FEIR 45: Regresión logística

Apuntes del curso FEIR3, curso 2014/15 actualizados. `Última actualización:` martes 02 abril 2019,18:57:29

María Elvira Ferre Jaén

Índice

1. Introducción

Nos basamos para documento de introducción al modelo de regresión logística en González-Revaldería, Fernández, García, & Queraltó (2007) y A. Field, Miles, & Field (2012).

La regresión logística es el conjunto de modelos estadísticos utilizados cuando se desea conocer la relación entre

- Una **variable dependiente cualitativa**, dicotómica (regresión logística binaria o binomial) o con más de dos categorías (regresión logística multinomial).
- Una o más variables explicativas independientes, llamadas **covariables**, ya sean cualitativas o cuantitativas.

Las covariables cualitativas deben ser dicotómicas, tomando valor 0 para su ausencia y 1 para su presencia. Si la covariable tuviera más de dos categorías debemos realizar una transformación de la misma en varias covariables cualitativas dicotómicas ficticias (*variables dummy*). Al hacer esta transformación cada categoría de la variable entraría en el modelo de forma individual.

Los modelos de regresión logística tienen **tres finalidades:**

- Cuantificar la importancia de la relación existente entre cada una de las covariables y la variable dependiente.
- Clarificar la existencia de interacción y confusión entre covariables respecto a la variable dependiente (es decir, los odds ratio para cada covariable).
- Clasificar individuos dentro de las categorías (presente/ausente) de la variable dependiente.

Por tanto, el objetivo de la **regresión logística** no es, como en regresión lineal, predecir el valor de la variable Y a partir de una o varias variables predictoras (X_s), sino que queremos predecir la **probabilidad** de que ocurra Y conocidos los valores de las variables X_s. La ecuación general es de la forma

$$P(Y) = \frac{1}{1 + e^{-(b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_n X_n)}},$$

donde $P(Y)$ es la probabilidad de que ocurra Y, e es la función exponencial y el resto de coeficientes son análogos a los de la regresión lineal.

En su forma más sencilla, cuando tenemos sólo una variable predictora X_1, la ecuación de la regresión logística viene dada por

$$P(Y) = \frac{1}{1 + e^{-(b_0 + b_1 X_1)}}.$$

Los valores posibles de estas ecuaciones varían entre 0 y 1. Un valor cercano a 0 significa que es muy improbable que Y haya ocurrido, y un valor cercano a 1 significa que es muy probable que tuviese lugar.

Como en la regresión lineal cada variable predictora de la ecuación logística tiene su propio coeficiente. Los valores de los parámetros se estiman utilizando el *método de máxima verosimilitud* que selecciona los coeficientes que hacen más probable que los valores observados ocurran.

1.1. Interpretación del modelo de regresión logística

El propósito del análisis es

- Predecir la probabilidad de que un evento ocurra para una persona dada (notación $P(Y_i)$). Para dicha i-ésima persona, Y será 0 (la respuesta no ocurre) o 1 (la respuesta ocurre), y el valor predicho, $P(Y)$, tendrá un valor 0 (no hay probabilidad de que el resultado ocurra) o 1 (el resultado seguro que ocurre).

- Determinar qué variables pesan más para aumentar o disminuir la probabilidad de que a alguien le suceda el evento en cuestión.

Para realizar el análisis nos basamos en las características que presentan los sujetos a los que, efectivamente, les ocurren o no estos sucesos.

Por ejemplo imaginemos que queremos predecir la probabilidad de *"estar desempleado" = 1* o *"no estarlo" = 0*, la regresión logística tomará en cuenta los valores que asumen en una serie de variables (edad, sexo, nivel educativo, posición en el hogar, origen migratorio, etc.) para los sujetos que están efectivamente desocupados (= 1) y los que no lo están (= 0).

En base a ello, el modelo predecirá para cada uno de los sujetos – independientemente de su estado real y actual – una determinada probabilidad de estar desocupado (es decir, de tener valor 1 en la variable dependiente). Es decir, si alguien es un joven, no amo de casa, con baja educación, de sexo masculino y origen emigrante (aunque esté ocupado) el modelo le predecirá una alta probabilidad de estar desocupado (puesto que la tasa de desempleo de el grupo así definido es alta), generando una variable con esas probabilidades estimadas. Y procederá a clasificarlo como desocupado en una nueva variable, que será el resultado de la predicción.

Además, analizará cuál es el peso de cada una de las variables independientes en el aumento o la disminución de esa probabilidad. Por ejemplo, cuando aumenta la educación disminuirá en algo la probabilidad de estar desocupado. En cambio, cuando el sexo pase de 0 ="mujer" a 1 = "varón", aumentará en algo la probabilidad de desempleo porque la tasa de desempleo de los jóvenes de sexo masculino es mayor que la de las mujeres jóvenes.

Obviamente cuanto más coincidan los estados pronosticados con los estados reales de los sujetos, mejor será el ajuste del modelo.

2. Regresión logística binaria

Los modelos de regresión logística binaria resultan ser los de mayor interés ya que la mayor parte de las circunstancias analizadas en las ciencias experimentales responden a este modelo (presencia o no de una cualidad, éxito o fracaso, etc.).

Como se ha visto, la variable dependiente será una variable dicotómica que se codificará como 0 ó 1 ("ausencia" y "presencia", respectivamente).

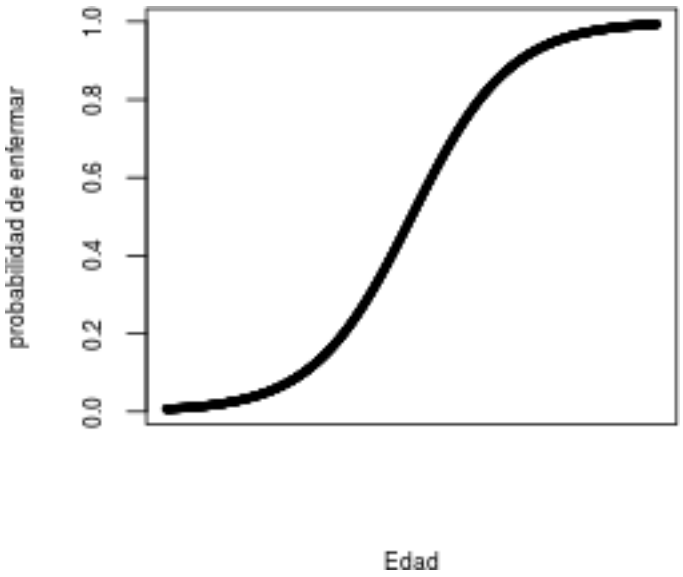
La ecuación de partida en los modelos de regresión logística es

$$P(Y = 1 \mid X) = \frac{exp(b_0 + \sum_{i=1}^{n} b_i x_i)}{1 + exp(b_0 + \sum_{i=1}^{n} b_i x_i)},$$

donde

- $P(Y = 1 \mid X)$ es la probabilidad de que Y tome el valor 1 (presencia de la característica estudiada)
- X es un conjunto de n covariables $x_1, \cdots, x_n$ que forman parte del modelo
- b_0 es la constante del modelo o término independiente
- b_i los coeficientes de las covariables.

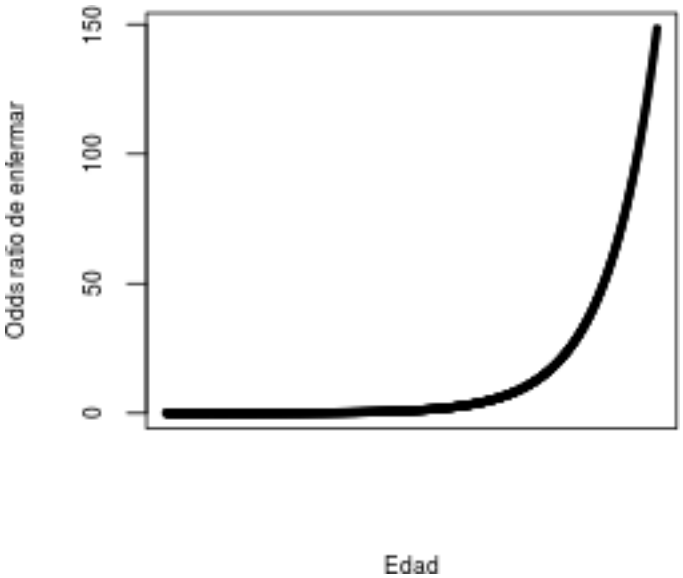
En la siguiente imagen vemos un ejemplo de esta distribución. Está representada la probabilidad de padecer una enfermedad coronaria en función de la edad. Como puede verse, la relación entre la variable dependiente (cualitativa dicotómica), y la covariable (edad, continua en este caso), no queda definida por una recta (lo que correspondería un modelo lineal), sino que describe una forma sigmoidea (distribución logística).



