Análisis Avanzado de Datos

Nicolás López

Motivación

Motivación

- Método comúnmente utilizado tanto en estadística clásica como en machine learning (ML), ¿por qué? ¿cuál es la intersección entre los dos mundos?
- Hace parte de una generalización del modelo RLS/RLM, denominados GLM (modelos lineales generalizados). ¿cómo se generaliza el concepto lineal?

Probabilidad y odds

Probabilidad y odds

Suponga que quiere conocer el fenómeno de reprobar o no la clase, para esto obtuvo una colección de 500 datos:

```
##
## A.Reprobar  B.Aprobar
##        200        300
```

La probabilidad de reprobar la materia es $\pi = 200/500$, mientras que el *odds* de reprobar está dado por $200/300$, y así, el *odds* de reprobar es de 2 a 3. Note que el *odds* no es una probabilidad, es una **razón**. Esta puede calcularse también como una razón entre probabilidades:

$$odds = \frac{200}{300} = \frac{200/500}{300/500} = \frac{\pi}{1-\pi}$$

Note que

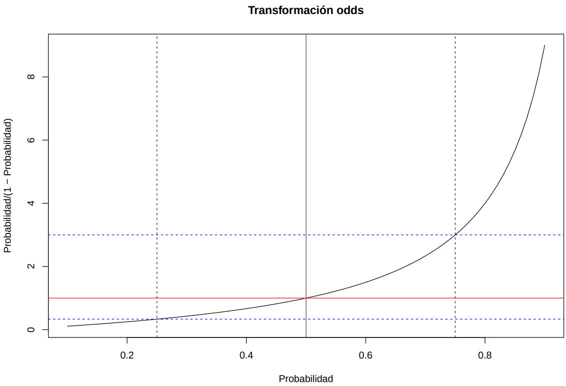
1. A medida que menos personas reprueban, *odds* disminuye (hacia 0).
2. A medida que más personas reprueban, *odds* incrementa (hacia ∞).
3. Si una mitad reprueba y la otra no, *odds* es 1.

Así

1. Cuando $\pi < 0.5$ hay menor probabilidad de reprobar, así, hay menor probabilidad de reprobar si el *odds* está en $(0, 1)$.
2. Cuando $\pi > 0.5$ hay mayor probabilidad de reprobar, así, hay mayor probabilidad de reprobar si el *odds* está en $(1, \infty)$.
3. Cuando $\pi = 0.5$ hay la misma probabilidad de reprobar y aprobar y el *odds* es igual a 1.

Gráficamente tenemos una **transformación monótona** de la probabilidad mediante el *odds*:

```
Probabilidad = seq(0.1,0.9,by=0.01)
plot(Probabilidad,Probabilidad/(1-Probabilidad),main="Transformación odds",type="l")
abline(v=0.5,lty=1,col="red")   ; abline(h=0.5/0.5,lty=1,col="red")
abline(v=0.25,lty=2,col="blue") ; abline(h=0.25/0.75,lty=2,col="blue")
abline(v=0.75,lty=2,col="blue") ; abline(h=0.75/0.25,lty=2,col="blue")
```

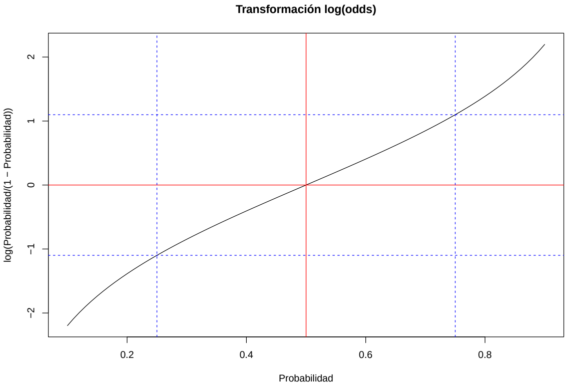


Sin embargo, esta transformación es **asimétrica respecto a la probabilidad**, ya que $\pi > 0.5 \longrightarrow$ odds $\in (1, \infty)$ pero $\pi < 0.5 \longrightarrow$ odds $\in (0, 1)$. Por ejemplo:

- Si $\pi = 0.25$, $\pi/(1-\pi) = 0.33$, a -0.66 del odds de 0.5 (1).
- Si $\pi = 0.75$, $\pi/(1-\pi) = 3.00$, a $+2.00$ del odds de 0.5 (1).

