

Examples of GLM

Logistic regression and Poisson regression

Cathy Maugis-Rabusseau

INSA Toulouse / IMT
GMM 116
cathy.maugis@insa-toulouse.fr

2023-2024

1 Logistic regression

2 Log regression

1 Logistic regression

- Modeling
 - Odds and odds ratio
 - Simple logistic regression
 - Multiple logistic regression

- A **binary** response variable Y
- Explanatory variables: $x^{(1)}, \dots, x^{(p)}$
- Example : Credit Card Default

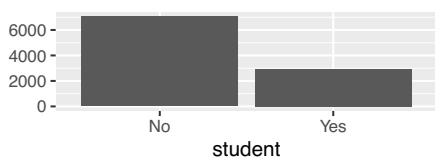
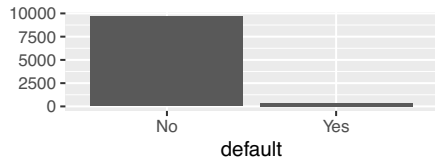
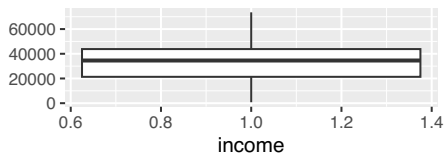
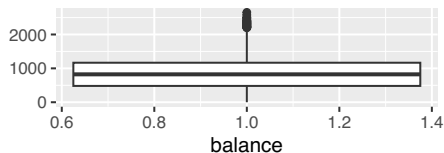
A simulated data set containing information on $n = 10000$ customers. The aim is to predict which customers will default on their credit card debt. We want to explain the binary variable *default* (1 if default, 0 otherwise) with the 3 following explanatory variables:

- *student*: A factor with levels No and Yes indicating whether the customer is a student
- *balance*: The average balance that the customer has remaining on their credit card after making their monthly payment
- *income*: Income of customer

Example

```
data(Default)  
attach(Default)  
summary(Default)
```

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



Modeling

loi de Bernoulli:

- Random component: $Y_i | \mathbf{x}_i \sim \mathcal{B}(\pi(\mathbf{x}_i))$, Y_1, \dots, Y_n indep.
- Link function g :

- **logistic function:**

π_i



$$g^{-1}(u) = F(u) = \frac{e^u}{1 + e^u} \iff g(\pi) = \ln\left(\frac{\pi}{1 - \pi}\right) = \text{logit}(\pi).$$

In this case, the model is called **logistic model**.

- **probit function:** F is the cdf of $\mathcal{N}(0,1)$ and $g = F^{-1}$ is the probit function.

↑ fonction de répartition.

- **Gompit or complementary log-log function:** F is the cdf of the Gompertz law

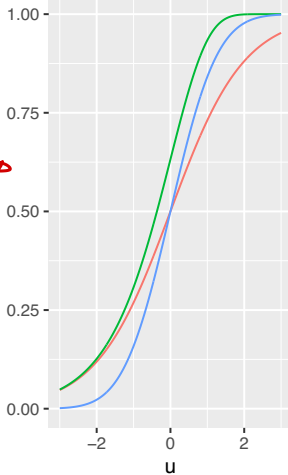
$$F(u) = 1 - \exp(-e^u) \iff g(\pi) = \ln[-\ln(1 - \pi)],$$

but this function is asymmetric.

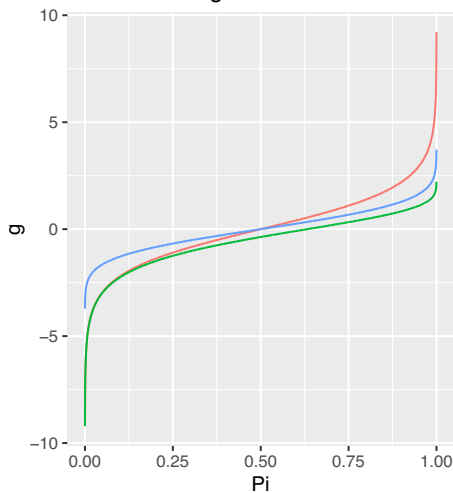
- *Partie linéaire* $g(\pi_i) = \mathbf{x}_i \boldsymbol{\beta}$

