Ejemplo: Default Data Set

El siguiente conjunto de datos se encuentra en la librería `ISLR`, tiene 10000 observaciones de individuos que usan tarjeta de crédito. Se conocen las siguientes características:

- **balance:** balance de la tarjeta de crédito.
- **income:** ingreso.
- **student:** variable binaria, indica si el individuo es estudiante.
- **default:**variable binaria que indica si el individuo no paga la tarjeta.

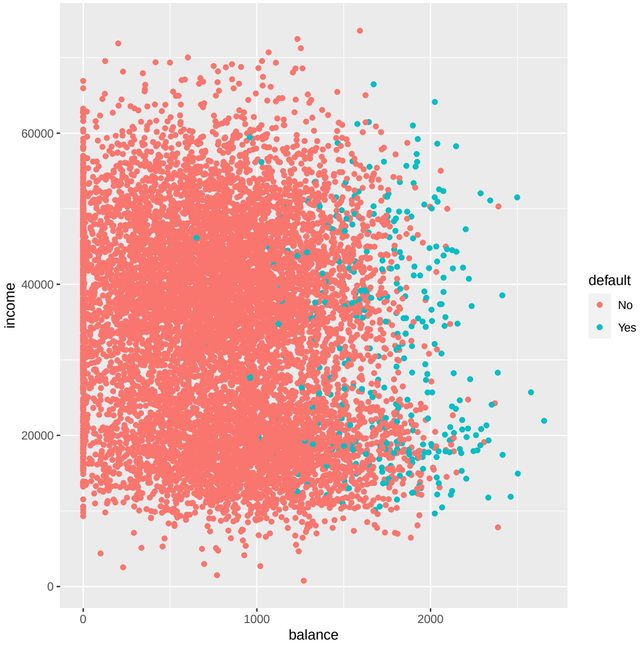
Objetivo: ajustar un modelo para predecir si una persona va a pagar la tarjeta de crédito.

Exploremos los datos

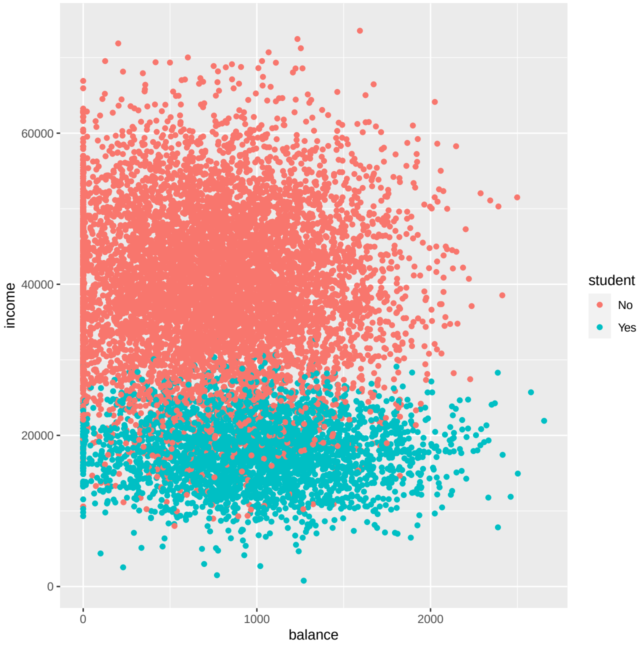
```
library(ISLR)
library(ggplot2)
summary(Default)
```

	default	student	balance	income
X	No :9667	No :7056	Min. : 0.0	Min. : 772
X.1	Yes: 333	Yes:2944	1st Qu.: 481.7	1st Qu.:21340
X.2			Median : 823.6	Median :34553
X.3			Mean : 835.4	Mean :33517
X.4			3rd Qu.:1166.3	3rd Qu.:43808
X.5			Max. :2654.3	Max. :73554

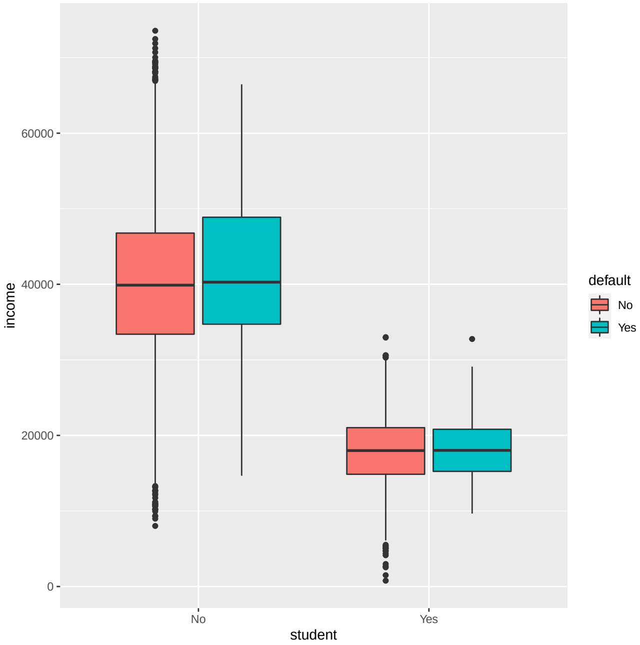
```
names(Default)
attach(Default)
```



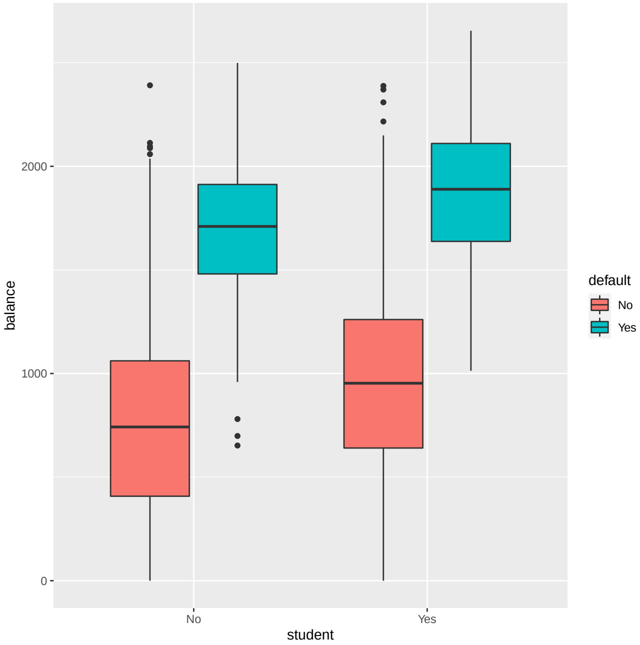
60000
40000
20000
0
income
0
1000
2000
balance
default
No
Yes



income
balance
60000
40000
20000
0
0
1000
2000
student
No
Yes



income
60000
40000
20000
0
No
Yes
student
default
No
Yes


balance
2000
1000
0
No
Yes
student
default
No
Yes

Modelos Logístico

Comenzaremos centrándonos en modelos donde la variable de respuesta es binomial

$$y_i = \begin{cases} 1 & \text{si la } i\text{-ésima observación tiene la característica estudiada.} \\ 0 & \text{en caso contrario.} \end{cases}$$