Al calcular el logaritmo del *odds* logramos una transformación monótona y simétrica de la probabilidad:

```
Probabilidad = seq(0.1,0.9,by=0.01)
plot(Probabilidad,log(Probabilidad/(1-Probabilidad)),main="Transformación log(odds)",type="l")
abline(v=0.5,lty=1,col="red")   ; abline(h=log(0.5/0.5),lty=1,col="red")
abline(v=0.25,lty=2,col="blue") ; abline(h=log(0.25/0.75),lty=2,col="blue")
abline(v=0.75,lty=2,col="blue") ; abline(h=log(0.75/0.25),lty=2,col="blue")
```



Y así,

- Si $\pi = 0.25$, $\log(\pi/(1-\pi)) = -1.09$, a -1.09 del log-odds de 0.5 (0).
- Si $\pi = 0.75$, $\log(\pi/(1-\pi)) = +1.09$, a $+1.09$ del log-odds de 0.5 (0).

Ahora, suponga que quiere conocer como se relaciona el reprobar o no la clase con la asistencia a todas las clases del semestre (variable predictora). Con la misma colección de 500 datos obtuvo la siguiente tabla de contingencia:

```
##
##              A.Reprobar B.Aprobar
##   A.Asiste          10       280
##   B.NAsiste        190        20
```

En este caso tenemos el *odds ratio* que permite determinar si existe una relación entre asistir a clase y aprobar.

El *odds ratio* se calcula como la razón de *odds* de cada subpoblación:

- Dado que una persona asiste a clase, su *odds* de reprobar es $10/280$ (y log odds de -3.33).
- Dado que una persona no asiste a clase, su *odds* de reprobar es $190/20$ (y log odds de $+2.25$).

Con lo cual

$$
\begin{aligned}
\text{odds ratio} &= \frac{\text{odds reprobar} \mid \text{Asiste}}{\text{odds reprobar}\mid \text{NAsiste}} \\
&= \frac{10/280}{190/20} \\
&= \frac{\frac{10}{290}\Big/\frac{280}{290}}{\frac{190}{210}\Big/\frac{20}{210}} \\
&= \frac{P(\text{Reprobar}|\text{Asiste})/P(\text{Aprobar}|\text{Asiste})}{P(\text{Reprobar}|\text{NAsiste})/P(\text{Aprobar}|\text{NAsiste})} = 0.003
\end{aligned}
$$

Los *odds* de reprobar la materia es 0.003 veces menor para estudiantes que asisten a clase.

Note que

- Si odds ratio $\in (0, 1)$ - Entre menor sea el *odds ratio*, menor ''riesgo" de reprobar - factor de protección.
- Si odds ratio $\in (1, \infty)$ - Entre mayor sea el *odds ratio*, mayor ''riesgo" de reprobar - factor de riesgo.
- A medida que el *odds ratio* se acerca a 1, la covariable no es informativa: da lo mismo en términos del *odds* de reprobar.

Sin embargo, *odds ratio* no da una significancia de la relación.

Por otra parte se puede encontrar el log(odds ratio), dado por

$$\begin{aligned}\log(\text{odds ratio}) &= \log(\text{odds reprobar} \mid \text{Asiste}) - \log(\text{odds reprobar} \mid \text{NAsiste})\\ &= \log(10/280) - \log(190/20)\\ &= -5.58\end{aligned}$$

Que como puede verse, mide la diferencia de los *odds* de reprobar e indica en cuánto asistir a la clase disminuye (en escala logarítmica) los *odds* de reprobar.