Link functions

CDF F



Link function g



1 Logistic regression

- Modeling
- Odds and odds ratio
- Simple logistic regression
- Multiple logistic regression

Odds and odds ratio

- The **odds** for an individual \mathbf{x} is

$$\text{odds}(\mathbf{x}) = \frac{\mathbb{P}(Y = 1|\mathbf{x})}{\mathbb{P}(Y = 0|\mathbf{x})} = \frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})} = \exp[\text{logit}(\pi(\mathbf{x}))]$$

- The **odds ratio** between two individuals \mathbf{x} and $\tilde{\mathbf{x}}$ is defined as the ratio between their odds:

$$\text{OR}(\mathbf{x}, \tilde{\mathbf{x}}) = \frac{\text{odds}(\mathbf{x})}{\text{odds}(\tilde{\mathbf{x}})}.$$

- The **odds ratio** allow to measure the effect of an explanatory variable on the binary response variable.


Odds and odds ratio

The odds ratios can be used in several ways:

- Comparison of success probabilities between two individuals:

$$\begin{cases} \text{OR}(\mathbf{x}, \tilde{\mathbf{x}}) > 1 & \Leftrightarrow & \pi(\mathbf{x}) > \pi(\tilde{\mathbf{x}}) \\ \text{OR}(\mathbf{x}, \tilde{\mathbf{x}}) = 1 & \Leftrightarrow & \pi(\mathbf{x}) = \pi(\tilde{\mathbf{x}}) \\ \text{OR}(\mathbf{x}, \tilde{\mathbf{x}}) < 1 & \Leftrightarrow & \pi(\mathbf{x}) < \pi(\tilde{\mathbf{x}}) \end{cases}$$

- Effect of an explanatory variable:

when $\text{logit}[\pi(\mathbf{x})] = \theta_0 + \theta_1 x^{(1)} + \dots + \theta_p x^{(p)}$ 

$$\text{OR}(\mathbf{x}, \tilde{\mathbf{x}}) = \prod_{j=1}^p \exp \left[\theta_j (x^{(j)} - \tilde{x}^{(j)}) \right].$$

If two individuals only differ on the j -th variable:

$$\text{OR}(\mathbf{x}, \tilde{\mathbf{x}}) = \exp \left[\theta_j (x^{(j)} - \tilde{x}^{(j)}) \right].$$

$$① \text{OR}(x, \tilde{x}) > 1$$

$$\frac{\pi(x)}{1-\pi(x)} > \frac{\pi(\tilde{x})}{1-\pi(\tilde{x})}$$

$$\pi(x) - \cancel{\pi(x)\pi(\tilde{x})} > \pi(\tilde{x}) - \cancel{\pi(x)\pi(\tilde{x})}$$

$$\pi(x) > \pi(\tilde{x})$$

$$② \text{OR}(x, \tilde{x}) = \frac{\text{odds}(x)}{\text{odds}(\tilde{x})}$$

$$\text{odds}(x) = \exp[\text{logit}(\pi(x))]$$

$$= \exp\left[\theta_0 + \sum_{j=1}^p \theta_j x^{(j)}\right]$$

$$\text{odds}(\tilde{x}) = \exp\left[\theta_0 + \sum_{j=1}^p \theta_j \tilde{x}^{(j)}\right]$$

$$\text{OR}(x, \tilde{x}) = \exp\left[\sum_{j=1}^p \theta_j (x^{(j)} - \tilde{x}^{(j)})\right]$$

Si on s'intéresse à θ_j :

$$x^{(k)} = \tilde{x}^{(k)} \quad \forall k \neq j$$

$$\text{OR}(x, \tilde{x}) = \exp\left[\theta_j (x^{(j)} - \tilde{x}^{(j)})\right]$$