Es decir, y_i es una realización de la variable aleatoria Y_i que tiene distribución $B(1, \pi_i)$, donde $P(Y_i = 1) = \pi_i$, luego

$$\begin{aligned} E(Y_i) &= \pi_i. \\ var(Y_i) &= \pi_i(1 - \pi_i). \end{aligned}$$

Observación: la media y la varianza dependen de π_i

La transformación logística

La primer propuesta sería estimar la π_i como una combinación lineal de variables regresoras

$$\pi_i = \mathbf{x_i}'\beta,$$

donde β es un vector de p regresores.
Problema π_i está entre $[0, 1]$ pero $\mathbf{x_i}'\beta$ no, por lo tanto se propone aplicar una transformación 1 a 1 entre el intervalo $[0, 1]$ y $\mathbb{R}_{>0}$.
Proponemos calcular la razón de probabilidad,

$$odds_i = \frac{\pi_i}{1 - \pi_i}$$

Razón de probabilidades

Es el cociente entre la probabilidad de que ocurra un evento A_i y la probabilidad de que ocurra su complemento A_i^C. Es fácil de interpretar.

- En un juego justo ($P(ganar) = 0.5$) la razón de probabilidades de ganar es 1. Es decir por cada vez que gano una pierdo.
- Si tengo $1/3$ de probabilidad de ganar, entonces la razón de probabilidades es $1/2$, por cada vez que gano 2 pierdo.
- Si la probabilidad de ganar es $1/37$ (la ruleta), la razón de probabilidades de ganar es $1/36$, por cada vez que gano 36 pierdo.

En algunos contextos es usual expresarse en términos de la razón de probabilidades, hay una biyección entre ambas formas, es importante tener en cuenta que la razón de probabilidades es no acotada.

La transformación logística

Continuamos ahora aplicando el logaritmo a la razón de probabilidades, puesto que es un transformación biyectiva entre $\mathbb{R}_{>0}$ y $\mathbb{R}$. Esta transformación se conoce como logit o log odds ratio

$$\eta_i = logit(\pi_i) = log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x_i}'\beta.$$

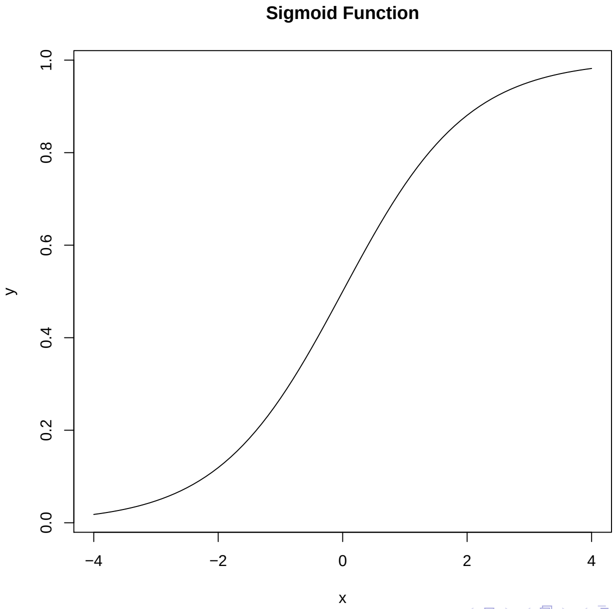
Si $logit(\pi_i) < 0$ entonces $\pi_i < 0.5$.

La transformación logística

La inversa de la función logit, la *antilogit* está dada por

$$\pi_i = antilogit^{-1}(\eta_i) = \frac{e^{\eta_i}}{1 + e^{\eta_i}}.$$

El modelo de regresión logística



El modelo de regresión logística

Sean $y_i, \ldots, y_n$ son realizaciones de una variables binomial $Y_i \sim Bi(1, \pi_i)$. Asumimos la siguiente relación subyacente entre el logit de π y los regresores $\mathbf{x_i}$ es lineal,

$$logit(\pi_i) = \mathbf{x_i}'\beta$$

El modelo que queda definido es un modelo lineal generalizado, con respuesta binomial y función de enlace logit.

El modelo de regresión logística: Interpretación

En cuanto a la interpretación, es análoga a la de un modelo lineal teniendo en cuenta que la respuesta ya no es más una media sino que es un logit.
β_j representa la modificación del cambio unitario de la variables x_j del logit de π_i, dejando fijas las otras variables. Puede resultar difícil de interpretar.
Si $\beta_j > 0$ la probabilidad de que ocurra el evento que estamos estudiando aumenta, mientras que si $\beta_j < 0$ la probabilidad disminuye.

El modelo de regresión logística: Interpretación

El *$odds_i$* resulta más sencillo de interpretar.

$$\frac{\pi_i}{1-\pi_i} = exp\{\mathbf{x_i}'\beta\}.$$

Esto define un modelo multiplicativo para la razón de probabilidades.
Por cada incremento unitario que se produzca en la la variable x_j, dejando fijas el resto de las variables, la razón de probabilidad se incremetará en $exp\{\beta_j\}$. Luego, $exp\{\beta_j\}$ representa una razón de probabilidad.

El modelo de regresión logística: Interpretación

Por otro lado, tenemos que

$$\pi_i = \frac{exp\{\mathbf{x_i}'\beta\}}{1 + exp\{\mathbf{x_i}'\beta\}}$$

no se puede interpretar en términos de aumentos unitarios de las variables regresoras. Para tener una interpretación aproximada derivamos respecto de x_{ij}

$$\begin{aligned}
\frac{\partial \pi_i}{\partial x_{ij}} &= \frac{\beta_j exp\{\mathbf{x_i}'\beta\}(1 + exp\{\mathbf{x_i}'\beta\}) - \beta_j exp\{\mathbf{x_i}'\beta\} exp\{\mathbf{x_i}'\beta\}}{(1 + exp\{\mathbf{x_i}'\beta\})^2} \\
&= \frac{\beta_j exp\{\mathbf{x_i}'\beta\}}{(1 + exp\{\mathbf{x_i}'\beta\})^2} \\
&= \beta_j \frac{exp\{\mathbf{x_i}'\beta\}}{1 + exp\{\mathbf{x_i}'\beta\}} \frac{1}{1 + exp\{\mathbf{x_i}'\beta\}} \\
&= \beta_j \pi_i (1 - \pi_i)
\end{aligned}$$

El modelo de regresión logística: Interpretación

El efecto de la variable x_j en la probabilidad π_i depende de β_j y del valor de la probabilidad. Para analizarlo se suele reemplazar π_i por el valor medio en la muestra, es decir la proporción de casos con el atributo que estamos estudiando en la muestra. El resultado aproxima el efecto de la covariable cerca de la respuesta media.

Estimación

Estimaremos los parámetros β por máxima verosimilitud. Consideremos el logartimo de la función de verosimilitud.

$$\mathscr{L}(\beta) = \log L(\beta) = \sum_{i=1}^{n} y_i \log(\pi_i) + (1 - y_i) \log(1 - \pi_i)$$

Deberíamos tomar la primer y segunda derivada respecto de β, y proceder a encontrar el máximo. Esta solución no se puede resolver en forma explicítica y para hallar el estimador de β se puede usar el algoritmo IRLS (iterated reweighted least squares), que era el mismo que utilizabamos en la materia anterior para encontrar estimadores MM.

