**Regresión logística**

La regresión logística es un tipo específico de Modelos Lineales Generalizados (GLM). Mientras los modelos lineales se utilizan para calcular la correlación entre una variable dependiente numérica (intervalo-razón) y una o más variables independientes (de cualquier tipo) y predecir nuevos valores; la regresión logística generaliza los conceptos de los modelos lineales regulares y se utilizan cuando la variable dependiente es categórica o numérica pero discreta. Si esta es binaria, la regresión logística es **binomial** o **dicotómica**, si la variable tiene tres o más niveles hablamos de una regresión logística **multinomial** o **politómica**.

**1. Supuestos**

La regresión logística no asume los supuestos de la regresión lineal, particularmente el de normalidad, Linealidad y homoscedasticidad. Puede manejar cualquier tipo de relación no necesariamente lineal, ya que aplica una transformación logarítmica no lineal.

* Las observaciones deben ser independientes entre ellas.
* Las variables independientes pueden ser continuas o discretas (categóricas, ordinales) y se pide poca o nula multicolinealidad[[1]](#footnote-2): las variables predictoras deben ser independientes unas de otras.
* No se exige linealidad entre la variable dependiente y las independientes, sin embargo sí requiere que la relación entre la variable dependiente y los log odds (ver más adelante) sea lineal.

Este modelo requiere muestras más grandes respecto a las que se usan para el modelo de regresión lineal, ya que la estimación por máxima verosimilitud es más débil que la estimación por mínimos cuadrados.

**2. Regresión logística**

La ecuación para una regresión logística va a ser semejante a la de una regresión lineal, pero el resultado ya no será el valor que toma la variable dependiente sino las probabilidades de un cierto valor posible de esa variable.

g(x)=b0 +b1x1+b2x2+(…)

Donde:

-b0: es el intercepto (el valor de mi variable dependiente cuando las independientes valen 0 o están en su nivel de base)

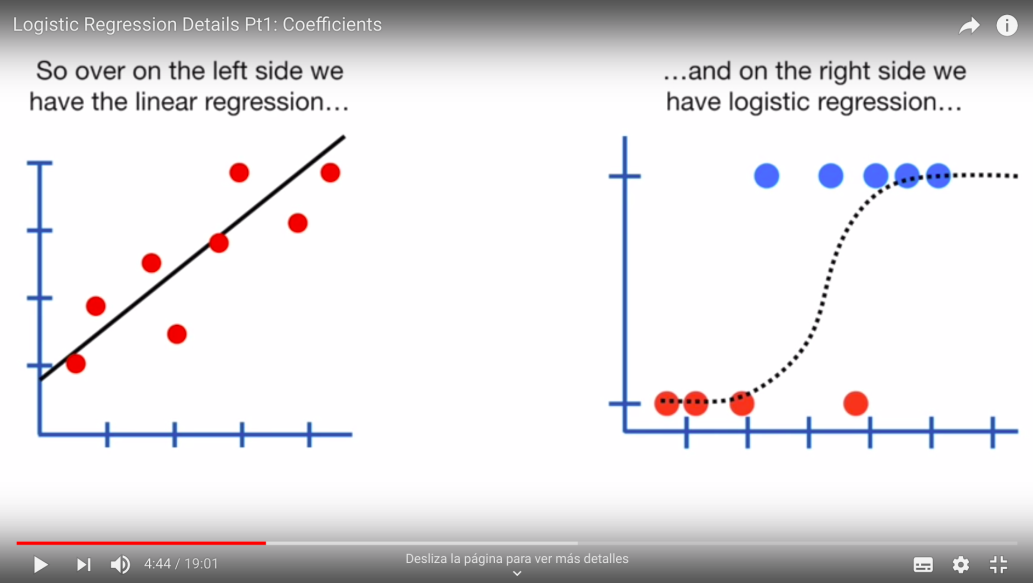
-b: es la pendiente de la recta (cuánto va a modificarse la probabilidad de mí variable dependiente cuando las variables independientes se modifican).

-X: el valor de las variables independientes.

Ahora bien, las probabilidades pueden estar representadas en diferentes escalas.