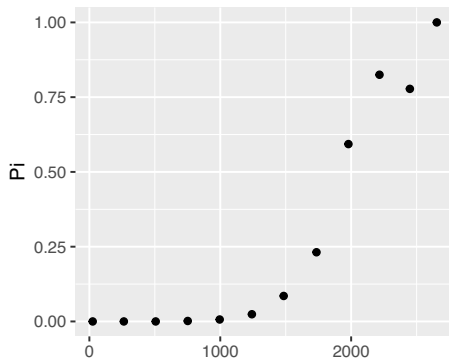
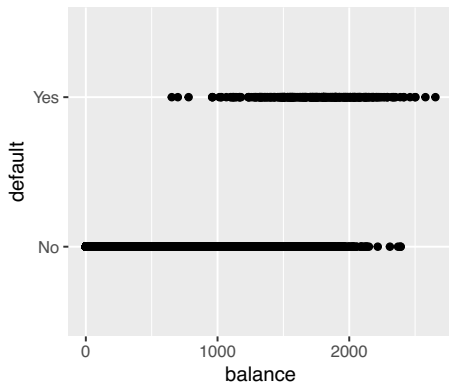
1 Logistic regression

- Modeling
- Odds and odds ratio
- Simple logistic regression
- Multiple logistic regression

with one quantitative explanatory variable

- Goal: explain *default* with the variable *balance*.
- Model:

$$Y_i \sim \mathcal{B}(\overset{\pi_i}{\pi_{\theta}(\mathbf{x}_i)}) \text{ with } \overset{\text{logit}(\pi_i)}{\ln\left(\frac{\pi_{\theta}(\mathbf{x}_i)}{1 - \pi_{\theta}(\mathbf{x}_i)}\right)} = \theta_0 + \theta_1 x_i.$$



Parameter estimation

- The parameter vector $\theta = (\theta_0, \theta_1)'$ is estimated by maximum likelihood.
- The likelihood: $L(\underline{Y}; \theta) = \prod_{i=1}^n \pi_{\theta}(\mathbf{x}_i)^{Y_i} [1 - \pi_{\theta}(\mathbf{x}_i)]^{1-Y_i}$
- The log-likelihood:

$$\begin{aligned} l(\underline{Y}; \theta) &= \sum_{i=1}^n \{ Y_i \ln[\pi_{\theta}(\mathbf{x}_i)] + (1 - Y_i) \ln[1 - \pi_{\theta}(\mathbf{x}_i)] \} \\ &= \sum_{i=1}^n \{ Y_i \ln[F(\theta_0 + \theta_1 x_i)] + (1 - Y_i) \ln[1 - F(\theta_0 + \theta_1 x_i)] \} \end{aligned}$$

$$g(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = x_i \theta$$
$$\pi_i = F(x_i \theta)$$

- System to be solved:

Rem: $\frac{F(u)}{F(u)} = 1 - F(u)$

cf poly
detail
calculs.

$$\begin{cases} \sum_{i=1}^n [Y_i - \pi_{\theta}(\mathbf{x}_i)] = 0 \Leftrightarrow \frac{\partial}{\partial \theta_0} l(\underline{Y}; \theta) = 0 \\ \sum_{i=1}^n x_i [Y_i - \pi_{\theta}(\mathbf{x}_i)] = 0 \Leftrightarrow \frac{\partial}{\partial \theta_1} l(\underline{Y}; \theta) = 0 \end{cases}$$

Parameter estimation

- In order to use a Newton-Raphson or a Fisher-scoring algorithm, the Hessian matrix or the Fisher information matrix.

$E\left[-\frac{\partial^2 l(\underline{Y}; \theta)}{\partial \theta_j \partial \theta_k}\right]$

$$\left\{ \begin{array}{l} \frac{\partial^2 l(\underline{Y}; \theta)}{\partial \theta_0^2} = - \sum_{i=1}^n F(\theta_0 + \theta_1 x_i) [1 - F(\theta_0 + \theta_1 x_i)] = - \sum_{i=1}^n \pi_i (1 - \pi_i) \\ \frac{\partial^2 l(\underline{Y}; \theta)}{\partial \theta_1^2} = - \sum_{i=1}^n x_i^2 F(\theta_0 + \theta_1 x_i) [1 - F(\theta_0 + \theta_1 x_i)] \\ \frac{\partial^2 l(\underline{Y}; \theta)}{\partial \theta_0 \partial \theta_1} = - \sum_{i=1}^n x_i F(\theta_0 + \theta_1 x_i) [1 - F(\theta_0 + \theta_1 x_i)] \end{array} \right.$$