Estimación

Veamos una idea del algoritmo, supongamos que tenemos una estimación inicial $\widehat{\beta}$.
Entonces repetimos hasta alcanzar convergencia los siguientes pasos:

1. Sea $\widehat{\eta}_i = \mathbf{x_i}'\widehat{\beta}$ para $i = 1, \ldots, n$.
2. Sea $\widehat{\mu}_i = logit^{-1}(\widehat{\eta}_i)$ para $i = 1, \ldots, n$.
3. Con estos valores calculamos

$$z_i = \widehat{\eta}_i + \frac{y_i - \widehat{\mu}_i}{\widehat{\mu}_i(1 - \widehat{\mu}_i)}.$$

4. Regresamos la variable z en las covariables por mínimos cuadrados pesados, $\widehat{\beta} = (X'WX)^{-1}X'Wz$, donde W es la matriz diagonal donde los pesos están dados por $w_{ii} = \widehat{\mu}_i(1 - \widehat{\mu}_i)$.

Para más detalles
`https://data.princeton.edu/wws509/notes/a2.pdf`

Estimación

Observaciones:

- La estimación incial de β se puede considerar, calculando

$$z_i = \log\left(\frac{y_i + 0.5}{1 - y_i + 0.5}\right)$$

  para evitar que numerador o denominador sea 0. Luego se regresa está cantidad, z_i versus x_i.
- IRLS es equivalente a minimizar por Newton-Rapshon.
- El estimador es consistente y su varianza asintótica es

$$var(\widehat{\beta}) = (X'WX)^{-1}.$$

Default Data

Ajustamos el modelo

$$\log\left(\frac{P(Default=1|Bal,Inc,St)}{1-P(Default=1|Bal,Inc,St)}\right) = \\ = \beta_0+\beta_1 Bal+\beta_2 Inc+\beta_3 St$$

```
salida1=glm(default~.,family = binomial,data =
Default)
salsum=summary(salida1)
```

Modelo ajustado

	Estimate	Std. Error	z value	Pr(>|z|)
			Bj/sd(Bj)	
(Intercept)	-10.87 Bo(hat)	0.49 sd(Bo)	-22.08	0.00
studentYes	-0.65 B1(hat)	0.24 sd(B1)	-2.74	0.01
balance	0.01 B2(hat)	0.00 sd(B2)	24.74	0.00
income	0.00 B3(hat)	0.00 sd(B3)	0.37	0.71

Table: Coeficientes estimados y errores estándares

Luego, el modelo queda

$$\log\left(\frac{P(Default = 1|Bal, Inc, St)}{1 - P(Default = 1|Bal, Inc, St)}\right) =$$
$$= -10.87 + 0.01 Bal + 0.00 Inc - 0.65 St$$

Default Data

- La variable *income* pareciera ser irrelevante.
- A mayor *balance* mayor probabilidad de defaultear la tarjeta, por cada $ que aumenta el *balance* el logaritmo de la razón de probabilidades de no pagar la tarjeta aumenta en 0.01, es decir, $e^{-0.01} = 0.99$, es el cambio en la razón de probabilidades.
- Los *estudiantes* tienen menor probabilidad de defaultear la tarjeta, el logaritmo de la razón de probabilidades en este caso es -0.65.
- Solamente se pueden interpretar en término de las razones de probabilidad.

Test de Bondad de Ajuste

Una vez ajustado el modelo es natural preguntarse cúan bien ajusta a los datos. Una medida de discrepancia es le estadístico deviance, que está dado por,

$$D = 2\sum_{i=1}^{n} y_i \log\left(\frac{y_i}{\widehat{\mu}_i}\right) + (1-y_i)\log\left(\frac{1-y_i}{1-\widehat{\mu}_i}\right)$$

donde y_i es el valor observado y $\widehat{\mu}_i$ es el valor estimado para la i-ésima observación.

La deviance es como la suma de los cuadrados de los residuos en regresion lineal [sumatoria(y - y(hat)/n]

Test de Bondad de Ajuste

Este estadístico proviene de plantear el test de cociente de verosimilitud generalizado para

$$H_0 : \beta \text{ tiene } q \text{ coordenadas no nulas}$$

versus

$$H_A : \beta \text{ puede tomar cualquier valor en el complemento de } H_0.$$

Luego D tiene distribución asintótica χ^2_{p-q} donde q es la cantidad de parámetros estimados por máxima verosimilitud bajo H_0 y p es la cantidad de parámetros estimados bajo $H_0 \cup H_A$.

Rechazo H_0 si $D > \chi^2_{p-q,1-\alpha}$

Default Data: test de bondad de ajuste

n-1 = Yo tenia 10.000 datos

```
summary(salida)
...
Null deviance:  2920.6 on 9999 degrees of freedom
Residual deviance:  1571.5 on 9996 degrees of freedom
AIC: 1579.5
Number of Fisher Scoring iterations:  8
```

Hacemos el test de bondad de ajuste, queremos

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \text{ vs } H_A : \exists \beta_j \neq 0, j = 1, 2, 3.$$

```
1-pchisq(salida1$null.deviance-salida1$deviance,
salida1$df.null-salida1$df.residual)
0
```

Rechazamos H_0 alguna variable es significativa para la regresión.

Test de Hipótesis para las variables individuales

Se puede ver que

$$\frac{\widehat{\beta}_j - \beta_j}{SE(\widehat{\beta})} \approx N(0,1).$$

Luego,

$$H_0 : \beta_j = 0 \quad \text{vs} \quad H_A : \beta_j \neq 0.$$

$$\text{Rechazo } H_0 \text{ si } \left| \frac{\widehat{\beta}_j}{SE(\widehat{\beta})} \right| > z_{1-\alpha/2}$$

Default Data: intervalos individuales

	Estimate	Std. Error	z value	Pr(>|z|)	Decisión
(Intercept)	-10.87	0.49	-22.08	0.00	Rech. H_0
studentYes	-0.65	0.24	-2.74	0.01	Rech. H_0
balance	0.01	0.00	24.74	0.00	Rech. H_0
income	0.00	0.00	0.37	0.71	No Rech. H_0

Table: Coeficientes estimados y errores estándares

Intervalos de confianza para los parámetros

Intervalo de confianza de nivel $1 - \alpha$ **para** β_j

$$\widehat{\beta}_j \pm z_{1-\alpha/2} SE(\widehat{\beta})$$

Intervalo de confianza de nivel $1 - \alpha$ **para** e^{β_j}

$$\left[\exp\left(\widehat{\beta}_j - z_{1-\alpha/2} SE(\widehat{\beta})\right), \exp\left(\widehat{\beta}_j + z_{1-\alpha/2} SE(\widehat{\beta})\right)\right]$$

Default Data:Intervalos de confianza para los parámetros

Calculamos los intervalos de confianza,

`confint(salida1)`

	2.5 %	97.5 %
(Intercept)	-11.86	-9.93
studentYes	-1.11	-0.18
balance	0.01	0.01
income	-0.00	0.00

Default Data:Intervalos de confianza para los parámetros

Calculamos los intervalos de confianza para e^{β_j}, representan los cambios que se producen en la razón de probabilidad por cada incremento unitario en x_j, dejando el resto de las variables fijas.