**3. Escalas**

A diferencia de la regresión lineal, las variables dependientes categoriales o discretas no toman cualquier valor entre -infinito y +infinito, sino entre 0 y 1 ya que representa la posibilidad de que un valor de X pertenezca a la categoría Y1 o Y2 (Y3, Y4…). Esta relación entre las variables se representa, ya no una lìnea recta, sino con una curva en forma de S.



Para poder aplicar las generalizaciones de los modelos lineales es necesario transformar el rango de predicciones (-infinito a +infinito) a un rango más acorde al tipo de variable. Esto se hace transformando las probabilidades de mí variable dependiente (0-1) en “log odds” o “logit” lo que da una escala de -infinito a +infinito. Los coeficientes en las tablas de R van a estar en esta escala.

*3.1. Odds*

Un odds simple es la razón de las chances de un evento (en nuestro caso, la posibilidad de observar uno de los niveles de nuestra variable dependiente) entre las chances de un no-evento (la posibilidad de no observar ese nivel). Es importante entender que las odds **no es lo mismo que probabilidad**: las probabilidades se calculan sobre todos los eventos posibles (la chances de observar el nivel 1 de nuestra variable + las chances de no observarla).

Por ejemplo, si estamos analizando el uso de verbos (verboA-verboB) en un determinado contexto a partir de un corpus y sabemos que el verboA aparece 3 veces en dicho contexto y un verboB aparece 2, las odds del verboA es:

odds(A)= 3/2 = 1.5

mientras que las probabilidades del verboA serían:

P(A)= 3/5 = 0.6

Ahora bien, las odds pueden ser calculadas a partir de las probabilidades:

Odds: P/ 1-P

donde P es la probabilidad del evento

y (1-P) es el cálculo para la probabilidad del no-evento ya que la P(e)+P(-e)=1

Entonces las odds del verboA también pueden calcularse como:

odds(A)= 0.6/ 1-0.6 =0.6/0.4= 1.5

|  |  |
| --- | --- |
| ODDS | Interpretación |
| Igual a 1 | Las chances para los dos eventos son iguales |
| Mayor a 1 | Las chances del evento (numerador) son mayores. |
| Mayor a 0 pero menor a 1 | Las chances del no-evento (denominador) son mayores. |

*3.2. Odds ratio*

Es la razón de dos odds.

OR= Odds1/ Odds2 = (Px/1-Px)/ (Py/1-Py) = (eventox/no-eventox)/ (eventoy/no-eventoy).

Las odds ratio se interpretan de la misma manera que las odds simples.

*3.3. Log odds o Logit*

Ahora bien, como se ve en el cuadro las odds que favorecen a un evento van de 1 a +infinito, pero las odds que favorecen a un no-evento se representan de 0 a 1. Esta asimetría dificulta la comparación de las odds en favor y en contra del evento que espero observar. Para solucionar esta asimetría se hace el logaritmo de las odds. Logit es el nombre que recibe el logaritmo de odds calculado a partir de probabilidades.

Esta escala toma valores entre -infinito a +infinito y está centrada en cero porque este valor es el logaritmo natural de 1 (es decir cuando las odds del evento y del no-evento son iguales).

Logit= log(odds) = log (p/1-p)

Si el número es negativo debemos interpretar que el evento es menos probable que el no-evento y a la inversa.

