineal.

Por tanto **se necesita una función de enlace para relacionar la media de la variable respuesta ( $\mu_i$ ) con el predictor lineal.**

# Modelo de regresión de Poisson

- Variable respuesta  $Y$  que representa el número de conteos de un suceso  
 $Y \sim \text{Poisson}(\mu)$  ( $\mu$  es el número medio de ocurrencias)
- Variables explicativas  $X_1, X_2, \dots, X_p$ .

El efecto de las variables explicativas sobre la respuesta se modeliza valorando su efecto sobre  $\mu$  y **la transformación que une el predictor lineal con la media de la variable respuesta es el logaritmo:**

$$\log(\mu_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip} \quad i = 1, \dots, n$$

- Si  $\beta_j \approx 0$ , el número de ocurrencias del suceso es independiente de la variable explicativa  $X_j$ .
- Si  $\beta_j > 0$ , el número de ocurrencias del suceso aumenta conforme aumentan los valores de la variable explicativa  $X_j$ .
- Si  $\beta_j < 0$ , el número de ocurrencias del suceso disminuye conforme aumentan los valores de la variable explicativa  $X_j$ .

## ¿Y si tenemos tasas de incidencia?

**El número de eventos observados puede depender de una variable “tamaño” que determine el número de oportunidades de que ocurra el evento:**

- No es igual detectar 5 casos de cáncer de próstata en un pueblo de 1000 habitantes que en una ciudad de 100000.
- El número de quejas que recibe un profesional sanitario dependerá del número de consultas realizadas.

A veces es posible analizar este tipo de datos con un modelo de regresión logística, por ejemplo si contamos el número de situaciones en las que ocurre el evento sobre un total posible (número de quejas sobre el total de consultas realizadas). Sin embargo, **si la proporción de eventos es pequeña, la aproximación por una Poisson es válida.**

# Modelo de regresión de Poisson para tasas

- Variable respuesta  $Y$  que representa el número de conteos de un suceso  $Y \sim \text{Poisson}(\mu)$  está influenciada por otra variable  $Z$  que actúa de “offset” (corrector) del valor de los conteos.
- Variables explicativas  $X_1, X_2, \dots, X_p$ .

El efecto de las variables explicativas sobre la respuesta se modeliza valorando su efecto sobre  $\mu$  y **la transformación que une el predictor lineal con la media de la variable respuesta es el logaritmo:**

$$\log(\mu_i) = \textcolor{red}{\log(Z_i)} + \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip} \quad i = 1, \dots, n$$

Esto es equivalente a modelizar el **riesgo o tasa de incidencia**  $(\frac{Y_i}{Z_i})$ .

# Interpretación de los parámetros del modelo

El resultado del ajuste del modelo nos da las estimaciones de los parámetros del predictor lineal:  $\beta_0, \beta_1, \dots, \beta_p$ .

**La interpretación de los parámetros asociados a cada variable explicativa suele realizarse a partir de su exponencial:**

- $\exp(\beta_j)$  representa el cambio en el riesgo de que ocurra el evento al aumentar una unidad la variable explicativa (si ésta es cuantitativa) o al pasar a la categoría correspondiente de la variable (si ésta es categórica).

Diremos que aumenta el riesgo cuando  $\exp(\beta_j) > 1$ , disminuye cuando  $\exp(\beta_j) < 1$  y existe igual riesgo cuando  $\exp(\beta_j) \approx 1$ .

# Valoración del ajuste del modelo

- El estadístico **DEVIANCE** mide la *desviación* entre el modelo ajustado y un modelo completo en el que el ajuste es perfecto. Cuanto mayor sea el estadístico **DEVIANCE** peor será el ajuste del modelo.
- En los programas estadísticos, junto con la DEVIANCE del modelo ajustado (`residual deviance`), también se muestra la DEVIANCE del modelo NULO (modelo sin ninguna variable explicativa), lo cual permite valorar la diferencia entre ambas.
- Si el ajuste del modelo es bueno, el estadístico DEVIANCE sigue una distribución  $\chi^2$  con  $(n - p - 1)$  grados de libertad (número de individuos - número de variables explicativas - 1). Si la DEVIANCE es mucho más grande que los grados de libertad, será indicación de que la hipótesis debe ser rechazada.
- Observaciones con **residuos** superiores a 2 en valor absoluto pueden indicar falta de ajuste.

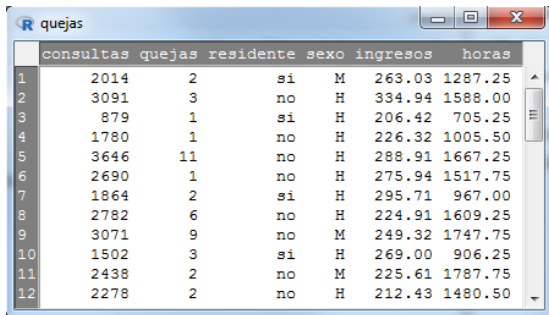


- Uno de los supuestos del modelo de regresión de Poisson es que **la media y la varianza de la variable respuesta son iguales**. En ocasiones, esto puede no cumplirse:
  - La varianza es mucho más grande que la media (SOBREDISPERSIÓN).
  - La varianza es más pequeña que la media (INFRADISPERSIÓN).
- Si la DEVIANCE es mucho más grande que los grados de libertad (falta de ajuste) puede indicar que existe un problema de sobredispersión/infradispersión.