	2.5 %	97.5 %
(Intercept)	0.00	0.00
studentYes	0.33	0.83
balance	1.01	1.01
income	1.00	1.00

Residuos

Hay varias propuestas diferentes para estudiar los residuos en estos modelos, típicamente para medir la influencia de una variable específica.

Residuos de Pearson

$$\frac{y_i - \widehat{\mu}_i}{\sqrt{\widehat{\mu}_i(1 - \widehat{\mu}_i)}}$$

En general no hay que esperar que sigan distribución normal.

Residuos

Los residuos parciales permiten analizar si es conveniente transformar un predictor. Luego para la i-ésima observación en la j-ésima variable tenemos,

$$r_{ij} = x_{ij}\widehat{\beta}_j + \frac{y_i - \widehat{\mu}_i}{\widehat{\mu}_i(1 - \widehat{\mu}_i)},$$

donde $\widehat{\mu}_i$ es la probabiliadad estimada de que $y_i = 1$.
Luego, graficar x_{ij} versus r_{ij} para estimar como convendría transformar x_j, si fuera necesario.

Análisis del modelo logístico

Se extienden automáticamente al análisis del modelo logístico algunos puntos estudiados para el modelo lineal:

- Regresores cuantitativos, interacciones.
- Robustez.
- Selección de modelos.
- Validación de modelos.

Default Data: Selección de variables

Utilizamos el criterio de Akaike para seleccionar variables, en este contexto. En este casos tenemos

$$AIC_q = \frac{-2}{n}\mathscr{L}(\beta) + \frac{2q}{n}$$

q indica el número de variables.

Default Data: Selección de variables

`step(salida1)`

Start: AIC= 1579.54

default $\sim$ student + balance + income

Df Deviance AIC

- income 1 1571.7 1577.7

<none> 1571.5 1579.5

- student 1 1579.0 1585.0

- balance 1 2907.5 2913.5

Step: AIC=1577.68

default $\sim$ student + balance

Df Deviance AIC <none> 1571.7 1577.7

- student 1 1596.5 1600.5

- balance 1 2908.7 2912.7

Call: glm(formula = default ~ student + balance, family = binomial, data = Default)
Coefficients:

(Intercept)	studentYes	balance
-10.749496	-0.714878	0.005738

Degrees of Freedom: 9999 Total (i.e. Null); 9997 Residual
Null Deviance: 2921
Residual Deviance: 1572 AIC: 1578

$$\log\left(\frac{P(Default = 1|Bal, St)}{1 - P(Default = 1|Bal, St)}\right) =$$
$$= -10.87 + 0.006 Bal - 0.71 St$$

Default Data

Por lo tanto tenemos que si un individuo es estudiante

$$\log\left(\frac{P(Default = 1|Bal, St = 1)}{1 - P(Default = 1|Bal, St = 1)}\right) = \\ = -10.87 + 0.006Bal - 0.71$$

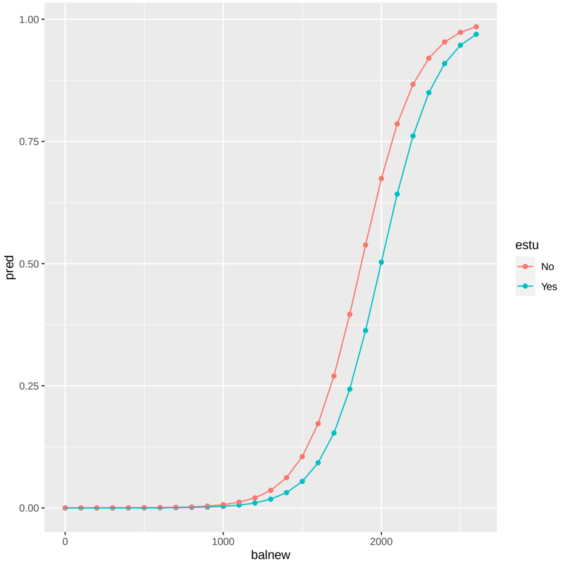
Mientras que si no es estudiante

$$\log\left(\frac{P(Default = 1|Bal, \overset{0}{St})}{1 - P(Default = 1|Bal, St)}\right) = \\ = -10.87 + 0.006Bal$$

Default Data: Gráfico de las probabilidades para estudiantes y no estudiantes

```
bal.new=seq(min(balance),max(balance),100)
new.est=data.frame(balance=bal.new,
student=factor(rep("Yes",length(bal.new))))
pred.est=predict(salida2,newdata =
new.est,type="response")
new.no.est=data.frame(balance=bal.new,
student=factor(rep("No",length(bal.new))))
pred.no.est=predict(salida2,newdata=new.no.est,
type="response")
df=data.frame(balnew=c(bal.new,bal.new),
pred=c(pred.est,pred.no.est),estu=rep(c("Yes","No"),
each=length(bal.new)))
ggplot(df, aes(x=balnew,y=pred,group=estu))+
geom_point(aes(color=estu))+
geom_line(aes(color=estu))
```

Default Data: Gráfico de las probabilidades para estudiantes y no estudiantes



Uso del modelo logístico

En general los modelos logísticos se utilizan en contextos de análisis de datos e inferencia donde el foco está puesto en entender el rol de las variables regresoras en la de respuesta.
Típicamente se busca el modelo más parsimonioso posible.

Clasificación a partir del modelo logístico

Supongamos que tenemos dos categorías **A** y **B**, a la categoría **A** le asignamos la etiqueta 1, y la **B** le asignamos la etiqueta 0.
Para cada observación $\mathbf{x}_i$ el modelo logístico asigna una probabilidad $\widehat{p}_i$,
El clasificador se define del siguiente modo

Si $\widehat{p}_i > p_0$ entonces a la i-ésima observación le le asignamos la etiqueta **A**.

Importante: para cada valor de p_0 tengo un clasificador distinto.
Cómo buscamos p_0?

Clasificación a partir del modelo logístico

Volvemos a nuestro ejemplo, veamos como funciona.
Si no tenemos información a priori entonces lo natural sería tomar $p_0 = 0.5$, sin embargo en muchos casos queremos privilegiar no cometer determinado error, en ese caso la el p_0 se determina teniendo en cuenta esta información.

Clasificación a partir del modelo logístico

```
predicciones=predict(salida2,type="response")
def.pred=rep(0,length(predicciones))
p0=0.5
def.pred[predicciones>p0]=1
table(default,def.pred)
```

	No	Yes
No	9628	39
Yes	228	105

Se clasificaron como defaulteadores a 39 individuos que no lo son y se clasificaron como no defaulteadores a 228 que lo son, esto es grave.