Si se divide la expresión por su complementario, es decir, si se construye su odds (la probabilidad de estar enfermo entre la probabilidad de estar sano), se obtiene una expresión de manejo matemático más fácil

$$\frac{P(Y=1 \mid X)}{1-P(Y=1 \mid X)} = \exp\left(b_0 + \sum_{i=1}^{n} b_i x_i\right)$$

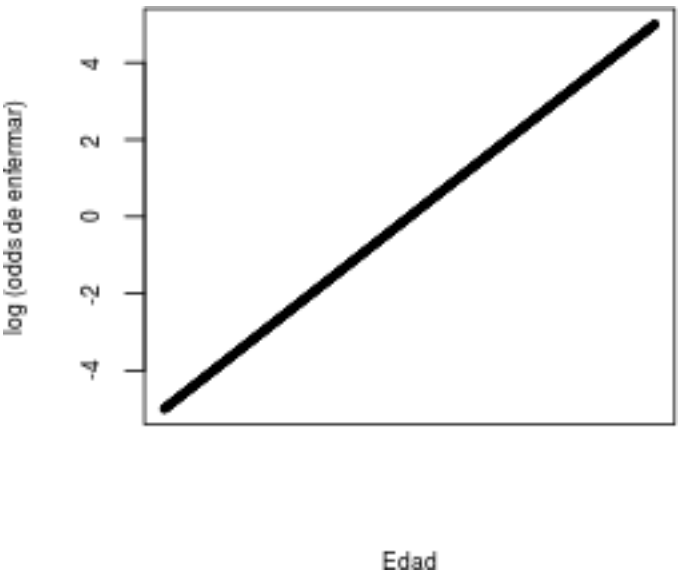
Pero esta expresión aún es difícil de interpretar. Su representación gráfica es



Si ahora realizamos su transformación con el logaritmo natural, se obtiene una ecuación lineal

$$\log\left(\frac{P(Y=1 \mid X)}{1-P(Y=1 \mid X)}\right) = b_0 + \sum_{i=1}^{n} b_i x_i$$

Se trata del llamado **logit**, esto es, el logaritmo natural del *odds* de la variable dependiente (p. ej, el logaritmo de la razón de proporciones de enfermar, de fallecer, de éxito, etc.). El término a la derecha de la igualdad es la expresión de una recta, idéntica a la del modelo general de regresión lineal.



No obstante la regresión logística presenta una **diferencia fundamental** respecto al modelo de regresión lineal.

En el modelo de regresión lineal se asume que los errores estándar de cada coeficiente siguen una distribución normal de media 0 y varianza constante (homocedasticidad), sin embargo, en el caso del modelo de regresión logística no pueden realizarse estas suposiciones pues la variable dependiente no es continua pues se trata de una variable dicotómica, sólo puede tomar dos valores, 0 ó 1, pero ningún valor intermedio.

2.1. Ajuste del modelo

El modelo debe ser aquél más reducido que explique los datos (*principio de parsimonia*), y que además sea técnicamente congruente e interpretable. Hay que tener en cuenta que un mayor número de variables en el modelo implicará mayores errores estándar. Deben incluirse todas aquellas variables que se consideren técnicamente importantes para el modelo, no debería dejarse de incluir toda variable que en un análisis univariado previo demostrara una relación "suficiente" con la variable dependiente.

Si estamos trabajando con variables ficticias y se decide incluir (o excluir) una de estas variables, todas sus correspondientes variables ficticias deben ser incluidas (o excluidas) en bloque.

Otro aspecto de interés es la significación que pudiera tener cada variable ficticia. No siempre todas las variables ficticias de una covariable son significativas. En estos casos es recomendable contrastar el modelo completo frente al modelo sin la covariable mediante la prueba de razón de verosimilitud (es decir, se sacarían del modelo en bloque todas las variables ficticias de la covariable de interés).

Una vez se dispone de un modelo inicial debe procederse a su reducción hasta obtener el modelo más reducido que siga explicando los datos. Para ello utilizaremos los métodos de selección paso a paso.

Cuando tengamos un modelo preliminar, se podrían incluir factores de interacción, es decir, estudiar cómo la asociación de dos o más covariables puede influir en la variable dependiente. Se recomienda la inclusión en el modelo inicial de todas las covariables necesarias más las interacciones de las mismas, o por lo menos, las interacciones de primer orden.

Si decidimos incluir un factor de interacción, hay que tener en cuenta que siempre deberán estar incluidas en el modelo las covariables que componen la interacción. Por ejemplo, si la interacción es *"hipertensión-diabetes"*, hay que meter también en el modelo las covariables *hipertensión* y *diabetes*. El modelo quedaría de la siguiente forma

$$b_0 + b_1\ hipertension + b_2\ diabetes + b_3\ hipertension : diabetes + \cdots$$

Por otra parte, y en relación con la inclusión de interacciones, hay que tener en cuenta que la inclusión de las mismas puede generar multicolinealidad, tanto más probable cuanto mayor sea el número de interacciones.

Se puede encontrar más información en González-Revaldería et al. (2007).

2.1.1. Ajuste del modelo en R

Para el desarrollo completo de este ejemplo nos hemos servido esencialmente del libro A. Field et al. (2012).

Comenzamos leyendo el conjunto de datos

```
enfermo <- read.table( "./files/eel.csv", sep = ";", head = TRUE )
head( enfermo )
```

```
##   Curado Tratamiento Duracion
## 1     No          No        7
## 2     No          No        7
## 3     No          No        6
## 4     Si          No        8
## 5     Si          Si        7
## 6     Si          No        6
```

```
str( enfermo )
```

```
## 'data.frame':    113 obs. of  3 variables:
##  $ Curado     : Factor w/ 2 levels "No","Si": 1 1 1 2 2 2 1 2 2 1 ...
##  $ Tratamiento: Factor w/ 2 levels "No","Si": 1 1 1 1 2 1 2 2 1 1 ...
##  $ Duracion   : int  7 7 6 8 7 6 7 7 8 7 ...
```

- Curado: sí o no (dependiente)
- Tratamiento: sí o no (predictora)
- Duración: días que el paciente estaba enfermo antes de comenzar el tratamiento (predictora).

En regresión logística para crear el modelo usamos el comando `glm()`, su forma general es:

```
glm(resultado ~ predictor(es), data = dataFrame, family = nombre de la distribución,
na.action = una acción ),
```

donde `family` es el nombre de la distribución (Gausiana, binomial, poisson, gamma). En el caso de la regresión logística usamos la opción `family=binomial()`.

```
logmodel <- glm( Curado ~ Tratamiento, data = enfermo, family = binomial( ) )
logmodel
```

```
##
## Call:  glm(formula = Curado ~ Tratamiento, family = binomial(), data = enfermo)
##
## Coefficients:
##   (Intercept)  TratamientoSi
##       -0.2877         1.2287
##
## Degrees of Freedom: 112 Total (i.e. Null);  111 Residual
## Null Deviance:       154.1
## Residual Deviance: 144.2     AIC: 148.2
```

Ya hemos creado el modelo de regresión logística. El siguiente paso es estudiar la calidad del modelo, su bondad.

2.2. Bondad de ajuste

La bibliografía utilizada en este apartado ha sido fundamentalmente A. Field et al. (2012).