Bajo la hipótesis nula de variables independientes, note el cálculo del *odds ratio*:

```
set.seed(1)
n_tot = 500
rval_reprobado  = runif(n_tot)
prop_reprobado  = 200/500
h0_reporbado    = ifelse(rval_reprobado < prop_reprobado,"A.Reprobar","B.Aprobar")

rval_asistencia = runif(n_tot)
prop_asistencia = 290/500
h0_asistencia   = ifelse(rval_asistencia < prop_reprobado,"A.Asiste","B.NAsiste")

h0_tablacont    = table(h0_asistencia,h0_reporbado)
print(h0_tablacont)
```

```
##              h0_reporbado
## h0_asistencia A.Reprobar B.Aprobar
##     A.Asiste          82       117
##     B.NAsiste        124       177
```

```
h0_or           = (h0_tablacont[1,1]/h0_tablacont[1,2])/(h0_tablacont[2,1]/h0_tablacont[2,2])
print(h0_or)
```

```
## [1] 1.000414
```

La distribución del log(odds ratio) bajo la hipótesis nula de variables independientes está dada por:

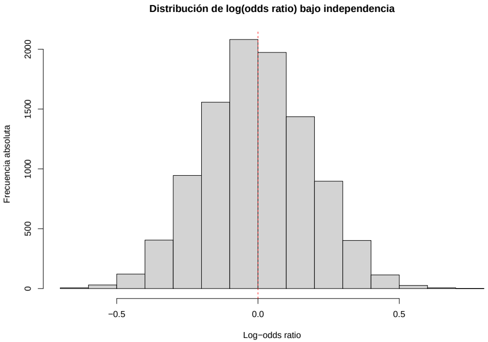
```
set.seed(1) ; h0_or = NULL ; n_tot = 500

for(i in 1:10000){
rval_reprobado  = runif(n_tot) ; prop_reprobado  = 200/500
h0_reporbado    = ifelse(rval_reprobado < prop_reprobado,"A.Reprobar","B.Aprobar")

rval_asistencia = runif(n_tot) ; prop_asistencia = 290/500
h0_asistencia   = ifelse(rval_asistencia < prop_reprobado,"A.Asiste","B.NAsiste")

h0_tablacont    = table(h0_asistencia,h0_reporbado)
h0_or[i]        = log((h0_tablacont[1,1]/h0_tablacont[1,2])/(h0_tablacont[2,1]/h0_tablacont[2,2]))
}

hist(h0_or,main = "Distribución de log(odds ratio) bajo independencia",
          xlab  = "Log-odds ratio", ylab ="Frecuencia absoluta")
abline(v=0,lty=2,col="red")
```



La transformación sigue una distribución normal, de hecho, de manera exacta, con media 0 y varianza

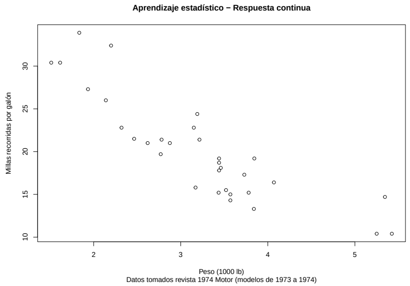
$$\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}$$

Es decir el log(odds ratio) es normal. Note que con esta distribución podemos determinar la significancia del log(odds ratio) calculado, a esto lo llamamos el test de Wald.

Regresión lineal vs regresión logística

Regresión lineal vs regresión logística

Si revisitamos la gráfica de dispersión de los datos de velocidad podemos establecer con claridad una relación entre estas dos variables.



Una relación entre las variables se da de la siguiente forma $Y = \beta_0 + \beta_1 X + \epsilon$.

- R^2 es la **proporción** de la varianza en Y explicada por el regresor X (similar al *odds ratio*/ *log-odds ratio*)
- F es la **relación** entre la varianza en Y explicada por el regresor X respecto a la que deja de explicar (similar al test de Wald).