Clasificación a partir del modelo logístico

`confusionMatrix(confmat)`

- Accuracy : 0.9733
- 95% CI : (0.9699, 0.9764)
- No Information Rate : 0.9856
- P-Value [Acc > NIR] : 1
- Kappa : 0.4288
- Mcnemar's Test P-Value : $< 2e - 16$
- Sensitivity : 0.9769
- Specificity : 0.7292
- Pos Pred Value : 0.9960
- Neg Pred Value : 0.3153
- Prevalence : 0.9856
- Detection Rate : 0.9628
- Detection Prevalence : 0.9667
- Balanced Accuracy : 0.8530
- 'Positive' Class : No

Clasificación a partir del modelo logístico

Si nuestro proposito es detectar fraude, tenemos que tener cortes menos permisivos, por ejemplo, si la fijamos en $p_0 = 0.4$ tenemos,

	No	Yes
No	9590	77
Yes	197	136

$$
\begin{aligned}
\text{Accuracy} : &\quad 0.9726 \\
\text{Sensitivity} : &\quad 0.9799 \\
\text{Specificity} : &\quad 0.6385
\end{aligned}
$$

lógicamente aumenta la especificidad y tanto la accuracy como la sensitividad disminuyen.
Una alternativa es encontrar el valor de p_0 que minimice la distancia a la esquina $(0, 1)$ en la curva ROC.

Clases desequilibradas

El problema de **distribución de clases desequilibradas,** se da cuando la proporción de clases de una categoría es sustancialmente menor que los de otras categorías. Casos típicos:

- transacciones fraudulentas.
- identificación de enfermedades raras.
- tasas de conversión de publicidad online.
- producción de artículos defectuosos.

Los modelos usuales pueden presentar problemas de sesgo en favor de las clases dominantes al tratar de minimizar el error global e inexactitud.

Clases desequilibradas

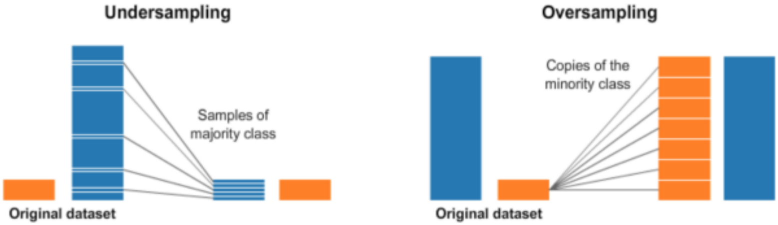
90 - 10 —> Se considera clase desbalanceada, siempre que se pueda, se deben evitar estas tecnicas

Estos metodos solo se aplican al set de entrenamiento, no en validacion ni testeo

Qué hacer en estos casos?

- A veces no es necesario hacer nada.
- Balancear al muestra de entrenamiento:
  - Submuestrear la muestra mayoritaria.
  - Sobremuestrear la muestra minoritaria.
  - Estrategia híbrida SMOTE, se generan en forma sintética observaciones distribuidas según la distribución de la clase minoritaria, sobremuestreandola y se submuestrea la clase mayoritaria.

Clases desequilibradas

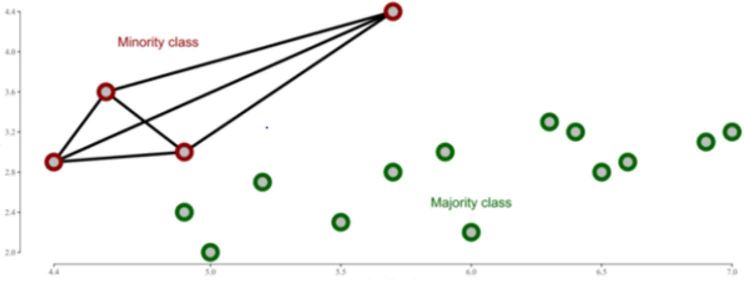


Clases desequilibradas

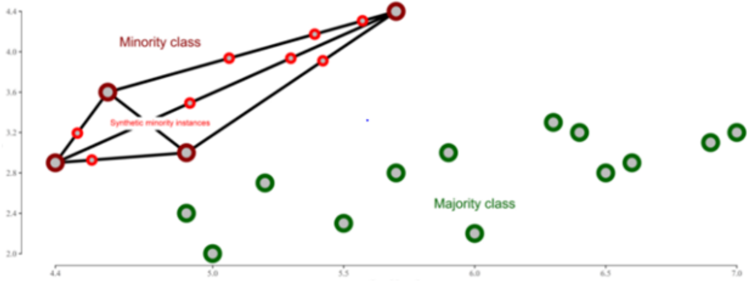
SMOTE (synthetic minority oversampling technique)
Sobremuestrear la clase minoritaria generando observaciones sintéticas interpolando linealmente observaciones de la clase minoritaria.

1. Llamamos $\mathcal{A}$ a la clase minoritaria. Para todo $x \in \mathcal{A}$ buscar los k vecinos más cercanos entre x y todas las observaciones del conjunto $\mathcal{A}$
2. Establecer a frecuencia de muestreo N de acuerdo con la proporción de desequilibrio. Para cada $x \in \mathcal{A}$, se seleccionan N observaciones entre sus vecinos más cercanos, y construyen el conjunto $\mathcal{A}_1$
3. Para cada $x_j \in \mathcal{A}_1$ generar una nueva observación $x' = x + \psi \left| x - x_k \right|$, donde $\psi \sim U[0, 1]$.

Clases desequilibradas



Clases desequilibradas



Clases desequilibradas

```
set.seed(135)
ind.def.si=which(default=="Yes")
ind.def.no=which(default=="No")
```

Tomamos la training con la misma proporción de datos de default que la muestra original.

```
train.index.ds=sample(ind.def.si,0.75*length(ind.def.si))
train.index.dn=sample(ind.def.no,0.75*length(ind.def.no))
train.index=sort(c(train.index.ds,train.index.dn))
train=Default[train.index,]
test=Default[-train.index,]
```

Clases desequilibradas

Veamos que se mantienen las proporciones
`summary(train)`

	default	student	balance	income
X	No :7250	No :5273	Min. : 0.0	Min. : 772
X.1	Yes: 249	Yes:2226	1st Qu.: 484.4	1st Qu.:21297
X.2			Median : 825.3	Median :34642
X.3			Mean : 836.9	Mean :33543
X.4			3rd Qu.:1165.8	3rd Qu.:43874
X.5			Max. :2654.3	Max. :73554

Clases desequilibradas

Balanceamos las dos clases en la muestra de entrenamiento

```
library(DMwR)
newtrain = SMOTE(default ~ balance+student, train,
perc.over = 1000,perc.under=100)
```

perc.over: controla la proporción de observaciones que hay que sobremuestrear de la muestra minoritaria. Por ej, `perc.over=600` indica que por cada observación hay que generar seis sintéticas.
perc.under: indica la proporción de observaciones de la muestra mayoritaria que hay que muestrear en relación a las observaciones que tenga la nueva muestra mayoritaria.

```
table(newtrain$default)
```