2.2.1. Criterio de máxima verosimilitud

Recordemos que regresión logística lo que hacemos es predicir la probabilidad ($P(Y)$) de que un evento (Y) ocurra para una persona (i) dada, basado en las observaciones de si el evento ocurre o no para esa persona (denotamos esto como Y_i, el resultado real para la i-ésima persona). Así, para esa i-ésima persona el suceso Y toma los valores 0 (no ocurre) o 1 (ocurre), y el valor predicho, $P(Y)$, variará entre 0 (no hay probabilidad de que el evento ocurra) y 1 (el evento ocurre con seguridad).

Por tanto, al igual que en regresión múltiple, podemos usar estos valores observados y predichos para evaluar el ajuste del modelo. La medida que usamos es **log-likelihood (logaritmo de la razón de verosimilitud)**:

$$log - likelihood = \sum_{i=1}^{N} [Y_i \log P(Y_i) + (1 - Y_i) \log (1 - P(Y_i))].$$

Se basa por tanto en la suma de las probabilidades asociadas con los resultados estimados y los valores reales.

El estadístico *log-likelihood* es análogo a la suma de cuadrados residual en la regresión múltiple en el sentido de que es un indicador cuánta información sin explicar queda en la variable respuesta tras haber ajustado el modelo. Grandes valores del *log-likelihood* indican un pobre ajuste del modelo, cuanto mayor sea este valor, más variabilidad sin explicar queda en el modelo.

2.2.2. Devianza

Otro indicador importante para estudiar el ajuste del modelo logístico es la **devianza** que se define como el doble logaritmo del estadístico de verosimilitud, es decir, $devianza = -2\times$ *log-likelihood* y se representa como `-2LL`.

La devianza tiene una distribución χ^2 y compara los valores de la predicción con los valores observados en dos momentos:

- El modelo sin variables independientes, sólo con la constante (*modelo referencia*).
- El modelo con las variables predictoras introducidas.

Simplemente tomamos la devianza del nuevo modelo y le restamos la devianza del modelo referencia. Esta diferencia se lo conoce como *ratio-likelihood* y tiene una distribución χ^2 con *k-1* grados de libertad, el número de parámetros del nuevo modelo, menos el número de parámetros del modelo referencia que es siempre 1.

$$\chi^2 = 2LL(nuevo) - 2LL(referencial)$$

$$gl = k_{nuevo} - 1$$

Por lo tanto, el valor de la devianza debiera disminuir sensiblemente entre ambas instancias e, idealmente, tender a cero cuando el modelo predice bien.

2.2.3. r y R^2

Cuando hablamos de regresión lineal el coeficiente de correlación, r y el de determinación, R^2, son medidas útiles para saber cómo de bien se ajusta el modelo a los datos. En regresión logística podemos calcular una medida análoga, conocida como **R-statistic**.

Se trata de la correlación parcial entre la variable resultado y cada una de las predictoras, y puede variar entre -1 y 1. Un valor positivo significa que al crecer la variable predictora, lo hace la probabilidad de que el evento ocurra. Un valor negativo implica que si la variable predictora decrece, la probabilidad de que el resultado ocurra disminuye. Si una variable tiene un valor pequeño de R entonces esta contribuye al modelo sólo una pequeña cantidad.

Una medida análoga al R^2 en la regresión logística puede ser

$$R_L^2 = \frac{2LL(nuevo) - 2LL(referencia)}{2LL(referencia)}$$

es la reducción proporcional en valor absoluto de *log-likelihood* y mide cuánto del error del ajuste disminuye al incluir las variables predictoras. Proporciona una medición de la significación real del modelo. Esta puede variar entre 0 (indicando que los predictores son inútiles prediciendo la variable respuesta) y 1 (indicando que el modelo predice perfectamente la respuesta).

2.2.4. Estadístico de Wadl

Como en la regresión lineal, queremos saber no sólo cómo de bueno es el modelo ajustándose a los datos, sino también la contribución individual de cada uno de las variables predictoras. Esta información la proporciona el *estadístico de Wald* (`z-statistic`) que sigue una distribución normal.

Al igual que el *estadístico* t en regresión lineal, el *estadístico* z nos dice si los coeficientes b_s para cada predictora son significativamente diferentes de cero. Si es distinto de cero asumimos que la variable predictora está haciendo una contribución significativa al modelo para predecir la respuesta (Y).

Su valor para un coeficiente concreto viene dado por el cociente entre el valor del coeficiente y su correspondiente error estándar. La obtención de significación indica que dicho coeficiente es diferente de 0 y merece la pena su conservación en el modelo.

Sin embargo, en modelos con errores estándar grandes, el estadístico de Wald puede proporcional falsos resultados. Si el coeficiente (b) del modelo es grande, el error estándar tiende a inflarse y esto incrementa la probabilidad de rechazar un predictor cuando en realidad está haciendo una contribución al modelo.

$$z = \frac{b}{SE_b}$$

2.2.5. Odds, odds-ratio y coeficientes

El *odds* de un suceso es el cociente sus probabilidades de ocurrencia entre sus probabilidades de no ocurrencia, bajo unas determinadas condiciones C.

$$odds_c(evento) = \frac{P(evento)}{1 - P(evento)}$$

La medida más crucial para la interpretación del modelo logístico es el valor del **odds ratio**, que es la exponencial del valor B (e.d. $\exp(B)$) del modelo de regresión, y se define como el indicador del cambio en los `odds` resultante del cambio de una unidad en el predictor.

Cuando la variable predictora es categórica el *odds ratio* de ocurrencia de un suceso es fácil de explicar, imaginemos que queremos predecir la probabilidad de que un sujeto tenga un accidente de tráfico al haber consumido drogas.

Los *odds* de tener un accidente es la probabilidad de tener un accidente dividido por la probabilidad de no tenerlo

$$odds_{Accidente} = \frac{P(accidente)}{P(no\ accidente)}.$$

Para calcular el cambio en el *odds* resultante del cambio de una unidad en la variable predictora, primero debemos calcular los *odds* de tener un accidente habiendo tomado drogas y después los *odds* de sufrir un accidente sin haber ingerido drogas.

Para obtener los *odds* de haber consumido drogas utilizamos las fórmulas de probabilidades

$$P(accidente) = \frac{1}{1 + e^{-(b_0 + b_1 X)}}$$

$$P(no\ accidente) = 1 - P(accidente)$$

En la primera ecuación tenemos tres incógnitas, los coeficientes b_0 y b_1, y el valor de la propia predictora. La predictora X la codificamos como 0 = consumo, 1 = no consumo, y con ella se estimarán los coeficientes.

A continuación calculamos lo mismo pero después de que el predictor hay aumentado en una unidad, es decir, calculamos el *odds* de tener un accidente cuando no se han consumido drogas. Así que el valor de X es ahora 1 (en vez de 0).

Tenemos por tanto los *odds* antes y después de aumentar la variable predictora una unidad. Con todo esto es fácil ahora calcular el **cambio proporcional en los odds** simplemente dividiendo los *odds* después del cambio entre los *odds* antes del cambio, es lo que se conoce como **odds-ratio**

$$oddsRatio = \frac{odds\ tras\ cambio\ en\ una\ unidad\ de\ X}{odds\ originales}$$

En el caso de que tuviéramos un modelo logístico con más de una variable, se llama **odds ratio** del predictor X_j al cociente del *odds* de ocurrencia al aumentar X_j en una unidad con respecto a no aumentarla, cuando se mantienen constantes el resto de las predictoras $X_1, \cdots, X_k$,

$$oddsRatio_(X_j) = \frac{odds\ tras\ cambio\ en\ una\ unidad\ de\ X_j}{odds\ originales}.$$

La interpretación del **odds-ratio** es que valores mayores que `1` indican que si el predictor aumenta los *odds* de la variable dependiente crecen. Inversamente, un valor menor que `1` indica que tal como el predictor aumente el *odds* del resultado decrece.

El *odds ratio* es una buena medida del tamaño del efecto. Más información en A. Field et al. (2012).

2.2.6. Bondad de ajuste en R

La referencia bibliográfica básica para el desarrollo de este apartado ha sido A. Field et al. (2012).