En ML buscamos estimar f en $Y = f(X) + \epsilon$, que para RLS resulta ser

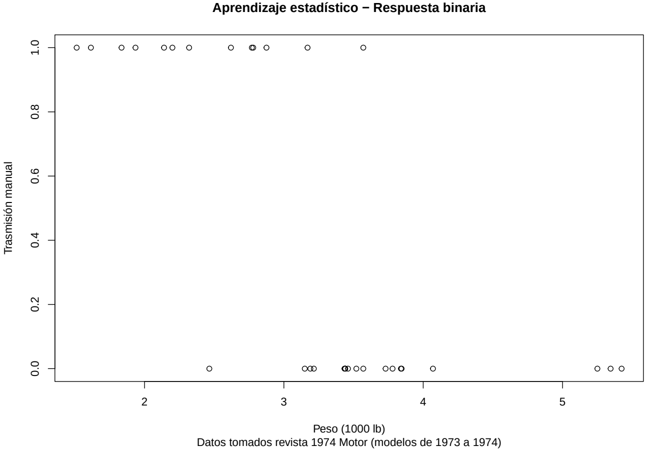
$$f(X) = \beta_0 + \beta_1 X$$

En RLM

$$f(X) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$$

Con X_i covariables discretas o continuas. Cada una con su interpretabilidad bajo el model ajustado.

La diferencia fundamental de la **regresión logística** con RLS/RLM es que nuestra variable respuesta es **binaria**:



En este caso tratamos ahora de un modelo lineal generalizado. Una nueva extensión del método de regresión lineal clásico.

Modelos lineales generalizados

Modelos lineales generalizados

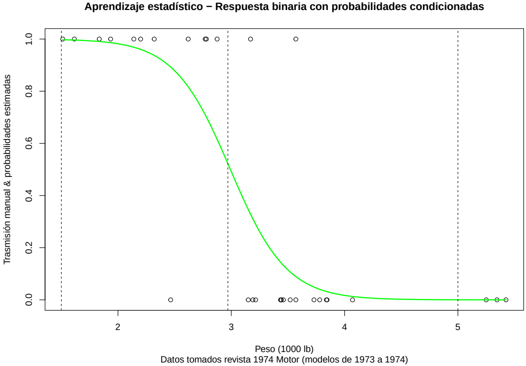
Los modelos lineales generalizados (GLM por sus siglas en inglés) están caracterizados por 3 componentes:

1. Componente aleatorio: variable respuesta Y con una distribución dada.
2. Componente sistemático: covariables para el modelo $X_1, \ldots, X_p$.
3. Función de enlace: Función que relaciona el valor esperado de Y con las covariables $X_1, \ldots, X_p$ de manera lineal.

Diferentes distribuciones de Y dan cabida a una función de enlace especial llamada función de enlace canónico.

Regresión logística

Anteriormente se modelaba el valor esperado de la variable respuesta, siendo esta continua. Nuevamente se modela $\mu(x) = E(Y|X = x)$, sólo que esta vez este valor se encuentra en $[0, 1]$



- Vehículo muy liviano, es altamente probable que sea manual ($y = 1$).
- Vehículo muy pesado, es altamente probable que sea automático ($y = 0$).

Destacando los componentes del GLM se tiene

1. Componente aleatorio: variable respuesta Y con una distribución binomial (éxito y fracaso). En total son 32 carros, entonces $(Y_1, ..., Y_{32})$ son v.a, que se asumen independientes.
2. Componente sistemático: covariables para el modelo $X_1, \ldots, X_p$. En este caso solo tenemos una covariable X igual al peso del vehículo. Una vez observada se relaciona linealmente mediante el predictor lineal:

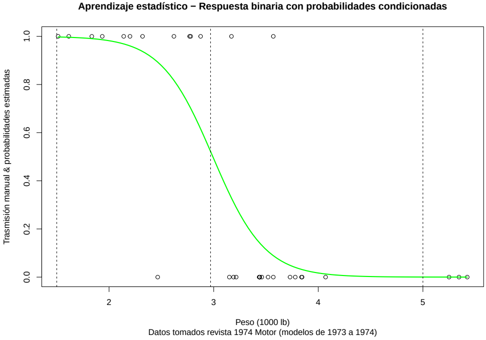
$$\beta_0 + \beta_1 x$$

3. Función de enlace: Función que relaciona el valor esperado de Y con las covariables de manera lineal.

$$g(\mu(x)) = \beta_0 + \beta_1 x$$

Interpretación del modelo

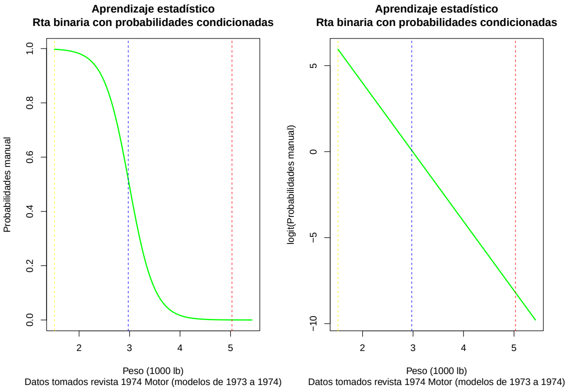
En RLS/RLM, recuerde que nuestra respuesta no está acotada, sin embargo, para el escenario logístico lo está: debe ser una probabilidad (de auto manual) que depende del peso x: $\pi(x)$



Podemos transformar $\pi(x)$ para tener un escenario no acotado como el de RLS/RLM mediante la función de enlace canónico para la distribución bernoulli

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

Con lo cual, $\pi(x)$, la probabilidad de que un carro sea manual dado su peso x, es modelada en el intervalo $(-\infty, +\infty)$.



Y los coeficientes del modelo se presentan en la escala $\text{logit}(\pi(x))$.

Para volver a la escala original (de logit a probabilidad), la función inversa del logit es

$$S(x) = \frac{1}{1 + \exp{(x)}}$$

La cual es llamada función sigmoide S o logística. Y con esto se tiene que

$$S(\text{logit}(\pi(x))) = \pi(x) = \frac{1}{1 + \exp{(\beta_0 + \beta_1 x)}}$$

Siendo así la probabilidad de Y modelada a través de $X = x$.

Al ajustar el modelo desde R

```
logistic_model <- glm(am ~ wt, data=mtcars, family=binomial)
summary(logistic_model)$coefficients
```

```
##             Estimate Std. Error   z value    Pr(>|z|)
## (Intercept) 12.04037   4.509706  2.669879 0.007587858
## wt          -4.02397   1.436416 -2.801396 0.005088198
```

Se tiene que:

$$\text{logit}(\pi(x)) = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = 12.04 - 4.02x$$

- Intercepto: En peso $x = 0$, $\text{logit}(\pi(x))$ es $12.04 > 0$, es decir, un carro de peso 0 aumenta (en escala logarítmica) el *odds* de ser manual.
- Pendiente: Al incrementar una unidad de peso, se espera una disminución en $\text{logit}(\pi(x))$ de $-4.024 < 0$, es decir, el incremento de peso disminuye (en escala logarítmica) el *odds* de ser vehículo manual.