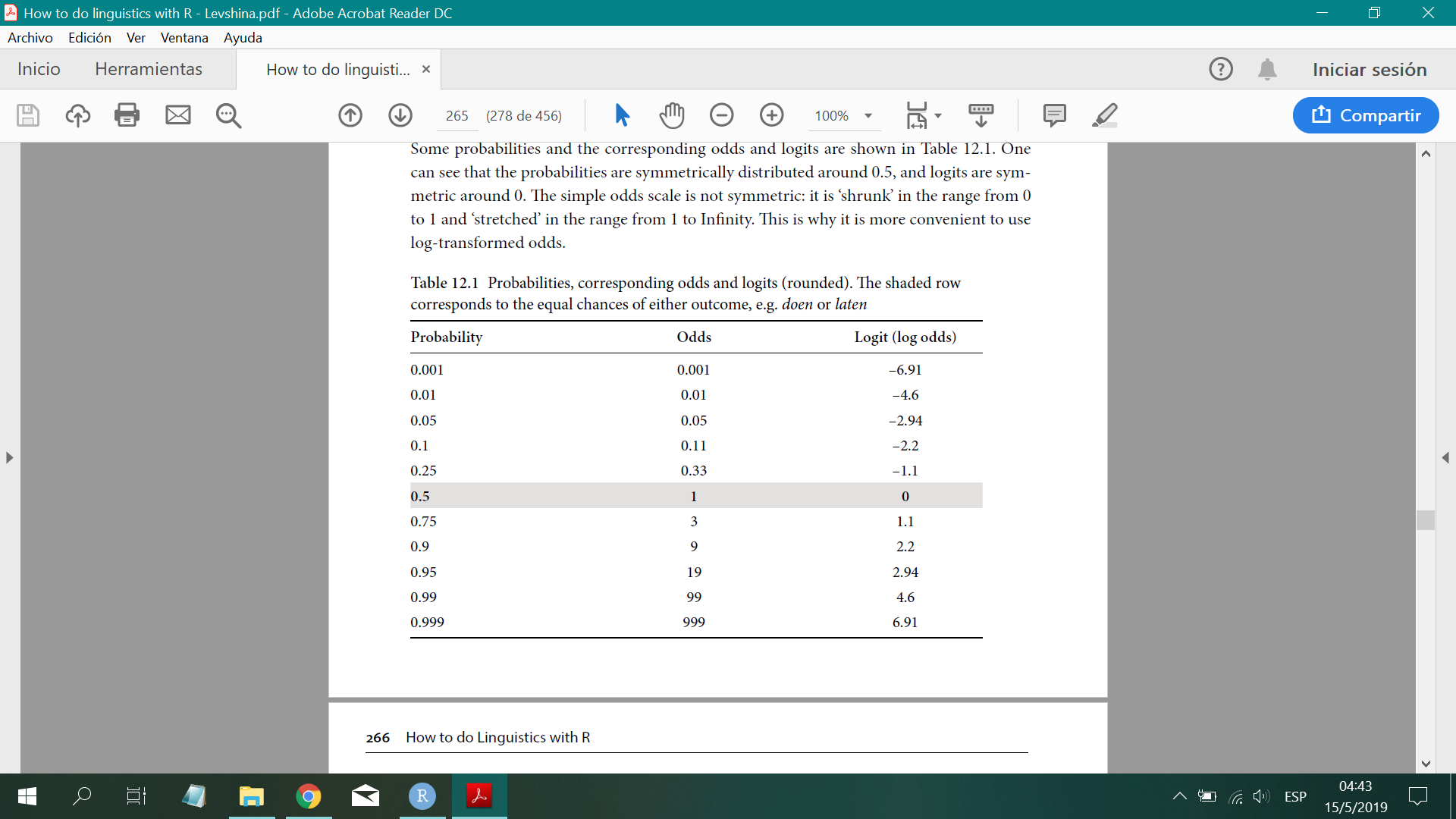
*2.3.2. Log odds ratio*

También se puede calcular el logaritmo de una odds ratio. Si el resultado es cercano a 0 no hay diferencias entre los eventos. Si es positivo Y refuerza las chances del evento en comparación con X. Si el resultado es negativo las chances del evento decrecen con Y.

log [ (Px/1-Px) / (Py/1-Py) ]

*3.4. Relación entre escalas*

Los resultados de una regresión logística entonces pueden ser presentados en escala de probabilidades, odds o logit. A partir de cada escala puede calcularse su equivalente en la otra.



LAVSHINA(p.265)