2.2.6.1. Devianza y χ^2

```
summary( logmodel )
```

```
##
## Call:
## glm(formula = Curado ~ Tratamiento, family = binomial(), data = enfermo)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5940  -1.0579   0.8118   0.8118   1.3018
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.2877     0.2700  -1.065  0.28671
## TratamientoSi   1.2287     0.3998   3.074  0.00212 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 154.08  on 112  degrees of freedom
## Residual deviance: 144.16  on 111  degrees of freedom
## AIC: 148.16
##
## Number of Fisher Scoring iterations: 4
```

- La **bondad de ajuste** global del modelo se evalúa mediante **la devianza** (-2 veces el logaritmo de la verosimilitud) y tenemos que valores grandes indican que los modelos estadísticos son pobres.

En el resumen del modelo `R` calcula la devianza nula (solo con la constante) y la devianza residual (todo el modelo). Para que el modelo sea bueno la devianza residual debe ser menor que la devianza nula ya que valores más bajo de -2LL indican que el modelo predice la variable respuesta con mayor precisión.

En este caso la devianza del modelo nulo es $-2LL = 154{,}08$, pero cuando añadimos `Tratamiento` este valor se reduce a 144,16, lo que nos dice que con esta variable el modelo mejora prediciendo si alguien está curado.

- Para saber **la eficacia** del modelo prediciendo la variable respuesta utilizamos el **estadístico chi-cuadrado**, que mide la diferencia entre el modelo en su estado actual y el modelo cuando sólo se incluyó la constante.

En este caso el valor del estadístico chi-cuadrado es igual al -2LL del modelo con Tratamiento menos el valor de `-2LL` cuando sólo tenemos la constante (154,08 a 144,16 = 9,92). Este valor lo podemos calcular automáticamente con `R` mediante la siguiente serie de comandos

```
dev <- logmodel$deviance
nullDev <- logmodel$null.deviance
modelChi <- nullDev - dev
modelChi
```

```
## [1] 9.926201
```

Para calcular la probabilidad asociada al estadístico chi-cuadrado utilizamos la función `pchisq(Chi, gl)`, cuyos argumentos son el estadístico chi-cuadrado y sus grados de libertad. La probabilidad que queremos es 1 menos el valor de la función `pchisq( )`.

```
chigl <- logmodel$df.null - logmodel$df.residual
chisq.prob <- 1 - pchisq( modelChi, chigl )
chisq.prob
```

```
## [1] 0.001629425
```

como la probabilidad es menor que 0,05, podemos rechazar la hipótesis nula de que el modelo es mejor

prediciendo la variable resultado que si elegimos por azar. Por tanto, podemos decir que, en general, el modelo tiene una aportación significativa en la perdición sobre la cura de un paciente que ha seguido un determinado tratamiento.

2.2.6.2. Coeficientes y z-statistic

Los valores `b`, los coeficientes, tienen la misma función que en la regresión lineal: son los valores que tenemos que sustituir en la ecuación del modelo para establecer la probabilidad de que un caso esté comprendido en una determinada categoría.

Este coeficiente en la regresión logística se puede interpretar como el cambio en el *logit* de la variable de resultado asociado al cambio de una unidad en la variable predictora, donde el logit es simplemente el logaritmo natural de las probabilidades de Y que ocurra.

```
summary(logmodel)$coefficients
```

Un estadístico crucial es el **estadístico de Wald (z-statistic)** que tiene una distribución normal y nos dice si el coeficiente b para ese predictor es significativamente diferente de cero para poder así suponer que la variable predictora está haciendo una contribución significativa a la predicción del resultado (Y).

Así, en este caso podemos afirmar que incluir la variable `Tratamiento` produjo una mejoría significativa en el ajuste del modelo, $\chi^2(1) = 9{,}93$, $p = 0{,}002$.

2.2.6.3. R^2

Podemos calcular un valor similar a la R^2 en los modelo lineales mediante la medida de *Hosmer & Lemeshow* (R^2_L).

```
R2.hl <- modelChi/logmodel$null.deviance
R2.hl
```

```
## [1] 0.06442071
```

2.2.6.4. Odds ratio

Para calcular el cambio en los *odds* resultantes del cambio de una unidad en la variable predictora, primero debemos calcular los *odds* para un paciente curado que no haya sido tratado y después los de un paciente que sí hay recibido tratamiento.

Para calcularlos utilizamos la ecuación

$$P(Y) = \frac{1}{1 + e^{-(b_0 + b_1 X_1)}}.$$

Los coeficientes b_0 y b_1 los obtenemos en la tabla resumen `summary(logmodel)`, vamos a calcular con ellos los *odds ratio*

```
# P(Y) = -0.288+1.229 x

# Odds paciente sin intervención X=0
b0 <- -0.288 + 1.229 * 0
cura0 <- 1 / ( 1 + exp( -b0 ) )
noCura0 <- 1 - cura0
```

```
odds0 <- cura0 / noCura0
odds0
```

```
## [1] 0.7497616
```

```
# Odds paciente con intervención X=1
b1 <- -0.288 + 1.229 * 1
cura1 <- 1 / ( 1 + exp( -b1 ) )
noCura1 <- 1 - cura1
odds1 <- cura1 / noCura1
odds1
```

```
## [1] 2.562543
```

```
# Odds-ratio
oddsRatio <- odds1 / odds0
oddsRatio
```

```
## [1] 3.41781
```

Podemos también calcular directamente los **odds ratio** haciendo la exponencial de los b-valores.

```
exp( logmodel$coefficients )
```

```
##   (Intercept) TratamientoSi
##      0.750000      3.416667
```

Vemos que se obtienen los mismos resultados, y se interpretan en relación al cambio en los odds (probabilidades).

Si el valor es mayor que 1, entonces indica que a medida que aumenta el predictor, las probabilidades de los resultados aumentan. A la inversa, un valor menor que 1 indica que a medida que aumenta el predictor, las probabilidades de los resultados disminuyen. En este ejemplo, podemos decir que las probabilidades de un paciente se cure tras haber sido tratado son 3,42 veces superiores a las de un paciente que no tratado.

Podemos además calcular los **intervalos de confianza** de los coeficientes

```
exp( confint( logmodel ) )
```

```
## Waiting for profiling to be done...
```

```
##                    2.5 %   97.5 %
## (Intercept)   0.4374531 1.268674
## TratamientoSi 1.5820127 7.625545
```

Lo importante de este intervalo de confianza es que no cruza 1 (los valores en cada extremo del intervalo son mayores que 1).

Esto es fundamental porque valores mayores que 1 significan que a medida que la variable predictora aumenta, también lo hacen las probabilidades de (en este caso) curarse. Los valores inferiores a 1 significan lo contrario: si la variable predictora aumenta, las probabilidades de curarse disminuye. El hecho de que tanto los límites inferiores y superiores de nuestro intervalo de confianza estén por encima de 1 nos da la confianza de que la dirección de la relación que hemos observado es cierta en la población.

2.3. Valores ajustados y residuos

Para el desarrollo de este apartado hemos empleado esencialmente el libro de texto A. Field et al. (2012).

Los valores ajustados en la regresión logística son un poco diferentes a los de la regresión lineal, pues predicen las probabilidades de Y dados los valores de cada predictor para cada participante.

También podemos predecir un grupo de pertenencia, basado en el resultado más probable para cada persona en el modelo.

Además, como en la regresión lineal, podemos examinar **los residuos** para asegurarnos de que el modelo se ajusta bien a los datos observados.

El principal propósito de examinar los residuos es

1) Aislar los puntos en los que el modelo se ajusta mal.
2) Aislar los puntos que ejercen una influencia excesiva sobre el modelo.

Para buscar los casos conflictivos tenemos que fijarnos es lo siguiente:

- Mirar los residuos estandarizados y asegurarnos de que no más de `5%` de los casos tiene un valor absoluto mayor de `2`, y que no más de un `1%` tiene valores absolutos más allá de `2.5`. Cualquier caso con valor superior a `3` podría ser un `valor atípico`.
- Calcular la media del estadístico `leverage` (número de predictores más 1, dividido por el tamaño muestral) y buscar valores mayores que dos o tres veces esa media.
- Buscar valores absolutos de `DFBeta` mayores que 1.