Al contar con una variable discreta en el modelo (como el caso de reprobar y asistencia a la clase), se tiene:

```
logistic_model2 <- glm(Reprobar ~ Asiste, data=data_course_b, family=binomial)
summary(logistic_model2)$coefficients
```

```
##              Estimate Std. Error    z value     Pr(>|z|)
## (Intercept)  2.251292  0.2350812   9.576657 1.002367e-21
## AsisteSi    -5.583496  0.3985405 -14.009861 1.356709e-44
```

El modelo ajustado es igual a:

$$\text{logit}(\pi(x)) = \begin{cases} 2.25 = \log(190/20), & \text{si } x = 0 \\ 2.25 - 5.58 = \log(190/120) - \log(10/280), & \text{si } x = 1 \end{cases}$$

- Intercepto: El *odds* (en escala logarítmica) de reprobar cuando una persona no asiste a clase es 2.25.
- Pendiente: Asistir a la clase disminuye (en escala logarítmica) los *odds* de reprobar en 5.58.

Estimación del modelo

Estimación del modelo

Ajustar el modelo de regresión logística no es posible mediante MCO, ya que el concepto de residuales:

$$e_i = y_i - \hat{y}_i$$

No se mantiene. Requerimos utilizar **máxima verosimilitud**.

Revisión de MLE - Probabilidad y verosimilitud

Anteriormente revisamos la *fmp*/*fdp*.

- El caso discreto caracteriza la medida en R mediante *fmp*, para todo real x:

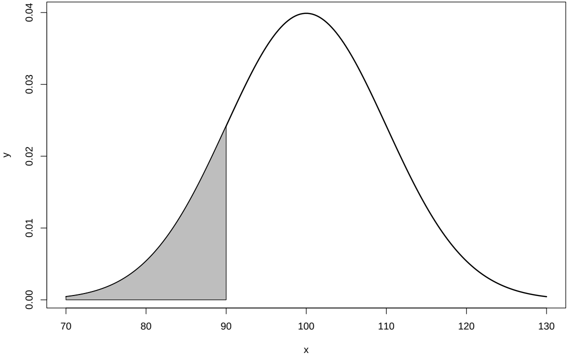
$$P(X = x) = F_X(x) - F_X(x^-)$$

- El continuo mediante la *fdp*, para todo real x:

$$F_X(x) = \int_{-\infty}^{x} f_X(x)dx$$

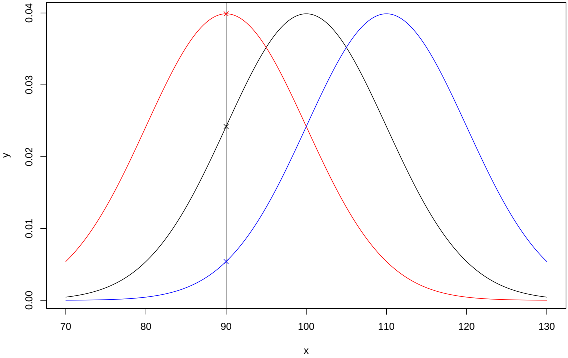
Recuerde que en ambos casos estas probabilidades son calculadas sobre **eventos** del experimento: $\{\omega \in \Omega : X(\omega) \in I\}$.

Por ejemplo, si una v.a. $X \sim N(\mu = 100, \sigma = 10)$, tenemos que la probabilidad del evento *obtener una observación menor o igual a 90* (es decir, $\{X \leq 90\}$) corresponde a:



Note que las probabilidades son calculadas una vez caracterizada/fijada la *fdp* (de manera semejante en el caso discreto).

Por otra parte, la verosimilitud no se calcula sobre eventos sino sobre valores en **el recorrido de la variable** y además **puede calcularse bajo múltiples fdp**:



Claramente 90 es mas **verosímil** en para la fdp de color rojo.

En resumen:

- La probabilidad se calcula con una *fdp*/*fmp* fija.
- La verosimilitud se le calcula a un dato fijo.

MLE para la regresión logística

El modelo subyacente de la regresión logística es el siguiente, para $i = 1, ...n$:

$$Y_i|(X_i = x_i) \sim Ber(\pi(x_i))$$

Con

$$\pi(x_i) = \frac{1}{1 + \exp\left(\beta_0 + \beta_1 x_i\right)} = S(\beta_0 + \beta_1 x_i)$$

O equivalentemente

$$\text{logit}(\pi(x_i)) = \beta_0 + \beta_1 x_i$$

Haciendo clara la linealidad sobre el logit y no sobre la probabilidad.