$\mathcal{I}_n(\theta) = \begin{pmatrix} \sum_{i=1}^n \pi_{\theta}(\mathbf{x}_i)(1 - \pi_{\theta}(\mathbf{x}_i)) & \sum_{i=1}^n x_i \pi_{\theta}(\mathbf{x}_i)(1 - \pi_{\theta}(\mathbf{x}_i)) \\ \sum_{i=1}^n x_i \pi_{\theta}(\mathbf{x}_i)(1 - \pi_{\theta}(\mathbf{x}_i)) & \sum_{i=1}^n x_i^2 \pi_{\theta}(\mathbf{x}_i)(1 - \pi_{\theta}(\mathbf{x}_i)) \end{pmatrix} = (X' W X),$

with $W = \text{diag}[\pi_{\theta}(\mathbf{x}_1)(1 - \pi_{\theta}(\mathbf{x}_1)), \dots, \pi_{\theta}(\mathbf{x}_n)(1 - \pi_{\theta}(\mathbf{x}_n))]$.

$\Rightarrow \hat{\theta}_{nv}$

$$I_n(\theta) = \begin{pmatrix} \sum_{i=1}^n \pi_i (1-\pi_i) & \sum_{i=1}^n x_i \pi_i (1-\pi_i) \\ \sum_{i=1}^n x_i \pi_i (1-\pi_i) & \sum_{i=1}^n x_i^2 \pi_i (1-\pi_i) \end{pmatrix}$$

$$X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

$$I_n(\theta) = X' W X$$

$$W = \begin{pmatrix} w_1 & & 0 \\ & \ddots & \\ 0 & & w_n \end{pmatrix}$$

$$\begin{aligned} X' W X &= \begin{pmatrix} 1 & 1 \\ x_1 & x_n \end{pmatrix} \begin{pmatrix} w_1 & 0 \\ & \ddots \\ 0 & w_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \\ &= \begin{pmatrix} \sum_{i=1}^n w_i & \sum_{i=1}^n x_i w_i \\ \sum_{i=1}^n x_i w_i & \sum_{i=1}^n x_i^2 w_i \end{pmatrix} \quad w_i = \pi_i (1-\pi_i) \end{aligned}$$

Example

```
glm.balance <- glm(default ~ balance, data = Default, family = binomial(link = "logit"))
summary(glm.balance)
```

Call:

```
glm(formula = default ~ balance, family = binomial(link = "logit"),
    data = Default)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.2697	-0.1465	-0.0589	-0.0221	3.7589

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.065e+01	3.612e-01	-29.49	<2e-16 ***
balance	5.499e-03	2.204e-04	24.95	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2920.6 on 9999 degrees of freedom
Residual deviance: 1596.5 on 9998 degrees of freedom
AIC: 1600.5

Number of Fisher Scoring iterations: 8

*z-test
test de nullité de param.*

*$\mathcal{D}(n_0)$ (n_0): $g(\pi_i) = \mu$ ✓
 $\mathcal{D}(n)$*

Lo pseudo R^2

$$= 1 - \frac{\mathcal{D}(n)}{\mathcal{D}(n_0)}$$



Example

```
import pandas as pd
import numpy as np
import statsmodels.api as sm
Defaultpy=r.Default
y=Defaultpy["default"].cat.codes
x=Defaultpy["balance"]
x_stat = sm.add_constant(x)
modelbalance = sm.Logit(y, x_stat).fit()
```

Optimization terminated successfully.

Current function value: 0.079823

Iterations 10

```
modelbalance.summary()
```

```
<class 'statsmodels.iolib.summary.Summary'>
"""
```

Logit Regression Results