Para solucionar problemas de sobredispersión/infradispersión se pueden utilizar **modelos de quasi-verosimilitud**, en los que se incorpora un nuevo parámetro en la variabilidad  $Var(Y_i) = \phi\mu_i$  y se utiliza una quasi-verosimilitud para estimar los parámetros del modelo.

## Ejemplo: Base de datos *quejas*

La base de datos *quejas* contiene información sobre el número de quejas a médicos del servicio de urgencias de un hospital. Se desea estudiar si existe relación entre el número de quejas que reciben los médicos con el hecho de que sean hombre o mujer, hayan sido residentes o no en un servicio de urgencias, sus ingresos y el número de horas y consultas que han realizado.



	consultas	quejas	residente	sexo	ingresos	horas
1	2014	2	si	M	263.03	1287.25
2	3091	3	no	H	334.94	1588.00
3	879	1	si	H	206.42	705.25
4	1780	1	no	H	226.32	1005.50
5	3646	11	no	H	288.91	1667.25
6	2690	1	no	H	275.94	1517.75
7	1864	2	si	H	295.71	967.00
8	2782	6	no	H	224.91	1609.25
9	3071	9	no	M	249.32	1747.75
10	1502	3	si	H	269.00	906.25
11	2438	2	no	M	225.61	1787.75
12	2278	2	no	H	212.43	1480.50

## Ejemplo: Base de datos *quejas*

- Variable respuesta  $Y_i$ : número de quejas que recibe el médico  $i$  del hospital ( $i = 1, \dots, n$ ).
- Variables explicativas: sexo, residente, ingresos, horas, consultas.

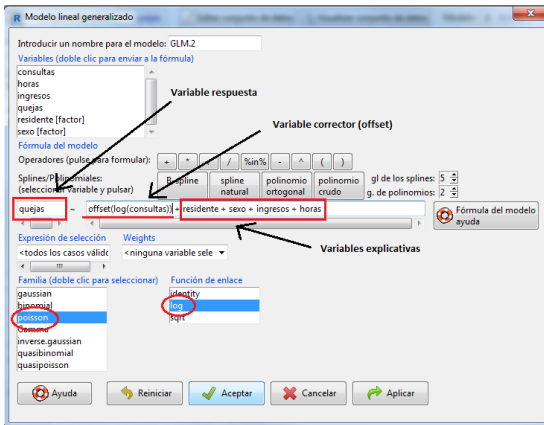
La variable *consultas* indica el número de consultas que ha realizado cada médico. Dicha variable actúa como corrector del valor de los conteos (*offset*), puesto que el número de quejas está limitado por el número de consultas realizadas.

**El hecho de que la variable respuesta  $Y_i$  sólo tome valores discretos que representan conteos permite utilizar la distribución de Poisson para modelizarla.** Para probabilidades de éxito pequeñas y totales grandes, la aproximación con una distribución de Poisson es válida.

# Modelo de regresión de Poisson con R Commander

Para ajustar un modelo de regresión de Poisson con R Commander seleccionaremos:

Estadísticos > Ajuste de modelos > Modelo lineal generalizado...



# Resultados

```
Call:
glm(formula = quejas ~ offset(log(consultas)) + residente + sexo +
    ingresos + horas, family = poisson(log), data = quejas)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.9434	-0.9490	-0.3130	0.7859	1.8037

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-7.9247861	0.8756756	-9.050	<2e-16 ***
residente[T.si]	-0.2090030	0.2011522	-1.039	0.2988
sexo[T.M]	-0.1954351	0.2181534	-0.896	0.3703
ingresos	0.0015761	0.0028295	0.557	0.5775
horas	0.0007019	0.0003505	2.002	0.0452 *

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 63.435 on 43 degrees of freedom  
Residual deviance: 54.518 on 39 degrees of freedom  
AIC: 187.3

Number of Fisher Scoring iterations: 5

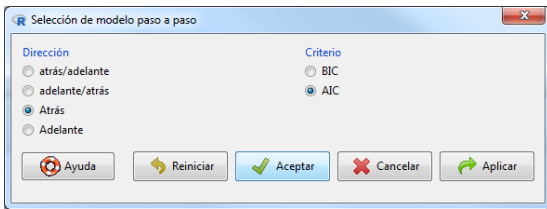
- En la columna  $\Pr(>|z|)$  se muestran los p-valores correspondientes a los contrastes de significación de los coeficientes del modelo. Si un p-valor es inferior a 0.05 (se rechaza  $H_0 : \beta_j = 0$ ), concluiremos que existe una relación significativa entre la variable explicativa y el número de quejas que reciben los médicos.

- Como podemos observar no todas las variables explicativas muestran una relación significativa con el número de quejas que reciben los médicos y, por tanto, podríamos plantearnos la eliminación de algunas variables del modelo.

# Construcción del mejor modelo

Tras haber ajustado el modelo con todas las variables explicativas, podemos aplicar un procedimiento automático de selección de variables paso a paso en **R Commander** seleccionando:

Modelo > Selección del modelo paso a paso



# Selección del mejor modelo

## Modelo > Selección del modelo paso a paso