Aunque aislemos numerosos valores atípicos o casos influyentes, no tenemos justificación para afirmar que al eliminarlos el modelo ajuste mejor. En lugar de ello, debemos inspeccionar estos casos atentamente y tratar de encontrar una buena razón de por qué son inusuales. Podría tratarse simplemente de un error al meter los datos, o podría ser que tuviera una buena razón para ser inusual, en tal caso sí podríamos excluir ese valor del modelo.

2.3.1. Obtención de residuos, valores predichos y estadísticos necesarios___

```
enfermo$probabilidades.predichas <- fitted( logmodel )
enfermo$studentized.residuals <- rstudent( logmodel )
enfermo$dfbeta <- dfbeta( logmodel )
enfermo$dffit <- dffits( logmodel )
enfermo$leverage <- hatvalues( logmodel )
```

2.3.2. Probabilidades predichas

```
head( enfermo[ , c( "Curado", "Tratamiento", "Duracion", "probabilidades.predichas" ) ] )
```

```
##   Curado Tratamiento Duracion probabilidades.predichas
## 1     No          No        7                0.4285714
## 2     No          No        7                0.4285714
## 3     No          No        6                0.4285714
## 4     Si          No        8                0.4285714
## 5     Si          Si        7                0.7192982
## 6     Si          No        6                0.4285714
```

Estos valores nos dicen que cuando un paciente no recibe tratamiento (Tratamiento = 0, no), hay una probabilidad de 0,429 de que se cure - básicamente, alrededor del 43 % de las personas se recuperan sin tratamiento alguno. Sin embargo, si al paciente se le aplica el Tratamiento (Tratamiento = 1, sí), hay una probabilidad de 0,719 de que mejore - alrededor del 72 % de las personas tratadas mejoran.

Si consideramos que una probabilidad de 0 indica que no hay oportunidad de mejorar, y una probabilidad de 1 que el paciente conseguirá definitivamente ponerse bien, los valores obtenidos proporcionan una fuerte evidencia de que tener un Tratamiento aumenta sus posibilidades de obtener una mejora.

Asumiendo que el modelo es riguroso y que Tratamiento tiene cierta importancia, entonces podríamos concluir que nuestro Tratamiento es el mejor predictor para ponerse mejor (curarse).

2.3.3. Valores influyentes y posibles atípicos

```
head( enfermo[ ,c( "leverage", "studentized.residuals", "dfbeta" ) ] )
```

```
##     leverage studentized.residuals dfbeta.(Intercept) dfbeta.TratamientoSi
## 1 0.01785714            -1.0643627      -3.886912e-02         3.886912e-02
## 2 0.01785714            -1.0643627      -3.886912e-02         3.886912e-02
## 3 0.01785714            -1.0643627      -3.886912e-02         3.886912e-02
## 4 0.01785714             1.3110447       4.782751e-02        -4.782751e-02
## 5 0.01754386             0.8160435      -3.039582e-17         3.225994e-02
## 6 0.01785714             1.3110447       4.782751e-02        -4.782751e-02
```

Los estadísticos residuales básicos para este ejemplo son bastante buenos: todos los casos tienen un `DfBeta` menor que 1, y los *leverage* están muy cerca del valor esperado de leverage `(k+1)/n`=0.0176991.

En definitiva, esto significa que no hay casos influyentes que tengan efecto sobre el modelo. Además, todos los residuos estandarizados tienen valores menores de $\pm$ 2, y así que parece que no tenemos por qué preocuparnos.

2.4. Resumen: etapas de la regresión logística

- Recodificar las variables independientes categóricas u ordinales en variables ficticias y la variable dependiente en 0 y 1.
- Evaluar el ajuste general del modelo final observando la *devianza* y su estadístico χ^2 asociado. Si la significación de la χ^2 es menor que 0,05, entonces el modelo se ajusta significativamente a los datos.
- Para cada variable en el modelo, mirar el valor del *estadístico de Wald* y su significación (debe ser menor que 0,05).
- Analizar la fuerza, sentido y significación de los coeficientes, sus exponenciales y estadísticos de prueba.
- Usar el *odds ratio* para interpretar el modelo. Los podemos obtener mediante el comando `exp(model$coefficients)`. Si el valor es mayor que 1 al aumentar la variable predictora el *odds* de la respuesta aumenta. Inversamente, un valor menor que 1 indica que si la variable predictora crece, el *odds* de la respuesta decrece. ¡OJO! Para que está interpretación sea cierta el intervalo de confianza del *odds ratio* no debe cruzar el 1.
- Realizar un diagnóstico del modelo para comprobar que es un modelo adecuado para predecir la variable respuesta, que se ajusta bien a los datos.

Información más detallada en A. Field et al. (2012).

2.5. Comparación y selección del modelo

La referencia principal aplicada en este apartado es A. Field et al. (2012).

2.5.1. Criterio de información

Como en la regresión lineal podemos usar el criterio de información de Aike (AIC) y el criterio de información de Bayes (BIC) para juzgar el ajuste del modelo. Estos criterios proporcionan una medida del ajuste de un modelo que penaliza al modelo que contiene más variables predictoras, pudiendo así comparar dos modelos.

- $AIC = -2LL + 2k$
- $BIC = -2LL + 2k \times \log(n)$, donde n es el número de casos del modelo.

2.5.2. Métodos paso a paso

Si usamos el método `hacia delante` el ordenador empieza con un modelo que incluye solo la constante y entonces añade predictores individuales al modelo basándose en la variable que mejore el `AIC` o `BIC`. El ordenador continua mientras ninguno de los predictores restantes haya disminuido el criterio.

El método `hacia atrás` usa el mismo criterio pero empieza el modelo incluyendo todas las variables predictoras. `R` comprueba si alguno de esos predictores puede ser retirado del modelo sin incrementar el criterio de información. Si se puede, esa variable se saca del modelo, y se analizan de nuevo el resto de variables.

Mejor que estos métodos es el de `ambas direccioes` que empieza como el método `hacia delante`, con solo la constante, pero cada vez que añadimos una variable, el ordenador comprueba si merece la pena eliminarla.

2.5.3. Selección del mejor modelo en R

Se tomaron datos al azar de 75 jugadores de baloncesto antes de realizar un triple en una competición para evaluar su ansiedad antes del lanzamiento. Los datos están en el archivo `triples.csv` y contiene cuatro variables:

- Marcado: está codificada como 0 = triple fallado y 1 = triple anotado
- PSWQ: grado de preocupaciones del jugador en su vida cotidiana
- Trayectoria: porcentaje de triples anotados por un jugador en particular en su carrera
- Ansiedad: estado de ansiedad antes de lanzar el triple

```
triples <- read.table( "./files/triples.csv", sep = ";", head = TRUE )
head( triples )
```

```
##   PSWQ Ansiedad Trayectoria Marcado
## 1   18       21          56      Si
## 2   17       32          35      Si
## 3   16       34          35      Si
## 4   14       40          15      Si
## 5    5       24          47      Si
## 6    1       15          67      Si
```

```
str( triples )
```

```
## 'data.frame':    75 obs. of  4 variables:
##  $ PSWQ       : int  18 17 16 14 5 1 4 12 11 15 ...
##  $ Ansiedad   : int  21 32 34 40 24 15 10 19 29 14 ...
##  $ Trayectoria: int  56 35 35 15 47 67 75 53 35 65 ...
##  $ Marcado    : Factor w/ 2 levels "No","Si": 2 2 2 2 2 2 2 2 2 2 ...
```

Modelo completo

```
modeloTriple <- glm( Marcado ~ Trayectoria + PSWQ + Ansiedad,
                     data = triples, family = binomial( ) )
summary( modeloTriple )
```

```
##
## Call:
## glm(formula = Marcado ~ Trayectoria + PSWQ + Ansiedad, family = binomial(),
##     data = triples)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.31374  -0.35996   0.08334   0.53860   1.61380
```

```
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -11.49256   11.80175  -0.974  0.33016
## Trayectoria   0.20261    0.12932   1.567  0.11719
## PSWQ         -0.25137    0.08401  -2.992  0.00277 **
## Ansiedad      0.27585    0.25259   1.092  0.27480
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 103.638  on 74  degrees of freedom
## Residual deviance:  47.416  on 71  degrees of freedom
## AIC: 55.416
##
## Number of Fisher Scoring iterations: 6
```

Tan solo obtenemos que la variable `PSWQ` es significativa, por ello vamos a aplicar el método paso a paso para ver si podemos reducir el modelo. Aún así vemos que la devianza del modelo es mucho menor que la devianza solo con la constante, lo cual significa que estas variables ayudan a predecir mejor la respuesta.