```
=====
Dep. Variable:          y      No. Observations:      10000
Model:                Logit      Df Residuals:        9998
Method:                MLE       Df Model:            1
Date:                Mar, 22 aoû 2023      Pseudo R-squ.:      0.4534
Time:                09:45:52      Log-Likelihood:     -798.23
converged:                True      LL-Null:           -1460.3
Covariance Type:      nonrobust      LLR p-value:       6.233e-290
=====
```

	coef	std err	z	P> z	[0.025	0.975]
const	-10.6513	0.361	-29.491	0.000	-11.359	-9.943
balance	0.0055	0.000	24.952	0.000	0.005	0.006

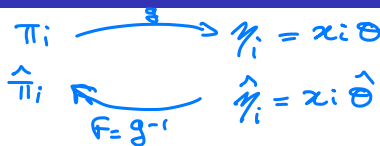
```
=====
```

Prediction

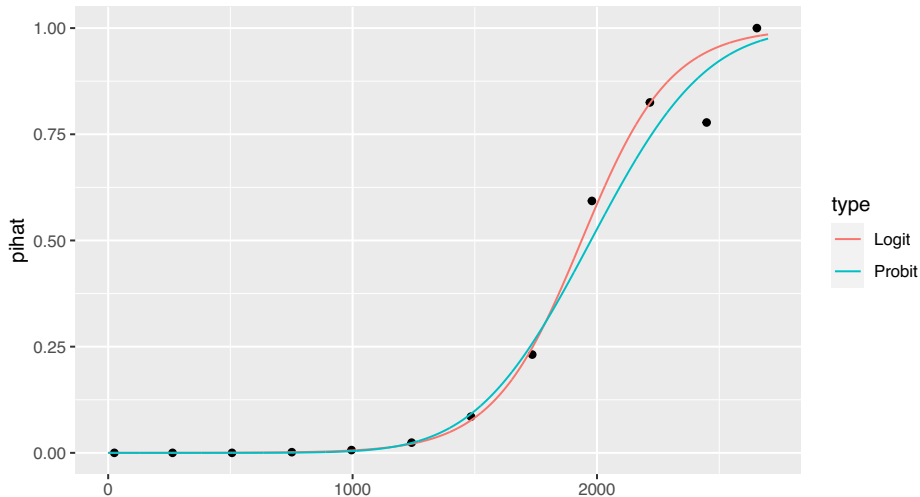
- Linear predictor: $\hat{\eta}_i = \hat{\theta}_0 + x_i \hat{\theta}_1$
- $\hat{\pi}(\mathbf{x}_i) = F(\hat{\eta}_i) = F(\hat{\theta}_0 + \hat{\theta}_1 x_i) = \pi_{\hat{\theta}}(\mathbf{x}_i)$.
- Adjusted values \hat{Y}_i using the Bayes' rule :

$$\hat{Y}_i = \begin{cases} 1 & \text{if } \hat{\pi}(\mathbf{x}_i) > s \\ 0 & \text{otherwise.} \end{cases}$$

- For a new individual $\mathbf{x}_0 = (1, x_0)$, the fitted model allows to predict
 - a proportion (probability) $\hat{\pi}(\mathbf{x}_0) = F(\hat{\theta}_0 + \hat{\theta}_1 x_0)$
 - a predicted response $\hat{Y}_0 = \mathbb{1}_{\hat{\pi}(\mathbf{x}_0) > s}$.
- The threshold by default is $s = 0.5$



Prediction



Confidence interval

$$I_n(\hat{\theta})^{1/2}(\hat{\theta} - \theta) \xrightarrow{n \rightarrow +\infty} \mathcal{N}_2(0_2, I_2)$$

pour n assez grand, $I_n(\hat{\theta})^{1/2}(\hat{\theta} - \theta) \approx \mathcal{N}_2(0_2, I_2)$
 $\hat{\theta} - \theta \approx \mathcal{N}_2(0_2, I_n(\hat{\theta}))$

- Goal: Construct a confidence interval for θ_j $\hat{\theta}_j - \theta_j \approx \mathcal{N}(0, [I_n(\hat{\theta})]_{jj})$
- Methods: Wald's method or method based on the likelihood ratio
- With Wald,

$$IC_{1-\alpha}(\theta_j) = \left[(\hat{\theta}_{ML})_j \pm \underbrace{z_{1-\alpha/2}}_{\uparrow} \sqrt{[I_n(\hat{\theta}_{ML})^{-1}]_{jj}} \right]$$

$1-\alpha/2$ quantile d'une $\mathcal{N}(0,1)$.

$$\mathbb{P}\left(-z \leq \frac{\hat{\theta}_j - \theta_j}{\sqrt{\dots}} \leq z\right) \xrightarrow{n \rightarrow +\infty} \mathbb{P}(-z \leq Z \leq z) = 1-\alpha.$$

$o\ddot{u} \ Z \sim \mathcal{N}(0,1).$