	Default
No	996
Yes	1245

Clases desequilibradas

Ahora ajustamos el modelo.

```
sal.smote=glm(default ~
balance+student,data=newtrain,family="binomial")
summary(sal.smote)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-8.36	0.39	-21.54	0.00
balance	0.01	0.00	22.93	0.00
studentYes	-0.18	0.15	-1.15	0.25

Null deviance: 3079.0 on 2240 degrees of freedom
Residual deviance: 1167.9 on 2238 degrees of freedom
AIC: 1173.9

Clases desequilibradas

La regresión es significativa, sin embargo el coeficiente que acompaña a `student` dejó de ser significativa.
El modelo quedaría en este caso

$$\log\left(\frac{P(Default = 1|Bal, St)}{1 - P(Default = 1|Bal, St)}\right) = \\ = -8.36 + 0.006Bal - 0.18St$$

Clases desequilibradas

Comparemos en la muestra de testeo, voy a tomar como valor de corte para predecir $p_0 = 0.5$

Predicho				
	Sin Equilibrar		Equilibrada	
	No	Yes	No	Yes
Real No	2407	10	2072	345
Real Yes	62	22	13	71

Si bien disminuye el accuracy, se puede ver que disminuye la probabilidad de clasificar a una observación como si no fuera a defaultear cuando lo va a hacer.

Clases desequilibradas

- nunca balancear la muestra de validación.
- si se está haciendo cross validation entonces hay que hacer el "equilibrado" de las clases dentro del procedimiento.
- medidas más informativas en estos casos son AUC y hacer el gráfico precisión-recal.

Clases desequilibradas

Problemas que pueden surgir:

- Sobre muestrear en exceso, disminuye la varianza, puede haber muchas observaciones iguales, overfitting.
- Sub muestrear en defecto, se disminuye mucho el tamaño muestral puede causar underfitting.
- Técnicas híbridas, suelen traer problemas si la clase menos representada tiene alta asimetría.

Regularización L_1 del modelo logístico

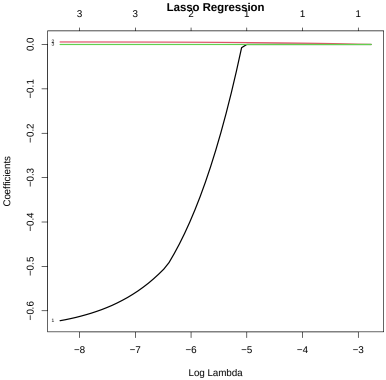
En forma análoga a la planteada en el contexto de regresión lineal, se propone penalizar la función de pérdida $\mathscr{L}$ mediante una penalidad en la norma L_1 de los coeficientes del vector de parámetros sin el intercept, $\|\beta\|_1 = \sum_{l=1}^{K} |\beta_l|$, proponiendo hallar el vector $\widetilde{\beta}$ tal que

$$\arg\max_{\beta} \mathscr{L}(\beta) - \lambda \|\beta\|_1 .$$

El parámetro λ se encuentra por cross-validatios y hay algoritmos que permiten encontrar la solución de este problema.

Default Data: lasso

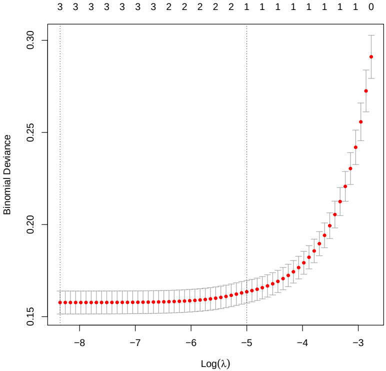
```
x=model.matrix(salida1)[,-1]
ajuste.lasso=glmnet(x, Default$default, family =
"binomial", alpha = 1)
plot(ajuste.lasso, label = T, xvar="lambda",lwd=2,
main="Lasso Regression")
```



Default Data: lasso

Vamos a cross validar el parámetro λ

```
cv.lasso= cv.glmnet(x, Default$default, alpha = 1,
family = "binomial")
plot(cv.lasso)
```



Default Data: lasso

```
lasso.cv <- glmnet(x, Default$default, alpha = 1,
family = "binomial", lambda = cv.lasso$lambda.1sd)
coef(lasso.cv,cv.lasso$lambda.1se)
```

(Intercept)	-8.708885883
studentYes	.
balance	0.004267818
income	.

$$\log\left(\frac{P(Default = 1|Bal)}{1 - P(Default = 1|Bal)}\right) = -8.71 + 0.0042Bal$$

Aplicación al problema de clasificación

Nos planteamos con el problema de clasificar a la observación $\mathbf{x_0}$ en uno de los dos grupos. Calculamos p_0, la probabilidad de que la nueva observación tenga la característica estudiada.

$$p_0 = \frac{e^{\mathbf{x_0}'\widehat{\beta}}}{1 + e^{\mathbf{x_0}'\widehat{\beta}}}.$$

Luego, si $p_0 > \tilde{p}$ decimos que tiene la característica estudiada.
El valor de corte $\tilde{p}$ se determina por cross validation.

Extensión del modelo lineal al caso de K categorías

Consideramos $k = 1, \ldots, K$ categorías, proponemos ajustar

$$\log \left(\frac{P(Y = k | X = \mathbf{x})}{P(Y = K | X = \mathbf{x})} \right) = \beta_{k0} + \mathbf{x}' \beta_k,$$

para $k = 1, \ldots, K - 1$, la probabilidad restante, para $k = K$ se calcula como el complemento.
Se plantea el modelo en término de $K - 1$ razones de probabilidad.
La elección del denominador es arbitraria.

Extensión del modelo lineal al caso de K categorías

Podemos reescribir el modelo en términos de las probabilidades

$$
\begin{aligned}
P(Y=k|X=\mathbf{x}) &= \frac{\exp\left(\beta_{k0}+\mathbf{x}'\beta_k\right)}{1+\exp\left(\sum_{l=1}^{K-1}\beta_{l0}+\mathbf{x}'\beta_l\right)}, \quad k=1,\ldots,K-1 \\
P(Y=K|X=\mathbf{x}) &= \frac{1}{1+\exp\left(\sum_{l=1}^{K-1}\beta_{l0}+\mathbf{x}'\beta_l\right)}
\end{aligned}
$$

El vector de parámetros es $\theta = \left\{\beta_{10}, \beta_1, \ldots, \beta_{(K-1)0}, \beta_{(K-1)}\right\}$
Debido al gran número de parámetros típicamente se usa para $K=2$, que es un caso muy simple.

Estimación y ajuste del modelo

El ajuste se realiza por máxima verosimilitud, teniendo en cuenta la probabilidad condicional de que $Y = k$ sabiendo que $X = \mathbf{x}$, Es decir,

$$\sum_{l=1}^{n} \log\left(P(Y = k|X = \mathbf{x}; \beta)\right)$$

Para calcular numéricamente el valor de $\widehat{\beta}$, utilizaremos nuevamente el algoritmo IRLS o de forma equivalente el algoritmo de Newton-Raphson.
La librería `glmnet` puede calcular los estimadores de β para el modelo sin regularizar.

Modelo logístico vs LDA

- La regresión logística es más general, no asumen nada sobre la distribución de las variables.
- Si las variables regresoras fueran normales, la regresión logística ignora estos datos y pierde eficiencia (tiene más varianza).
- la regresión logística es más robusta, ya que a puntos alejados les da menos peso, mientras que en LDA, todas las observaciones tienen igual peso al estimar medias y matriz de covarianza.
- predictores cualitativos, nunca son normales, conviene la regresión logística.
- en la práctica los dos métodos dan resultados similares.

Regresión Logística

Ventajas:

- computacionalmente eficiente.
- no requiere que las variables estén escaladas.
- es fácil de interpretar.
- es importante que los regresores no estén correlacionados, prestar atención a la ingeniería de variables.
- muy utilizada, puede ser un buen benchmark para comparar otras metodologías.

Regresión Logística

Desventajas:

- suele tener peor perfomance predictiva que otros algoritmos más complejos.
- no es el más adecuado para resolver problemas no lineales, ya que la superficie de separación es lineal.
- no es bueno cuando los regresores no están altamente correlacionados a la variable de respuesta.

Modelos Probit

En principio cualquier función que mapée probabilidades en los números reales prodría servir.
Sea $F(\cdot)$ una f.d.a. de una variable aleatoria definida en la recta real, por ejemplo $F = \Phi$ luego

$$\eta_i = F^{-1}(\pi_i),$$

para $\eta_i \in \mathbb{R}$. En este caso la función de enlace será F^{-1}, la regresión probit típica toma como función de enlace Φ^{-1}.
Específicamente, supongamos que el término de error $\epsilon_i \sim N(0, \sigma^2)$, asumiendo $Y_i^* = \mathbf{x}_i'\beta + \epsilon_i$

$$\begin{aligned}
\pi_i &= P(Y_i^* > 0) \\
&= P(\epsilon_i > -\mathbf{x}_i'\beta) = P(\epsilon_i/\sigma > -\mathbf{x}_i'\beta/\sigma) \\
&= 1 - \Phi(-\mathbf{x}_i'\beta/\sigma) = \Phi(\mathbf{x}_i'\beta/\sigma)
\end{aligned}$$

Modelo Probit

- No se puede identificar β en forma independiente de σ. Existen dos alternativas para poder interpretarlo, fijar $\sigma = 1$ o interpretar β en términos de desvíos estándares.
- En un rango amplio de valores esta función se parece bastante a la funcíon *logit*, ambas pasan por el punto $(0, 1/2)$, ajustando adecuadamente el σ y β prácticamente coindiden.

Default Data: Modelo Probit