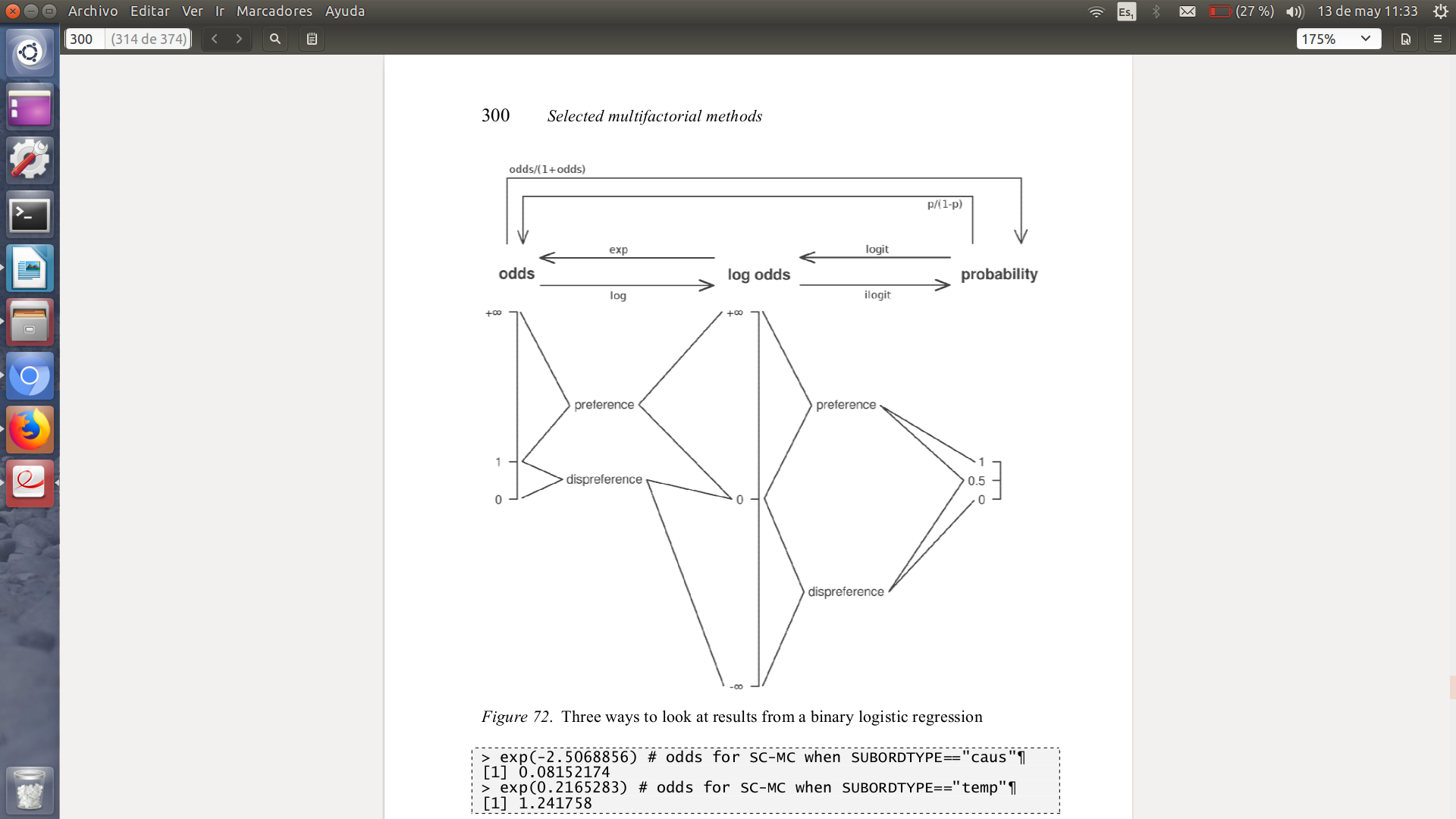


Figura 2. “This graph represents the three perspectives on the results next to each

other and it represents the possible numerical ranges of the three ways on

the y -axes: odds range from 0 to +∞, log odds from -∞ to +∞, and probabil-

ities from 0 to 1. Each of these perspectives expresses preference, dispref-

erence, and lack of effect in different ranges. In numerical odds space, no

preference is 1, in log odds space it’s 0, and for predicted probabilities it’s

of course 0.5 (since we have two options). For odds, preferences are re-

flected by odds greater than 1, by positive log odds, and by predicted prob-

abilies of > 0.5, and the opposites reflect dispreferences.” (GRIES).

Pero los log odds son simétricos y más sencillos de interpretar, por lo que generalmente se utiliza esta escala. Teniendo en cuenta esto podemos re-escribir la ecuación de una regresión logística como:

log(P/1-P) = b0+b1x1+b2x2 (...)