```
# likelihood ratio  
confint(glm.balance)
```

	2.5 %	97.5 %
(Intercept)	-11.383288936	-9.966565064
balance	0.005078926	0.005943365

```
# Wald  
confint.default(glm.balance)
```

	2.5 %	97.5 %
(Intercept)	-11.359186056	-9.943475172
balance	0.005066999	0.005930835

```
ci = modelbalance.conf_int(0.05)  
print(ci)
```

	0	1
const	-11.359208	-9.943453
balance	0.005067	0.005931

Test for the nullity of θ_j (Z-test)

- Hypotheses: $\mathcal{H}_0 : \theta_j = 0$ against $\mathcal{H}_1 : \theta_j \neq 0$
- Based on the result

$$\mathcal{I}_n(\hat{\theta}_{ML})^{1/2}(\hat{\theta}_{ML} - \theta) \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \mathcal{N}(0_2, I_2),$$

we prove that, under \mathcal{H}_0 ,

$$\frac{\hat{\theta}_j}{\hat{\sigma}_j} \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \mathcal{N}(0, 1) \text{ with } \hat{\sigma}_j = \sqrt{[\mathcal{I}(\hat{\theta}_{ML})^{-1}]_{jj}}$$

- Reject zone (asymptotic test of level α):

$$\mathcal{R}_\alpha = \left\{ \left| \hat{\theta}_j / \hat{\sigma}_j \right| > z_{1-\alpha/2} \right\} \xrightarrow[n \rightarrow +\infty]{} \mathbb{P}(|Z| > z) = \alpha.$$

where $z_{1-\alpha/2}$ is the $(1 - \alpha/2)$ -quantile of $\mathcal{N}(0, 1)$. $\tilde{z} \sim \mathcal{N}(0, 1)$.

```
summary(glm.balance)
```

```
Call:
glm(formula = default ~ balance, family = binomial(link = "logit"),
    data = Default)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.2697	-0.1465	-0.0589	-0.0221	3.7589

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.065e+01	3.612e-01	-29.49	<2e-16 ***
balance	5.499e-03	2.204e-04	24.95	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2920.6 on 9999 degrees of freedom
 Residual deviance: 1596.5 on 9998 degrees of freedom
 AIC: 1600.5

Number of Fisher Scoring iterations: 8

Since $p\text{-values} < 2e - 16$, we reject the nullity of both parameters.



```
modelbalance.summary()
```

```
<class 'statsmodels.iolib.summary.Summary'>
"""
```

Logit Regression Results

```
=====
Dep. Variable:          y      No. Observations:          10000
Model:                Logit      Df Residuals:              9998
Method:                MLE      Df Model:                  1
Date:                Mar, 22 aoû 2023      Pseudo R-squ.:          0.4534
Time:                09:45:53      Log-Likelihood:         -798.23
converged:                True      LL-Null:                 -1460.3
Covariance Type:        nonrobust      LLR p-value:           6.233e-290
=====
```

	coef	std err	z	P> z	[0.025	0.975]
const	-10.6513	0.361	-29.491	0.000	-11.359	-9.943
balance	0.0055	0.000	24.952	0.000	0.005	0.006

```
=====
```

Possibly complete quasi-separation: A fraction 0.13 of observations can be perfectly predicted. This might indicate that there is complete quasi-separation. In this case some parameters will not be identified.

```
"""
```