Por lo cual

$$P(Y=y|X=x)=\begin{cases} S(\beta_0+\beta_1 x), & \text{si } y=1 \\ 1-S(\beta_0+\beta_1 x), & \text{si } y=0 \end{cases}$$

De manera más concisa, para una observación (x_i, y_i) se tiene que

$$P(Y=y_i|X=x_i)=S(\beta_0+\beta_1 x_i)^{y_i}[1-S(\beta_0+\beta_1 x_i)]^{1-y_i}$$

Con lo que la verosimilitud bajo independencia está dada por

$$L(\beta_0,\beta_1|(x_1,y_1),...,(x_n,y_n))=\prod_i S(\beta_0+\beta_1 x_i)^{y_i}[1-S(\beta_0+\beta_1 x_i)]^{1-y_i}$$

- Puede agregar más variables para pronosticar la transmisión del vehículo: tanto discretas como continuas. La verosimilitud se mantiene.
- Recuerde el problema del *signo zodiacal* al añadir covariables.

Regresión logistica en problemas de prediccíon

Regresión logistica en problemas de prediccíon

Note que el resultado del modelamiento se encuentra en $[0,1]$, pero en clasificación, el resultado se encuentra en $\{0,1\}$. Suponga que quiere determinar si el modelo ajustado puede ser utilizado para determinar si el vehículo es automático o manual dependiendo de su peso.

```
logistic_model <- glm(am ~ wt, data=mtcars, family=binomial)
pred_resp      <- predict(logistic_model,type="response")
```

Así tenemos $\hat{\pi}(x_i)$ para $i = 1, ..., 32$. Necesitamos un valor umbral a partir del cual tomar la decisión de que el carro sea manual $(y = 1)$.

Si fuera 0.5 tenemos

```
umbral = 0.5
pred_resp_val = ifelse(pred_resp > umbral,'P1.Manual','P2.Auto')
obs_resp_val  = ifelse(mtcars$am == 1,'01.Manual','02.Auto')

tabla_resumen = table(obs_resp_val,pred_resp_val)
tabla_resumen
```

```
##              pred_resp_val
## obs_resp_val P1.Manual P2.Auto
##    01.Manual        11       2
##    02.Auto           1      18
```

Con esta tabla podemos calcular la bondad del umbral seleccionado a través de dos números resumen

$$\text{Sensibilidad}(0.5) = \frac{\text{Manuales } (y = 1) \text{ predichos correctamente}(0.5)}{\text{Manuales totales}} = \frac{11}{13}$$

$$\text{Especificidad}(0.5) = \frac{\text{Automáticos } (y = 0) \text{ predichos correctamente}(0.5)}{\text{Automáticos totales}} = \frac{18}{19}$$

Obtenidos como

```
sens  = tabla_resumen[1,1] / (tabla_resumen[1,1] + tabla_resumen[1,2])
sens
```

```
## [1] 0.8461538
```

```
espec = tabla_resumen[2,2] / (tabla_resumen[2,1] + tabla_resumen[2,2])
espec
```