```
salida3=glm(default ∼ ., family = binomial(link =
"probit"),data = Default)
sumsal3=summary(salida3)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.48	0.24	-22.96	0.00
studentYes	-0.30	0.12	-2.49	0.01
balance	0.00	0.00	24.77	0.00
income	0.00	0.00	0.51	0.61

Default Data: Modelo Probit

```
(Dispersion parameter for binomial family taken to be
1)
Null deviance:  2920.6 on 9999 degrees of freedom
Residual deviance:  1583.2 on 9996 degrees of freedom
AIC: 1591.2
Number of Fisher Scoring iterations:  8
```

El análisis de la regresión probit es análogo al de la regresión logística, para tener un vistazo de como se interpretan los coeficientes y test de hipótesis asociados se puede mirar
`https://stats.idre.ucla.edu/r/dae/probit-regression/`

Otras funciones de enlace: modelos *probit*

- *logit* fácil de interpretar, odds ratio.
- *probit* fácil de interpretar en términos de distribución acumulada.
- *logit* computacionalmente más sencilla.

https://pdfs.semanticscholar.org/7218/daab6499b46759f0a16d173d01d348bed906.pdf

Modelo Poisson

Recordemos que una variable aleatoria Y_i tiene distribución de Poisson si su probabilidad puntual esta dada por

$$P(Y = y) = \frac{\lambda^y}{y!} e^{-\lambda}, \text{ con } \lambda > 0.$$

Además, $E(Y) = var(Y) = \lambda$.

Modelo Poisson

Vinculación con procesos estocásticos.
Supongamos que determinado evento ocurre al azar siguiendo los siguientes patrones:

- la probabilidad de que ocurra un evento en un intervalo de tiempo fijo es proporcional a la longitud del intervalo.
- la probabilidad de dos o más ocurrencias en un intervalo de tiempo corto es prácticamente cero.
- la probabilidad de ocurrencias en intervalos de tiempos disjuntos son mutuamente independientes.

luego, la distribución de probabilidades del número de eventos que ocurre en un intervalo de tiempo fijo es $\mathscr{P}(\lambda t)$, donde t es la longitud del intervalo.

Modelo Poisson

Una propiedad importante de la distribución Poisson, es que dadas $Y_1 \sim \mathscr{P}(\lambda_1)$ e $Y_2 \sim \mathscr{P}(\lambda_2)$, independientes, entonces $Y_1 + Y_2 \sim \mathscr{P}(\lambda_1 + \lambda_2)$.
Esto permite trabajar de manera sencilla, tanto con datos individuales como con datos agrupados.

$$Y_{ij} = \text{cantidad de eventos ocurridos para la observación } j \text{ en el grupo } i.$$

Luego, si las observaciones son independientes $Y_{ij} \sim \mathscr{P}(\lambda_i)$, para $j = 1, \ldots, n_i$, tenemos que $Y_i \sim \mathscr{P}(\lambda_i n_i)$.

Modelo Poisson: modelo log-lineal

Sean $y_1, \ldots, y_n$ realizaciones independientes con distribución $Y_i \sim \mathscr{P}(\mu_i)$, luego dejamos que la media (y la varianza) dependan de las variables regresoras,

$$\mu_i = \mathbf{x}_i' \beta,$$

Observación: $\mu_i \in \mathbb{N}$ y $\mathbf{x}_i' \beta \in \mathbb{R}$, es un contradicción.
Segunda propuesta

$$\eta_i = \log(\mu_i) = \mathbf{x}_i' \beta,$$

Interpretación: El coeficiente β_j representa el cambio medio en el logaritmo de la media por cada unidad que cambia el predictor x_j. Al aumentar x_j en una unidad de produce un incremento de β_j en el logaritmo de la media.

Modelo Poisson: modelo log-lineal

En forma análoga,

$$\mu_i = \exp(\mathbf{x}_i'\beta)$$

Cada coeficiente de en la regresión exponencial $\exp(\beta_j)$ representa un efecto multiplicativo del j-ésimo predictor sobre la media.
El incremento unitario en x_j multiplica la media por un factor $\exp(\beta_j)$.
Es usual que los datos de conteo, los efectos de los predictores sean a menudo multiplicativos, en lugar de aditivo. Es decir, normalmente se observan pequeños efectos para pequeños registros y grandes efectos para registros grandes. Si esto ocurre, tenemos este modelo resulta natural y sencillo.

Estimación

Los parámetros se estiman por máxima verosimilitud.
Es fácil ver que

$$
\begin{aligned}
\mathscr{L}(\beta) = \log Lik(\beta) &= \sum_{i=1}^{n} y_i \log \mu_i - \mu_i \\
&= \sum_{i=1}^{n} y_i \mathbf{x}_i' \beta - \exp(\mathbf{x}_i' \beta)
\end{aligned}
$$

Derivando en β e igualando a cero, tenemos

$$\mathbf{X}' y = \mathbf{X}' \widehat{\mu}$$

$\mathbf{X}$ es la matriz de diseño.