Test for the nullity of θ_j



- It is also possible to consider a submodel testing
- Example for the nullity of θ_1

```
anova(glm(default~1,data=Default,family=binomial(link="logit")), glm.balance, test="Chisq")
```

Analysis of Deviance Table

Model 1: default ~ 1

Model 2: default ~ balance

	Resid.	Df	Resid.	Dev	Df	Deviance	Pr(>Chi)
1	9999			2920.7			
2	9998		1596.5	1	1324.2	< 2.2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

$$T = \mathcal{Q}(n) - \mathcal{Q}(n) \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \chi^2(k-1)$$

```
from scipy.stats import chi2
LR_stat = (-2)* (modelbalance.llnull - modelbalance.llf);
df = 1
pvalue = 1 - chi2(df).cdf(LR_stat);
print(LR_stat)
```

1324.1980279638472

```
print(pvalue)
```

0.0

With a qualitative explanatory variable

- Goal: Explain *default* with the variable *student* (2 levels).
- Possible models:

$$\begin{aligned}\text{logit} [\pi_{\theta}(\mathbf{x}_i)] &= \theta_0 + \theta_1 \mathbb{1}_{x_i=1} + \theta_2 \mathbb{1}_{x_i=0} \\ &= (\theta_0 + \theta_2) + (\theta_1 - \theta_2) \mathbb{1}_{x_i=1} + 0 \mathbb{1}_{x_i=0}.\end{aligned}$$

- \Rightarrow non-identifiable model \Rightarrow constraint on parameters.
- For example, if we assume that $\theta_2 = 0$, the model is

$$\text{logit}[\pi_{\theta}(\mathbf{x}_i)] = \theta_0 + \theta_1 \mathbb{1}_{x_i=1}$$

- We can then use the same reasoning to estimate the parameters, construct confidence intervals, test the nullity of each parameter, ...

- With R, the constraint by default is $\theta_2 = 0$:

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

Call:

```
glm(formula = default ~ student, family = binomial, data = Default)
```

Deviance Residuals:

	Min	1Q	Median	3Q	Max
	-0.2970	-0.2970	-0.2434	-0.2434	2.6585

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-3.50413	0.07071	-49.55	< 2e-16 ***
studentYes	0.40489	0.11502	3.52	0.000431 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2920.6 on 9999 degrees of freedom
 Residual deviance: 2908.7 on 9998 degrees of freedom
 AIC: 2912.7

Number of Fisher Scoring iterations: 6

Example



```
y=Defaultpy["default"].cat.codes
x=Defaultpy["student"].cat.codes
x_stat = sm.add_constant(x)
modelstudent = sm.Logit(y, x_stat).fit();
```

```
Optimization terminated successfully.
      Current function value: 0.145434
      Iterations 7
```

```
modelstudent.summary()
```

```
<class 'statsmodels.iolib.summary.Summary'>
```

```
"""
```

Logit Regression Results

```
=====
Dep. Variable:          y      No. Observations:      10000
Model:                Logit      Df Residuals:         9998
Method:                MLE      Df Model:             1
Date:                Mar, 22 aoû 2023      Pseudo R-squ.:      0.004097
Time:                09:45:54      Log-Likelihood:     -1454.3
converged:                True      LL-Null:           -1460.3
Covariance Type:        nonrobust      LLR p-value:       0.0005416
=====
```

	coef	std err	z	P> z	[0.025	0.975]
const	-3.5041	0.071	-49.554	0.000	-3.643	-3.366
0	0.4049	0.115	3.520	0.000	0.179	0.630

```
"""
```

Example

- Model: $Y_i \sim \mathcal{B}(\pi_\theta(\mathbf{x}_i))$ with

$$\text{logit}[\pi_\theta(\mathbf{x}_i)] = \theta_0 + \theta_1 \mathbb{1}_{\text{student}_i=1}$$

- Odds and odds ratio:

$$\begin{cases} \text{odds}(\text{"student"}) = e^{\theta_0 + \theta_1} = 0.045 \\ \text{odds}(\text{"non-student"}) = e^{\theta_0} = 0.030 \\ \text{OR}(\text{"student"}, \text{"non-student"}) = e^{\theta_1} = 1.5 \end{cases}$$