Ajustar un modelo es, entonces, encontrar los valores de b0 y b1, b2(…) para cada valor que puede tomar X.

**4. Regresión logística binaria**

*4.2. Ajustar un modelo*

PAQUETES: “rms”, “visreg”, “car”

Hay dos funciones para crear una regresión logística en el paquete “rms”, cada una te da diferentes estadísticas:

>modelo.lrm= lrm(VD~ VI1+VI2+VI…, data=nombre)

>modelo.lrm

>modelo.glm=glm(VD~VI1+VI”+VI…, data=nombre, family=binomial)

>summary(modelo.glm)

\*Tabla de LRM

-A la izquierda se reporta el número total de observaciones y la frecuencia de cada nivel de la variable dependiente.

-La columna “Model Likelihood Ratio Test” nos dice si el modelo es significante en general (este test es similar al F-test de ANOVA): grados de libertad, p-value.

La hipótesis nula es que la desviación del modelo no es diferente a la desviación de un modelo sin ninguno de los predictores propuestos, de modo que p<0.05 para que el modelo sea significativo.

-La estadìstica más reportada para la regresión logística es el índice de concordancia C: es la proporción de las veces que el modelo predice mayor probabilidad de un valor que otro en los casos donde efectivamente se observa ese valor. Hosmer & Lemeshow (2000) proponen que:

C= 0.5 ---> el modelo no discrimina

0.7<= C <0.8 ---> es una discriminación aceptable

0.8<=C< 0.9 ---> discriminación excelente

C>= 0.9 ---> discriminación excepcional

-Otra estadística muy reportada es el pseudo-R2 de Nagelkerke: toma un valor entre 0 (sin poder predictivo) y 1 (predicción perfecta). Tiende a ser más bajo que el R2 incluso cuando la calidad de los modelos es comparable.

-Coeficientes:

\*Intercepto: es el valor en log(odds) de la variable dependiente cuando todos los predictores están en su nivel referencial o cero. [[2]](#footnote-3) El algoritmo compara el segundo nivel de la variable dependiente con el primero (el valor referencial).

\*Coeficientes de los predictores: están representados en log odds ratios, comparan las odds del valor de la dependiente por cada nivel de un predictor con el nivel referencial de ese predictor. Si el coeficiente es positivo el nivel especificado en la tabla aumenta las chances de valor no referencial de la variable dependiente. Si el coeficiente es negativo, el nivel especificado disminuye las odds del valor no-referencial.

-S.E.: El error estándar. Usualmente errores estándar muy altos son señal de esparcimiento de la data o multicolinealidad.

-Wald es el test que lleva ese nombre: es la razón entre el estimativo y el error estándar, y se utiliza para calcular valores-p

-Valor-P: cuán confiado se puede estar sobre el coeficiente estimado si la hipótesis nula de la “no diferencia entre un valor dado y el valor referencial de los predictores” puede ser rechazada.

En relación a la ecuación:

g(x)=b0 +b1x1+b2x2+(…)

El logit (las probabilidades en log odds de un valor de la variable dependiente) se va a calcular reemplazando las b por los coeficientes de la tabla, y las X por 1 si la variable es un predictor relevante (significativo en p), o por 0 si no lo es.

\*Si un 95% intervalo de confianza incluye un cero el efecto no es significante. Esto va a estar en log(odds)ratios, se puede pasar a simples odds ratios, en cuyo caso el efecto no es significativo si el intervalo incluye 1.

*4.3. Selección de variables*

Existen dos métodos para seleccionar las variables de un modelo

-Ajustar un modelo con todas las variables independientes de interés y ver si son significativas.

-Stepwise (paso a paso): se ajusta un modelo de una variable, y uno de dos y se los compara (forward).O se ajusta un modelo con todas las variables y uno con todas menos una y se los compara (backward)

Se los puede comparar con ANOVA o con la funciòn DROP1 (toma un término por vez y testea el cambio de ajuste del modelo)

*4.5. Interacciones*

A este modelo se le podrían agregar interacciones entre las variables independientes. Estas se pueden testear también comparando con ANOVA dos modelos, uno con interacción y otro sin.

La comparación con ANOVA es útil porque cuando un predictor tiene más de un nivel no todas las comparaciones aparecen en la tabla.