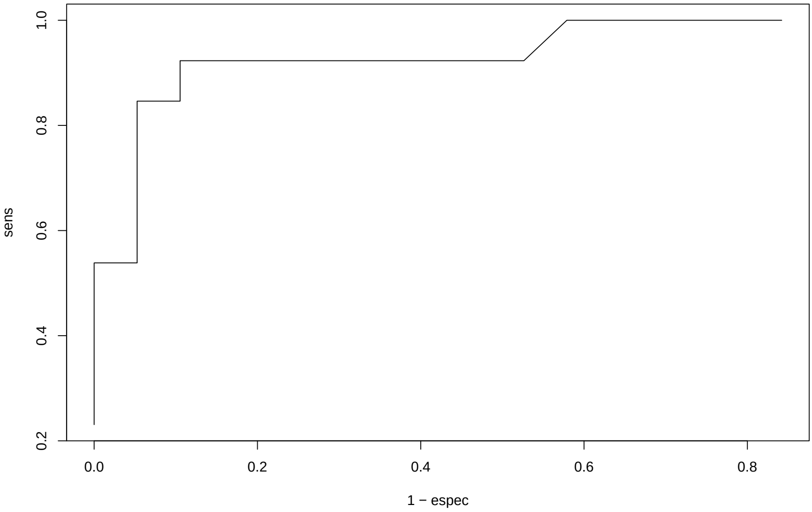
```
## [1] 0.9473684
```

Tomando una secuencia de umbrales podemos encontrar una curva que resume la bondad del modelo en hacer predicciones (con los datos de entrenamiento en este caso)

```
umbrales     = seq(0.01,0.99,by = 0.01)
obs_resp_val = ifelse(mtcars$am == 1,'01.Manual','02.Auto')
sens = espec = NULL
for(i in 1:length(umbrales)){
  umbral       = umbrales[i]
  pred_resp_val = ifelse(pred_resp > umbral,'P1.Manual','P2.Auto')
  tabla_resumen = table(obs_resp_val,pred_resp_val)
  sens[i]  = tabla_resumen[1,1] / (tabla_resumen[1,1] + tabla_resumen[1,2])
  espec[i] = tabla_resumen[2,2] / (tabla_resumen[2,1] + tabla_resumen[2,2])
}
```

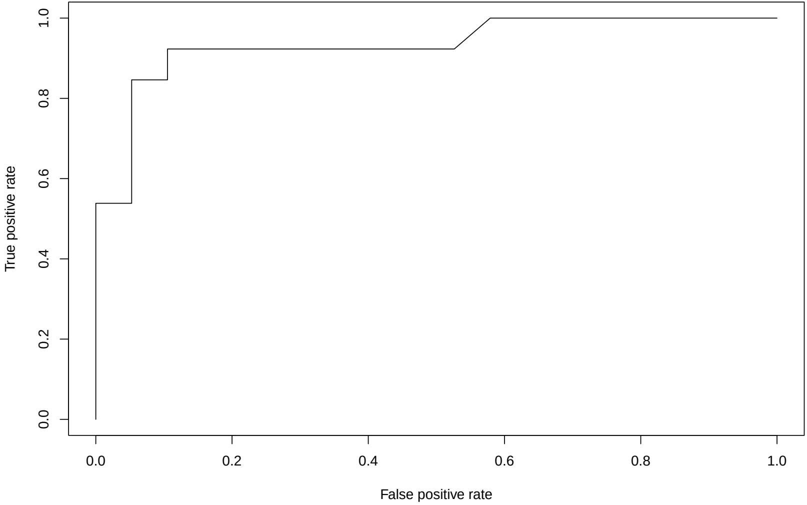
Se tiene con esto la curva ROC mediante

```
plot(1-espec,sens,type = 'l')
```



Puede calcularse de manera automática mediante

```
library(ROCR)
pred <- ROCR::prediction(pred_resp, mtcars$am)
perf <- ROCR::performance(pred, "tpr", "fpr")
plot(perf, colorize=FALSE)
```



1. Comparar las curvas ROC para múltiples modelos de predicción, logísticos o no, permite evaluar el poder de predicción de cada uno y seleccionar el más apropiado. Es importante realziar esta comparación en el marco de la validación externa.
2. Entre mayor área abarque la curva ROC, mejor será el poder de discriminación de las dos categorías. El AUPRC resume en un sólo número esta noción.

Referencias

Referencias

1. Hastie, T., Tibshirani, R., Friedman, J. The Elements of Statistical Learning. Springer.
2. Garet, Witten, Hastie, Tibshirani. Introduction to Statistical Learning with R.