Capítulo 1
Regresión Logística

Si un banco le da un préstamo a una persona ¿qué tan probable es que esa persona le pague? ¿Cómo modelamos esta probabilidad? De esto, en muchas ocasiones, se encarga la regresión logística.
Hagamos un ejemplo concreto, vamos trabajar con una base de datos que llamamos `base_cred` y la leemos `python` cómo sigue.
Primero instalamos la biblioteca que nos permite leer el archivo e inmediatamente llamamos a la base

```
import pandas as pd
base_cred=pd.read_csv('incumplimiento.csv')
```

y después de eso podemos imprimir la base escribiendo

```
base_cred
```

pero esto puede resultar muy engorroso a la hora de visualizar la base, así que basta ver las primeras 5 entradas,

```
base_cred.head(5)
```

```
Folio Incumplimiento  Estudiante    Saldo          Ingreso
     1      No           No          729.526495     44361.625074
     2      No           Yes         817.180407     12106.134700
     3      No           No         1073.549164     31767.138947
     4      No           No          529.250605     35704.493935
     5      No           No          785.655883     38463.495879
```

Ahora, sí queremos ver las últimas entradas de la base, escribimos `tail` en lugar de `head`.
También podemos ver la información general de nuestra base

```
base_cred.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 5 columns):
Folio              10000 non-null int64
Incumplimiento     10000 non-null object
Estudiante         10000 non-null object
Saldo              10000 non-null float64
Ingreso            10000 non-null float64
dtypes: float64(2), int64(1), object(2)
memory usage: 390.7+ KB
```

Después del breve resumen de la base que hemos importado al código, ¿cómo accedemos a cada de una de las variables? Dos maneras comunes

```
base_cred.NombredelaVariable
```

```
base_cred['NombredelaVariable']
```

y así podemos hace un análisis descriptivo de la gráfica. Antes debemos de instalar las siguientes bibliotecas

```
from matplotlib import*
from pylab import*
```

Intuitivamente ¿que variables creen que influyan en la probabilidad de incumplimiento? ¿el Ingreso? ¿el Saldo?, sería bueno graficar `Ingreso vs Saldo` y ver, a partir allí quienes incumplieron. Para esto necesitamos acceder a los datos *créditos que incumplieron e identificarlos y créditos que cumplieron e identificarlos*

```
base_cred[base_cred.Incumplimiento=='No']
base_cred[base_cred.Incumplimiento=='Yes']
```

Ya que los tenemos identificados, ahora seleccionamos `Saldo` de los que incumplieron y `Saldo` de los que cumplieron. Hacemos lo mismo con Ingresos. Esto nos va a facilitar "clasificar" a los cumplidores de los incumplidores en relación con el saldo y el ingreso.

```
s1=base_cred[base_cred.Incumplimiento=='No'].Saldo
i1=base_cred[base_cred.Incumplimiento=='No'].Ingreso
s2=base_cred[base_cred.Incumplimiento=='Yes'].Saldo
i2=base_cred[base_cred.Incumplimiento=='Yes'].Ingreso
```

y luego graficamos

```
s1=base_cred[base_cred.Incumplimiento=='No'].Saldo
i1=base_cred[base_cred.Incumplimiento=='No'].Ingreso
s2=base_cred[base_cred.Incumplimiento=='Yes'].Saldo
i2=base_cred[base_cred.Incumplimiento=='Yes'].Ingreso

figsize(16,8)
scatter(s1,i1,marker='o',s=40,edgecolors='g',facecolors='g',alpha=0.2,label='Bien
    portados')
scatter(s2,i2,marker='o',s=40,edgecolors='red',facecolors='red',alpha=0.5,label='
    Mal portados')
legend()
xlabel('Saldo')
ylabel('Ingreso')
```

y lo que debemos obtener es la figura 1.1 ¿qué observamos en esta gráfica? Nos sugiere que el `Saldo` del crédito tiene un peso importante en la probabilidad de incumplimiento. Así que, tal vez debamos de incluirlo como variable explicativa, de este modo un primer objetivo es modelar