Aplicamos el **método paso a paso** hacia atrás

```
step( modeloTriple, direction = "backward" )
```

```
## Start:  AIC=55.42
## Marcado ~ Trayectoria + PSWQ + Ansiedad
##
##               Df Deviance    AIC
## - Ansiedad     1   48.662 54.662
## <none>             47.416 55.416
## - Trayectoria  1   50.074 56.074
## - PSWQ         1   61.141 67.141
##
## Step:  AIC=54.66
## Marcado ~ Trayectoria + PSWQ
##
##               Df Deviance    AIC
## <none>             48.662 54.662
## - Trayectoria  1   60.516 64.516
## - PSWQ         1   61.173 65.173

##
## Call:  glm(formula = Marcado ~ Trayectoria + PSWQ, family = binomial(),
##     data = triples)
##
## Coefficients:
## (Intercept)  Trayectoria         PSWQ
##      1.2803       0.0648      -0.2301
##
## Degrees of Freedom: 74 Total (i.e. Null);  72 Residual
## Null Deviance:       103.6
## Residual Deviance: 48.66     AIC: 54.66
```

Comenzamos con un `AIC=55.42` y la función considera la eliminación de cada una de las variables para finalmente sacar del modelo la variable `ansidad` por ser la que produce un `AIC` más bajo (`54.66`). En el

siguiente paso se considera la eliminación de alguna de las dos restantes variables, pero R decide quedarse con ellas ya que su eliminación supone un aumento, en el mejor de los casos, del AIC a 64.516.

Nos quedamos por tanto con el modelo $Marcado = 0{,}0648 \times Trayectoria - 0{,}2301 \times PSWQ$ que nos dice que el éxito al marcar el triple se ve influenciado positivamente por la trayectoria profesional del jugador y negativamente por el nivel de preocupación que tenga el jugador en su vida diaria.

Bondad del modelo

```
modeloTriple2 <- glm( Marcado ~ Trayectoria + PSWQ, family = binomial(),
                      data = triples )
summary( modeloTriple2 )
```

```
##
## Call:
## glm(formula = Marcado ~ Trayectoria + PSWQ, family = binomial(),
##     data = triples)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2212  -0.3306   0.1038   0.5046   1.6067
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.28031    1.67017   0.767  0.44333
## Trayectoria  0.06480    0.02209   2.934  0.00335 **
## PSWQ        -0.23009    0.07983  -2.882  0.00395 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 103.638  on 74  degrees of freedom
## Residual deviance:  48.662  on 72  degrees of freedom
## AIC: 54.662
##
## Number of Fisher Scoring iterations: 6
```

En este caso tanto la variable Trayectoria como la variable PSWQ tienen un estadístico z significativo, tienen una aportación significativa al modelo.

```
dev <- modeloTriple2$deviance
nullDev <- modeloTriple2$null.deviance
modelChi <- nullDev - dev
modelChi
```

```
## [1] 54.97669
```

Como el valor es positivo quiere decir que la devianza del modelo es menor que la devianza nula por lo que las variables del modelo mejoran la predicción de la variable respuesta. Sin embargo su valor es elevado lo que implique que en el modelo quedan residuos sin explicar. Idealmente queríamos que este valor fuese lo más cercano posible a cero. Calculamos ahora su significación.

```
chidf <- modeloTriple2$df.null - modeloTriple2$df.residual
chisq.prob <- 1 - pchisq( modelChi, chidf )
chisq.prob
```

```
## [1] 1.1533e-12
```

El valor es $< 0{,}05$ lo que quiere decir que el modelo con las variables predictoras es significativamente mejor que aquel solo con la constante. Nuestro modelo en su conjunto es significativamente bueno prediciendo la ansiedad de los juradores.

Podemos decir por tanto que la variable `Trayectoria` y la variable `PSWQ` provocan una mejoría significativa en el ajuste del modelo, $Chi^2(1) = 54{,}97669$, $p < 0{,}001$.

2.6. Supuestos del modelo

La regresión logística comparte algunos supuestos con la regresión usual

- **Linealidad:** en regresión lineal asumimos que la variable respuesta tiene una relación lineal con las variables predictoras. En regresión logística la respuesta es categórica y por ello este supuesto se viola. Por ello por lo que utilizamos el *logit* de los datos. Así, el supuesto de linealidad en regresión logística es que existe una relación lineal entre cada variable predictora continua y el logaritmo de la variable respuesta.
- **Independencia de los errores:** los distintos casos de los datos no deben estar relacionados, por ejemplo, no podemos medir a la misma gente en diferentes puntos del tiempo.
- **Multicolinealidad:** aunque no es un supuesto como tal, la multicolinealidad es un problema como en la regresión lineal. Las variables predictoras no deben estar altamente correlacionadas.

Vamos a utilizar el ejemplo de *triples* para estudiar más detenidamente estos supuestos y ver cómo comprobarlos en `R`.

2.6.1. Linealidad

Para contrastar este supuesto necesitamos ejecutar la regresión logística pero incluyendo como predictores las iteraciones entre cada predictor y el logaritmo de sí mismo. Creamos la iteración de cada término con su logaritmo mediante el comando

```
triples$logPSWQInt <- log(triples$PSWQ)*triples$PSWQ
triples$logAnsInt <- log(triples$Ansiedad)*triples$Ansiedad
triples$logTrayeInt <- log(triples$Trayectoria)*triples$Trayectoria
```

Para realizar el contraste rehacemos el análisis exactamente de la misma manera que anteriormente excepto porque metemos todas las variables de golpe en un solo bloque y añadimos las nuevas iteraciones

```
linealidad <- glm( Marcado ~ PSWQ + Ansiedad + Trayectoria + logPSWQInt + logAnsInt
                   + logTrayeInt, data = triples, family = binomial( ) )
summary( linealidad )
```

```
##
## Call:
## glm(formula = Marcado ~ PSWQ + Ansiedad + Trayectoria + logPSWQInt +
##     logAnsInt + logTrayeInt, family = binomial(), data = triples)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0685  -0.3846   0.1116   0.5460   1.8272
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.87885   14.92410  -0.260    0.795
## PSWQ        -0.42233    1.10267  -0.383    0.702
## Ansiedad    -2.64485    2.79702  -0.946    0.344
```

```
## Trayectoria  1.66601    1.48202   1.124    0.261
## logPSWQInt   0.04393    0.29675   0.148    0.882
## logAnsInt    0.68077    0.65277   1.043    0.297
## logTrayeInt -0.31855    0.31731  -1.004    0.315
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 97.283  on 70  degrees of freedom
## Residual deviance: 45.909  on 64  degrees of freedom
##   (4 observations deleted due to missingness)
## AIC: 59.909
##
## Number of Fisher Scoring iterations: 7
```

Sólo estamos interesados en los términos de las iteraciones. Cualquier iteración que sea significativa quiere decir que el efecto principal ha violado el supuesto de linealidad del logaritmo.

En este caso las tres iteraciones tienen valores de significación (columna `Pr(>|z|)`) mayores de 0.05, indicando que el supuesto de linealidad de se cumple para `PSWQ`, `Ansiedad` y `Trayectoria`.