Estimación

La estimación se realiza utilizando IRLS, donde

$$z_i = \widehat{\eta}_i + \frac{y_i - \widehat{\mu}_i}{\widehat{\mu}_i}$$

y la diagonal de la matriz W está dada por,

$$w_{ii} = \widehat{\mu}_i$$

Las estimaciones iniciales se pueden obtener aplicando la función de enlace a los datos, es decir tomar el logaritmo de la respuesta y regresándolo en los predictores via OLS. Para evitar problemas si alguna de las cuentas (variables poisson) es cero se le puede sumar una constante fija a todos las respuestas.

Bondad de ajuste del modelo

En este caso la deviance está dada por

$$D = 2\sum_{i=1}^{n} y_i \log\left(\frac{y_i}{\widehat{\mu}_i}\right) - (y_i - \widehat{\mu}_i)$$

- el primer término es igual al de la regresión logística.
- el segundo término suele ser cero.
- queremos que la deviance sea chica, implica verosimilitud alta.

Si n es suficientemente grande la distribución asintótica es χ^2_{n-p}.

Ejemplo

Consideremos los siguientes datos

Variables regresoras:

- **math:** variable continua, indica la nota en el examen final de matemática.
- **prog:** variable categórica con tres niveles que indica en que programa está inscripto el alumno, *General*, *Academic* o *Vocational*

Variable de respuesta:

- **m_awards:** número de premios que ganó un estudiante de secundario en un año.

Ejemplo

Miramos los datos

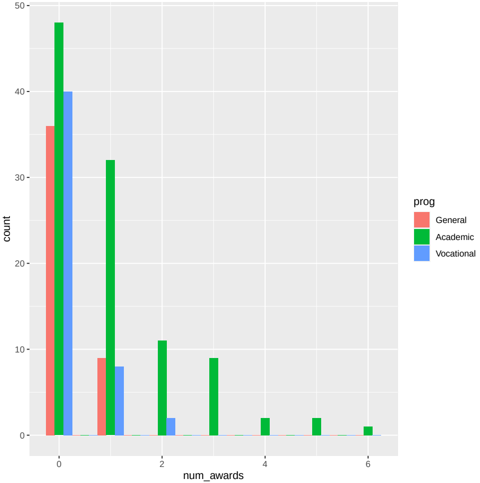
```
head(p)
```

	id	num_awards	prog	math
1	45	0	Vocational	41
2	108	0	General	41
3	15	0	Vocational	44
4	67	0	Vocational	42
5	153	0	Vocational	40
6	51	0	General	42

Ejemplo

Analizamos las medias y desvíos por programa

General	Academic	Vocational
0.20 (0.40)	1.00 (1.28)	0.24 (0.52)



Ejemplo

```
m1= glm(num_awards ~ prog + math, family="poisson",
data=p)
salpremio=summary(m1)
Deviance Residuals:
```

Min	1Q	Median	3Q	Max
-2.204	-0.844	-0.511	0.256	2.680

parece haber asimetría en los residuos, la mediana esta alejada del cero.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.25	0.66	-7.97	0.00
progAcademic	1.08	0.36	3.03	0.00
progVocational	0.37	0.44	0.84	0.40
math	0.07	0.01	6.62	0.00

Ejemplo

- El coeficiente que acompaña a la variable *math* es 0.07, en promedio se espera que por cada unidad que aumente en esta variable el logaritmo de *num_awards* aumente en 0.07.
- En promedio los individuos de programas *Académicos* tienen 1.1 $log(num\_awards)$ más que los de programas *General*.
- En promedio los individuos de programas *Vocational* tienen 0.37 $log(num\_awards)$ más que los de programas *General*.

Ejemplo

```
 (Dispersion parameter for poisson family taken to be
1)
 Null deviance:  287.67 on 199 degrees of freedom
 Residual deviance:  189.45 on 196 degrees of freedom
 AIC: 373.5
 Number of Fisher Scoring iterations:  6
```

Podemos testear el ajuste general del modelo, en este caso vemos que la deviance residual bajo en realación a la deviance bajo la hipótesis nula. Queremos verificar que no haya sobredispersión de los datos.

Para eso, aproximadamente podemos ver m1$deviance/m1$df.residual=189.6/196 ¡1 no indica sobredispersión del ajuste

Ejemplo

El test formal sería

```
pchisq(deviance, df.residual, lower.tail=FALSE)
0.6182274
```

El ajuste es adecuado.

Modelo de Poisson

El modelo quedaría

$$
\begin{aligned}
E(\texttt{num\_awards}) &= \exp\left(-5.47 + 0.07\texttt{math} +\right. \\
&+ \left. 1.08\texttt{Academic} + 0.37\texttt{Vocational}\right)
\end{aligned}
$$

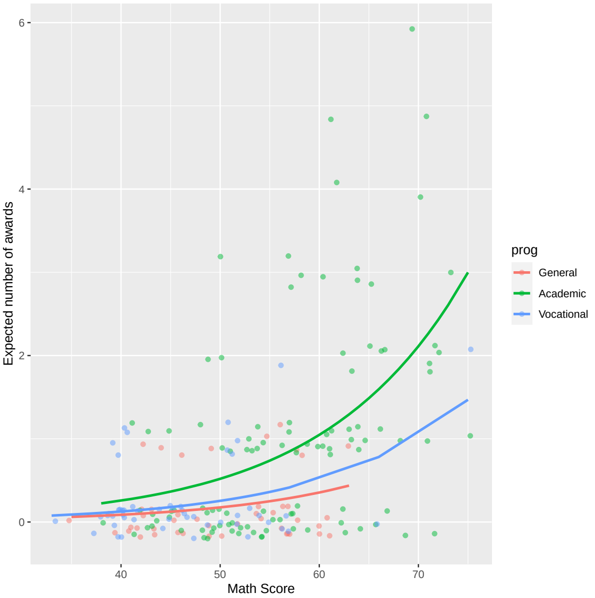
Modelo de Poisson

Predicción, dejemos fija la variable `math` en su valor medio, para cada uno de los programas.

General	Academic	Vocational
0.2114	0.6249	0.3060

- número medio predicho de eventos para el nivel `General` es 0.21
- número medio predicho de eventos para el nivel `Academic` es 0.62
- número medio predicho de eventos para el nivel `Vocational` es 0.31

Modelo de Poisson



Bibliografía

- F. Harrell.
  Regression Modelling Strategies.
  Springer-Verlag, 2015.
  Capítulo 10.
- T. Hastie, R. Tibshirani, J. Friedman.
  Elements of Statistical Learning, 2nd Ed.
  Springer-Verlag, 2009.
  Capítulo 4.
- J. Garret, D. Witten, T. Hastie, R. Tibshirani.
  An Introduction to Statistical Learning.
  Springer-Verlag, 2013.
  Capítulo 4.
- K. Murphy.
  Machine Learning. A probabilistic perspective.
  The MIT press, 2012.
  Capítulo 8 y 9.