$$Pr(Incumplimiento = Si|Saldo)$$

Haciendo una identificación adecuada, vamos a trabajar con el modelo

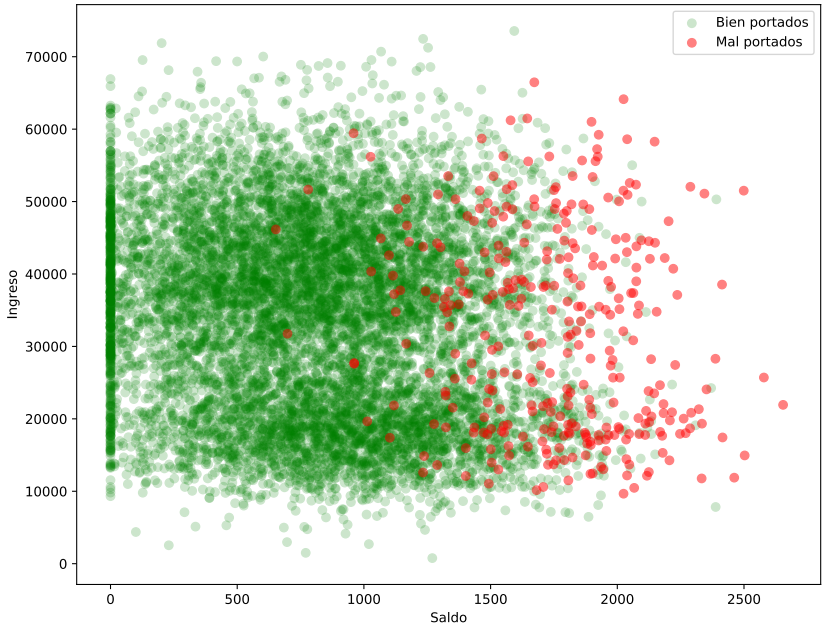

Figura 1.1: De verde aparecen las personas que incumplieron y de rojo las que no

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \tag{1.1}$$

La figura 1.1 nos sugiere que una de las variables que influye en la probabilidad de incumplimiento es `Saldo`, entonces lo que sigue ahora es graficar

`Saldo vs Incumplimiento`

Y esto lo hacemos con el siguiente código,

```
scatter(base_cred.Saldo,base_cred.Incumplimiento,)
xlabel('Saldo')
ylabel('Incumplimiento')
```

y nos devuelve la figura 1.2, qué como vemos, esta a su vez nos sugiere que debemos ajustarle una función logística. Esto lo hacemos como sigue (siempre teniendo en cuenta quien es la variable de respuesta y quien es la variable explicativa). Hacemos el ajuste,

```
import sklearn.linear_model as skl_lm
base_cred['In_num']=base_cred.Incumplimiento.factorize()[0]
y=base_cred.In_num#Variable de Respuesta
x_1 = base_cred.Saldo.values.reshape(-1,1)#Variable Explicativa
```

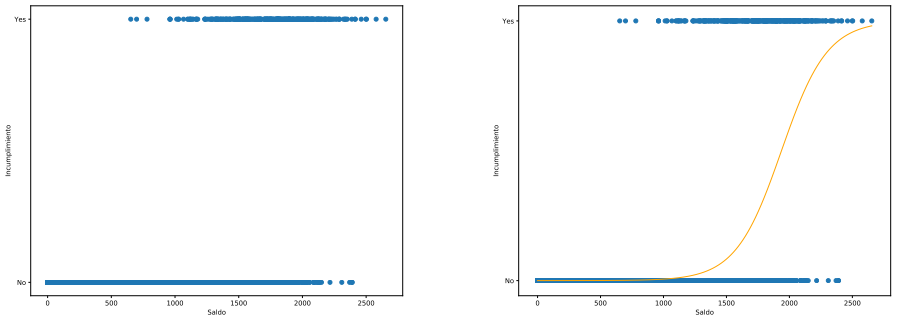

Figura 1.2: Con y sin curva ajustada

```
x_test = np.arange(base_cred.Saldo.min(), base_cred.Saldo.max()).reshape(-1,1)
clf = skl_lm.LogisticRegression(solver='newton-cg')
clf.fit(x_1,y)
prob=clf.predict_proba(x_test)
```

y luego graficamos la curva ajustada,

```
plot(prob[:,1],color='orange')
scatter(base_cred.Saldo,base_cred.Incumplimiento,)
xlabel('Saldo')
ylabel('Incumplimiento')
```

El resultado de la curva ajustada lo podemos ver en la parte de la derecha de la figura 1.2. Y entonces muchos se preguntarán ¿Es un buen ajuste? Po's quien sabe, de eso precisamente trata la tarea.

Tarea 2(Regresión Logística): Primera parte

Problema 1. *La idea de este ejercicio es ver si la curva* naranja *ajusta bien en el siguiente sentido.*

Algoritmo 1. ¿Qué tal el ajuste?

1. De la base original, calcule la proporción de incumplimiento (si etiquetamos con 1 si incumplio y con 0 si no, entonces lo que estoy pidiendo es simplemente la proporción de 1's.).
2. Calcule $\hat{\beta}_0$ y $\hat{\beta}_1$ usando `python` con `clf` (esto ya lo hicimos en clases).
3. Para cada crédito, calcule
$$f(\hat{\beta}_0 + \hat{\beta}_1 Saldo_i),$$
donde $f(x) = \frac{e^x}{1+e^x}$.
4. Luego defina, $\hat{Y}_i = 1$ si $f(\hat{\beta}_0 + \hat{\beta}_1 Saldo_i) > 0{,}5$. Y calcule la proporción de 1's y compare con la proporción original.

5. Luego calcule la proporción en donde sí coincidieron los 1's, la proporción en donde sí coincidieron los 0's, la proporción en donde en lugar de 1's salieron 0, y por último la proporción en donde en lugar de 0's salieron 1's. Explique detalladamente sus conclusiones.

Problema 2. *Luego, considere el modelo en donde tomamos como variables explicativas:* `Saldo, Estudiante e Ingreso` *y nuevamente vea el "ajuste" cómo en el algoritmo 1.*