```
Step: AIC=184.13
```

```
quejas ~ horas + offset(log(consultas))
```

	Df	Deviance	AIC
<none>	57.347	184.13	
- horas	1	63.435	188.22

```
Call: glm(formula = quejas ~ horas + offset(log(consultas)), family = poisson(log),  
data = quejas)
```

Coefficients:

(Intercept)	horas
-7.7422837	0.0007728

```
Degrees of Freedom: 43 Total (i.e. Null); 42 Residual
```

```
Null Deviance: 63.44
```

```
Residual Deviance: 57.35 AIC: 184.1
```

El modelo que proporciona un mejor ajuste en base al estadístico *AIC* es aquel que tiene la variable explicativa: horas.

# Resultados del modelo final

**Ajustando de nuevo el modelo con las variables explicativas seleccionadas, se obtiene:**

```
Call:
glm(formula = quejas ~ offset(log(consultas)) + horas, family = poisson(log),
    data = quejas)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9381  -1.0523  -0.3266   0.8939   1.9462

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.7422837   0.5083727  -15.230   <2e-16 ***
horas        0.0007728   0.0003252    2.376   0.0175 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 63.435  on 43  degrees of freedom
Residual deviance: 57.347  on 42  degrees of freedom
AIC: 184.13

Number of Fisher Scoring iterations: 5
```

- El p-valor asociado al contraste de significación de la variable “horas” es  $0.0175 < 0.05$ , por tanto, concluimos que **existe asociación entre el número de horas que trabajan los médicos y el número de quejas que reciben**. La estimación del coeficiente que acompaña a la variable “horas” es positiva, lo que nos indica que el número de quejas que reciben los médicos aumenta conforme aumenta también el número de horas que trabajan.
- No se observa una relación estadísticamente significativa entre el número de quejas que reciben los médicos y si éstos han sido o no residentes en un servicio de urgencias, su sexo y sus ingresos.



# Interpretación de los parámetros del modelo

La interpretación de los parámetros del modelo se realizará considerando la exponencial de los coeficientes estimados:

```
> exp(coef(GLM.2)) # Exponentiated coefficients
(Intercept)      horas
0.0004340791 1.0007731411
```

- $\exp(\beta_1) = 1.0008$ : indica que el riesgo de que un médico reciba una queja se multiplica por 1.0008 por cada incremento de una unidad en el número de horas que trabaja.

# Intervalos de confianza

Podemos obtener los **intervalos de confianza** para los parámetros del modelo seleccionando en **R Commander**:

Modelos > Intervalos de confianza

```
> Confint(GLM.2, level = 0.95, type = "LR")
              Estimate      2.5 %      97.5 %
(Intercept) -7.7422837321 -8.7817505177 -6.787055165
horas        0.0007728424  0.0001543203  0.001430378
```

Como ya vimos anteriormente, también es posible determinar si existe asociación entre una variable explicativa y la proporción de quejas que reciben los médicos a partir de los intervalos de confianza para los coeficientes del modelo ( $\beta_j$ ). **Si el intervalo no contiene al cero, concluiremos que existe una asociación significativa entre la variable explicativa y la proporción de quejas que reciben los médicos.**

# Evaluación del modelo de regresión de Poisson

Si observamos los *residuos* del modelo, vemos que:

- **Residuos deviance:**

Deviance Residuals:					
Min	1Q	Median	3Q	Max	
-1.9381	-1.0523	-0.3266	0.8939	1.9462	

No hay ninguna observación que presente un residuo grande (mayor o menor que 2).

# Evaluación del modelo de regresión de Poisson

Si observamos la *DEVIANCE residual* del modelo, vemos que:

- **Deviance residual:**

```
Null deviance: 63.435  on 43  degrees of freedom  
Residual deviance: 57.347  on 42  degrees of freedom  
AIC: 184.13
```

Para contrastar si el ajuste del modelo es bueno ( $\text{Deviance}=57.347 \sim \chi^2(42)$ ) podemos ejecutar el siguiente comando en la consola de R Commander:

```
> 1 - pchisq(57.347, 42)  
[1] 0.0574926
```

El p-valor del contraste  $\chi^2$  para la deviance residual no es inferior a 0.05, por tanto, no puede rechazarse que el ajuste del modelo es bueno. Sin embargo, la diferencia entre la deviance del modelo ajustado y la deviance del modelo nulo no es muy grande. Podríamos probar a añadir interacciones de orden 2 entre las variables y ver si se reduce la deviance residual del modelo.

## Ejercicio práctico