2.6.2. Multicolinealidad

Se dice que existe multicolinealidad cuando dos o más de las covariables del modelo mantienen una relación lineal. Cuando la colinealidad es perfecta, es decir, cuando una covariable puede determinarse según una ecuación lineal de una o más de las restantes covariables, es imposible estimar un único coeficiente de todas las covariables implicadas. En estos casos debe eliminarse la covariable que actúa como dependiente.

Normalmente lo que se hallará será una multicolinealidad moderada, es decir, una mínima correlación entre covariables. Si esta correlación fuera de mayor importancia, su efecto sería el incremento exagerado de los errores estándar, y en ocasiones, del valor estimado para los coeficientes de regresión, lo que hace las estimaciones poco creíbles. Podemos verificar este supuesto con los estadísticos de VIF y tolerancia, en las matrices de correlación, etc.

Vamos a estudiar este supuesto en nuestro modelo inicial que ya sabemos que una de las variables (Ansiedad) no aporta nada al modelo.

```
library( car )
```

```
## Loading required package: carData
```

```
vif( modeloTriple )
```

```
## Trayectoria        PSWQ    Ansiedad
##   35.227113    1.089767   35.581976
```

```
1/vif( modeloTriple )
```

```
## Trayectoria        PSWQ    Ansiedad
##  0.02838723  0.91762767  0.02810412
```

Estos valores indican que hay un problema de colinealidad: un VIF más de 10 se considera problemático El resultado de este análisis es bastante tajante: existe colinealidad entre la Ansiedad y Trayectoria, y esta dependencia convierte el modelo en sesgado.

Si identificamos multicolinealidad no hay mucho que podamos hacer, la solución no es fácil:

- Podemos intentar eliminar la variable menos necesaria implicada en la colinealidad, a riesgo de obtener un modelo menos válido. Sin embargo, un problema común es no saber qué variable debemos omitir. Cualquiera de las variables problemáticas puede ser omitida, no hay fundamentos estadísticos para suprimir una variable en vez de otra.

- Se recomienda que si eliminamos una variable predictora, ésta se reemplace por otra igualmente importante que no tenga una colinealidad tan fuerte.
- Se puede intentar cambiar la escala de medida de la variable en conflicto (es decir, transformarla). Sin embargo estas transformaciones hacen al modelo muy dependiente de los datos actuales, invalidando su capacidad predictiva.
- También se puede recurrir a aumentar la muestra para así aumentar la información en el modelo y ver si la multicolinealidad puede disminuir, aunque no siempre será posible.
- La última posibilidad, aunque más compleja cuando hay varios predictores, es hacer un análisis factorial y usar las puntuaciones del factor resultante como predictor.

2.6.3. Problemas en la regresión logística

La regresión logística tiene además problemas propios, no son supuestos si no cosas que pueden ir mal.

- Si creamos una tabla con todos los posibles valores de todas las variables entonces idealmente debemos tener datos en cada celda de la tabla. Si no los tenemos debemos estar atentos a posibles errores típicos grandes.
- Si podemos predecir la variable resultado perfectamente a partir de una (o una combinación) de variables predictoras entonces tenemos el problema de *separación completa* y este crea grandes errores también. El problema es que al estar la mayoría de los datos situados en los extremos, los datos intermedios `R` no sabe cómo ajustarlos y opta por tomar la curva lo más vertical posible provocando esos errores.

2.7. Ejemplo completo de reg. logística binaria. Interpretación

Para llevar a cabo este ejemplo hemos utilizado como referencia Group (2014b).

Un investigador está interesado en saber qué efecto tienen variables como *PAU* (nota en Selectividad), *bach* (nota media bachiller) y el *prestigio* del instituto al que asistieron en la admisión de los estudiantes en una determinada universidad. La variable respuesta, admitir/no admitirlo, es una variable binaria.

```
univ <- read.table( "./files/universidad.csv", sep = ";", head = TRUE )
```

```
str( univ )
```

```
## 'data.frame':    400 obs. of  4 variables:
##  $ admitido : int  0 1 1 1 0 1 1 0 1 0 ...
##  $ pau      : int  380 660 800 640 520 760 560 400 540 700 ...
##  $ bach     : num  3.61 3.67 4 3.19 2.93 3 2.98 3.08 3.39 3.92 ...
##  $ prestigio: Factor w/ 4 levels "1","2","3","4": 3 3 1 4 4 2 1 2 3 2 ...
```

Este conjunto de datos tiene una variable respuesta binaria llamada *admitido* y tenemos tres variables predictoras: *pau*, *bach* y *prestigio*. Vamos a tratar a las variables *pau* y *bach* como continuas y la variable *prestigio* una categórica que toma valores de 1 a 4. Las instituciones con prestigio de 1 tienen el mayor reconocimiento, mientras que aquellos con un prestigio de 4 tienen la reputación más baja.

Descriptivos básicos

```
summary( univ )
```

```
##     admitido          pau            bach       prestigio
##  Min.   :0.0000   Min.   :220.0   Min.   :2.260   1: 61
##  1st Qu.:0.0000   1st Qu.:520.0   1st Qu.:3.130   2:151
##  Median :0.0000   Median :580.0   Median :3.395   3:121
##  Mean   :0.3175   Mean   :587.7   Mean   :3.390   4: 67
##  3rd Qu.:1.0000   3rd Qu.:660.0   3rd Qu.:3.670
```

```
##  Max.   :1.0000   Max.   :800.0   Max.   :4.000
```

```
sapply( univ, sd )
```

```
##    admitido         pau        bach   prestigio
##   0.4660867 115.5165364   0.3805668   0.9444602
```

2.7.1. Creación el modelo

Como la variable *prestigio* tiene 4 categorías y en el modelo de regresión logística solo podemos meter variables dicotómicas, tendríamos que crear tres variables ficticias a partir de ella para poder usarla en el modelo logístico. Pero como R es muy listo, si no lo hacemos y metemos la variable *prestigio* directamente en el modelo, él crea por sí solo las variables ficticias.

```
modelUni <- glm( admitido ~ pau + bach + prestigio, data = univ,
                 family = "binomial" )
summary( modelUni )
```

```
##
## Call:
## glm(formula = admitido ~ pau + bach + prestigio, family = "binomial",
##     data = univ)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6268  -0.8662  -0.6388   1.1490   2.0790
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.989979   1.139951  -3.500 0.000465 ***
## pau          0.002264   0.001094   2.070 0.038465 *
## bach         0.804038   0.331819   2.423 0.015388 *
## prestigio2  -0.675443   0.316490  -2.134 0.032829 *
## prestigio3  -1.340204   0.345306  -3.881 0.000104 ***
## prestigio4  -1.551464   0.417832  -3.713 0.000205 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 499.98  on 399  degrees of freedom
## Residual deviance: 458.52  on 394  degrees of freedom
## AIC: 470.52
##
## Number of Fisher Scoring iterations: 4
```

Como vemos en el resumen del modelo R ha creado las variables `prestigio2`, `prestigio3` y `prestigio4`, todas ellas variables ficticias que muestran la diferencia entre cada grupo y el grupo de referencia, que en este caso es `prestigio1`. Así el valor para, por ejemplo, `prestigio3` indica la diferencia del *logit* de admisión de una persona que fue a un instituto con prestigio uno a una que fue a una institución con prestigio tres.

Nos fijamos pues en los coeficientes, sus errores estándar, el estadístico z y los p-valores asociados. Tanto la variable *pau* como *bach* son estadísticamente significativas, al igual que los tres términos para prestigio.

Los coeficientes de la regresión logística nos dan el cambio en el logaritmo de las cuotas de admisión (*logit*) como resultado de un aumento de una unidad en cada una de las variables predictoras.

Para cada cambio en una unidad en *pau*, el logaritmo de las probabilidades de admisión (en comparación con la no admisión) aumentan en 0,002. Para un aumento de una unidad en el *bach*, el logaritmo de las probabilidades de admisión en la escuela aumentan un 0,804.

La variable *prestigio* tiene una interpretación ligeramente diferente. Por ejemplo, habiendo asistido a una universidad con un *prestigio* 2, si la frente a un institución con *prestigio* 1, cambia el logaritmo de las probabilidades de admisión a $-0{,}675$, al estudiar en un instituto de peor nivel (2) disminuye la probabilidad de entrar en esa universidad.