*4.6. Overfitting*

Al igual que en la regresión lineal, se puede validar el modelo con bootstrapping: se utiliza una muestra para analizar y validar los resultados en vez de ajustar un modelo con una muestra y validarlo con otra. Su principio básico es: un nuevo data set es organizado como muestra a partir de los datos originales de forma random y con reemplazo (una misma observación puede ser utilizada más de una vez) y entonces el análisis estadístico es efectuado otra vez. Cuando el procedimiento se repite muchas veces se pueden obtener valores estimativos confiables de la precisión del modelo.

La columna de “optimism” muestra las diferencias entre el entrenamiento y el test. Altos números indican overfitting. Una de las reglas es que la pendiente de optimismo (el optimismo en los estimados de los coeficientes de regresión de las variables independientes) no deberían exceder 0.05.

Si hay evidencia de overfitting se puede intentar agregar más observaciones o utilizar algún tipo de penalización.

**5. Regresión logística multinomial**

Las regresiones logísticas multinomiales son aquellas cuya variable dependiente categorial tiene más de dos niveles.

Existen dos formas de ajustar estos modelos:

-Ajustar dos modelos binarios donde uno de los niveles es la referencia para la comparación (nivel 1 vs nivel dos y nivel 1 vs nivel 3)

-Comparar las odds de cada nivel vs todos los otros niveles (nivel 1 vs nivel2+3 / nivel 2 vs nivel1+3/ nivel 3 vs nivel1+2)

*5.1. Modelos con nivel de referencia*

Para este tipo de modelos utilizaremos la función mlogit, la cual necesita que los datos estén en un formato especial llamado “long” donde cada observación se repite varias veces por cada nivel de la variable dependiente.

mmodelo=mlogit(VD~1| VI1+VI2+(...), data=nombre, reflevel= n)

summary(mmodelo)

#1| crea un modelo multinomial.

# “reflevel” especifica el nivel de referencia con el que se compararán los otros. Por default R selecciona el primero.

En el output encontramos:

* La frecuencia de cada nivel
* Estadísticas de un buen ajuste:
* Log-likelihood: la cantidad de desviación que queda, mientras más chico sea el valor absoluto mejor ajustado estará el modelo
* McFadden R2: análogo a R2 en la regresión lineal. Valores de 0.2-0.4 se consideran buenos (equivalentes a 0.7-0.9 de los modelos lineales)
* Likelihood ratio test: testea la significancia general del modelo (cuando H0 es que ninguno de los coeficientes es diferente de cero). Un valor pequeño de p muestra que el modelo es significante.
* Coeficientes: cada predictor aparece màs de una vez dado que se compara el nivel nombrado con el nivel de referencia de la variable dependiente cuando se modifica la variable independiente nombrada. Estos valores son log odds ratios.  
  Un coeficiente positivo significa que el nivel del predictor aumenta las chances del nivel dependiente nombrado en comparación con las chances del nivel de referencia. Un coeficiente negativo significa que el predictor disminuye las chances de el nivel dependiente nombrado.

*5.2. Modelos de tipo “uno vs el resto”*

En estos modelos cada nivel de la variable dependiente es comparado con todos los otros niveles.

mmodelo2=polytomous(VD~VI1+VI2+ (...), data=nombre)

summary(mmodelo2)$statistics

Summary nos muestra (en simple odd ratio):

-DF

-AIC.model

-BIC.model

-loglikelihood

-deviance

-R2. likelihood (Mcfadden)

-R2 nagelerke

-Crosstable: las filas son observaciones y las columnas predicciones

-Accuracy: Cuán correctamente el modelo predice los niveles de VD (predicciones correctas/ observaciones totales)

-recall.predicted: La proporción de instancias predichas por cada nivel de VD. “De todas las instancias reales de X en qué porcentaje se predijo X”.

-precision.predicted: cuàntas veces se predijo correctamente (en %). “De todas las X qué se predijo cuántas fueron correctas”.

1. Existen varias reglas (VIF deben ser menores a 5) (VIF no deberían ser mayores a 10) [↑](#footnote-ref-2)
2. R utiliza como valor referencial de los predictores, por default (teatment coding) el primer nivel en orden alfabético. [↑](#footnote-ref-3)