Todos los casos se enfrentan con la variable de referencia, en este caso contra *prestigio* 1. Para saber el cambio entre, por ejemplo, haber ido a un instituto de prestigio 2 frente a uno de prestigio 3 tenemos que restar los valores de los coeficientes. Así, $logit_{2-3} = -0{,}675443 - (-1{,}340204) = 0.664761$, es más probable entrar si venimos de un instituyo con prestigio dos que si venimos de uno con prestigio tres.

Debajo de la tabla de los coeficientes están los índices de ajuste, como residuos, las devianzas y el AIC. Más adelante veremos cómo utilizar estos valores para evaluar el ajuste del modelo.

2.7.2. Bondad de ajuste

```
dev <- modelUni$deviance
nullDev <- modelUni$null.deviance
modelChi <- nullDev - dev
modelChi
```

```
## [1] 41.45903
```

Como el valor es positivo quiere decir que la devianza del modelo es menor que la devianza nula por lo que las variables del modelo mejoran la predicción de la variable respuesta.

```
chidf <- modelUni$df.null - modelUni$df.residual
chisq.prob <- 1 - pchisq( modelChi, chidf )
chisq.prob
```

```
## [1] 7.578194e-08
```

Obtenemos un estadístico χ^2 con valor 41,45 y un p-valor $< 0{,}01$, lo que indica que el efecto general del modelo es estadísticamente significativa.

En particular podemos calcular la aportación particular de la variable *prestigio* en el modelo mediante la función `wald.test(b, Sigma, Terms)`, donde `b` son los coeficientes del modelo, `Sigma` es la matriz del covarianza del modelo y en `Terms` indicamos los términos del modelo que queremos estudiar.

```
library( aod )
wald.test( b = coef( modelUni ), Sigma = vcov( modelUni ), Terms = 4:6 )
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 20.9, df = 3, P(> X2) = 0.00011
```

La prueba χ^2 con valor 20,9 y tres grados de libertad se asocia con un p-valor 0,00011, lo que indica que el efecto general de `prestigio` es estadísticamente significativo.

2.7.3. Odds ratio

Calculamos directamente los odds ratio y sus respectivos coeficientes y los ponemos juntos en una única matriz.

```
exp( cbind( OR = coef( modelUni ), confint( modelUni ) ) )
```

```
## Waiting for profiling to be done...
```

```
##                    OR       2.5 %    97.5 %
## (Intercept) 0.0185001 0.001889165 0.1665354
## pau         1.0022670 1.000137602 1.0044457
## bach        2.2345448 1.173858216 4.3238349
## prestigio2  0.5089310 0.272289674 0.9448343
## prestigio3  0.2617923 0.131641717 0.5115181
## prestigio4  0.2119375 0.090715546 0.4706961
```

Podemos decir que, al aumentar *bach* en una unidad, las probabilidades de ingresar en la universidad (frente a no ser admitido) aumenta un factor de 2,23. Para obtener más información sobre cómo interpretar los odds ratio visitar la página web Group (2014a).

2.7.4. Predicción

Comenzamos calculando la probabilidad predicha de admisión para cada valor de *prestigio*, manteniendo *pau* y *bach* en sus valores medios.

```
univ_pre1 <- with( univ, data.frame( pau = mean( pau ), bach = mean( bach ),
                                     prestigio = factor( 1:4 ) ) )
univ_pre1$prestigioPred <- predict( modelUni, newdata = univ_pre1, type = "response" )

head(univ_pre1)
```

```
##     pau   bach prestigio prestigioPred
## 1 587.7 3.3899         1     0.5166016
## 2 587.7 3.3899         2     0.3522846
## 3 587.7 3.3899         3     0.2186120
## 4 587.7 3.3899         4     0.1846684
```

Vemos que la probabilidad predicha de ser aceptado en un la universidades más prestigiosas (Prestigio = 1) es de 0,52 y de 0,18 para las instituciones de prestigio más bajo (Prestigio = 4), manteniendo *pau* y *Bach* en sus valores medios.

Podemos de manera similar crear una tabla de probabilidades predichas variando el valor de *pau* y *prestigio*. Creamos 100 valores de *pau* entre para cada valor de prestigio (es decir, 1, 2, 3 y 4).

```
univ_pre2 <- with( univ, data.frame( pau = rep( seq( from = 200, to = 800, length= 100 ), 4),
                                     bach = mean( bach ), prestigio = factor( rep( 1:4, each = 100 ) ) ) )

univ_pre2$prestigioPred2 <- predict( modelUni, newdata = univ_pre1, type = "response" )

head( univ_pre2 )
```

```
##        pau   bach prestigio prestigioPred2
## 1 200.0000 3.3899         1      0.5166016
## 2 206.0606 3.3899         1      0.3522846
## 3 212.1212 3.3899         1      0.2186120
## 4 218.1818 3.3899         1      0.1846684
## 5 224.2424 3.3899         1      0.5166016
## 6 230.3030 3.3899         1      0.3522846
```

Puede encontrar una ampliación del ejemplo en Group (2014b).

2.8. Resumen de código en R

```
# Leer los datos de un fichero .csv
df <- read.table("files/45A-file.csv", sep = ";", head = TRUE)
### Primera aproximación a los datos
str(df)
summary(df)

# Modelo de regresión logística
logmodel <- glm( var1 ~ var2, data = df, family = binomial( ) )
logmodel

## resumen del modelo
summary(logmodel)

# Bondad de ajuste del modelo
## Devianza y Chi2
dev <- logmodel$deviance
nullDev <- logmodel$null.deviance
modelChi <- nullDev - dev
modelChi

chigl <- logmodel$df.null - logmodel$df.residual
chisq.prob <- 1 - pchisq( modelChi, chigl )
chisq.prob

# R^2
R2.hl <- modelChi/logmodel$null.deviance
R2.hl

# Odds ratio
exp( logmodel$coefficients )
## intervalos de confianza
exp( confint( logmodel ) )


# Diagnóstico del modelo
df$probabilidades.predichas <- fitted( logmodel )
df$studentized.residuals <- rstudent( logmodel )
df$dfbeta <- dfbeta( logmodel )
df$dffit <- dffits( logmodel )
df$leverage <- hatvalues( logmodel )

head( df[ , c( "Curado", "Tratamiento", "Duracion", "probabilidades.predichas" ) ] )
head( df[ ,c( "leverage", "studentized.residuals", "dfbeta" ) ] )

# Selección del modelo
modelog <- glm(var1 ~ var2 + var3 + var4, data = df, family = binomial( ) )
summary( modelog )
step( modelog, direction = "backward" )
```

```
# Supuestos del modelo
## Linealidad
df$logvar3Int <- log(df$var3)*df$var3
df$logvar4Int <- log(df$var4)*df$var4
df$logvar2Int <- log(df$var2)*df$var2

linealidad <- glm( var1 ~ var2 + var3 + var4 + logvar2Int + logvar3Int
+ logvar4Int, data = df, family = binomial( ) )
summary( linealidad )

## multicolinealidad
library( car )
vif( modelog )
```

Volver al índice del curso

Referencias y bibliografía

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using r* (1st edition.). Sage Publications Ltd.

González-Revaldería, J., Fernández, J. M. P., García, B. P., & Queraltó, J. M. (2007). Curso de estadística para el laboratorio clínico. módulo 3: Regresión logística. Sociedad Española de Bioquímica Clínica y Patología Molecular. Retrieved October 24, 2014, from http://www.seqc.es/es/Varios/7/40/Modulo_3:_Regresion_logistica_y_multiple/

Group, U. S. C. (2014a). FAQ: How do i interpret odds ratios in logistic regression? Institute for Digital Research; Education. Retrieved October 24, 2014, from http://www.ats.ucla.edu/stat/r/dae/logit.htm

Group, U. S. C. (2014b). R data analysis examples: Logit regression. Institute for Digital Research; Education. Retrieved October 24, 2014, from http://www.ats.ucla.edu/stat/r/dae/logit.htm