Thus a student is 1.5 times more likely to be in default than a non-student.

odds("student") = 1.5 odds("non-student")

```
exp(c(glm.student$coefficients, sum(glm.student$coefficients)))
```

```
(Intercept) studentYes  
0.03007299 1.49913321 0.04508342
```

1 Logistic regression

- Modeling
- Odds and odds ratio
- Simple logistic regression
- Multiple logistic regression

- A binary response variable Y
- p regressors $x^{(1)}, \dots, x^{(p)}$
- Example: $p = 3$ regressors
 - 1 qualitative variable ($student = x^{(1)}$)
 - 2 quantitative variables ($balance = x^{(2)}$, $income = x^{(3)}$)

Model without interaction

- Model: $Y_i \sim \mathcal{B}(\pi_\theta(\mathbf{x}_i))$ indep. with

$$\text{logit}[\pi_\theta(\mathbf{x}_i)] = \theta_0 + \theta_1 \mathbb{1}_{x_i^{(1)}=1} + \theta_2 x_i^{(2)} + \theta_3 x_i^{(3)}$$

```
glm.additif<-glm(default~.,data=Default,family=binomial(link="logit"))
summary(glm.additif)
```

Call:

```
glm(formula = default ~ ., family = binomial(link = "logit"),
    data = Default)
```

Deviance Residuals:

	Min	1Q	Median	3Q	Max
	-2.4691	-0.1418	-0.0557	-0.0203	3.7383

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.087e+01	4.923e-01	-22.080	< 2e-16 ***
studentYes	-6.468e-01	2.363e-01	-2.738	0.00619 **
balance	5.737e-03	2.319e-04	24.738	< 2e-16 ***
income	3.033e-06	8.203e-06	0.370	0.71152

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2920.6 on 9999 degrees of freedom
Residual deviance: 1571.5 on 9996 degrees of freedom
AIC: 1579.5

Model without interaction



```
Defaultpy=r.DefaultBIS
y=Defaultpy["default"]
x=Defaultpy[Defaultpy.columns.drop("default")]
x_stat = sm.add_constant(x)
modeladditif = sm.Logit(y, x_stat).fit()
```

```
Optimization terminated successfully.
      Current function value: 0.078577
      Iterations 10
```

```
modeladditif.summary()
```

```
<class 'statsmodels.iolib.summary.Summary'>
"""
```

Logit Regression Results

```
=====
Dep. Variable:          default    No. Observations:          10000
Model:                  Logit      Df Residuals:              9996
Method:                  MLE        Df Model:                  3
Date:                   Mar, 22 aoû 2023    Pseudo R-squ.:          0.4619
Time:                   09:45:54    Log-Likelihood:         -785.77
converged:              True        LL-Null:                 -1460.3
Covariance Type:        nonrobust    LLR p-value:             3.257e-292
=====
```

	coef	std err	z	P> z	[0.025	0.975]
const	-10.8690	0.492	-22.079	0.000	-11.834	-9.904
student	-0.6468	0.236	-2.738	0.006	-1.110	-0.184
balance	0.0057	0.000	24.737	0.000	0.005	0.006
income	3.033e-06	8.2e-06	0.370	0.712	-1.3e-05	1.91e-05

```
=====
```

Possibly complete quasi-separation: A fraction 0.15 of observations can be

Test of nullity of θ_j

- Test of nullity \implies Z-test

In our example, $p\text{-value}=0.71152$ for the nullity of $\theta_3 \implies$ we can remove the variable *income* from the model

- We can reach the same conclusion by a sub-model test

```
glm.sansincome<-glm(default~student+balance,data=Default,family=binomial(link="logit"))
anova(glm.sansincome,glm.additif,test="Chisq")
```

Analysis of Deviance Table

Model 1: default ~ student + balance

Model 2: default ~ student + balance + income

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	9997	1571.7			
2	9996	1571.5	1	0.13677	0.7115

Test the nullity of θ_2 and θ_3 simul.

- We want to test the nullity of θ_2 and θ_3 simultaneously
- Method 1: sub-model testing

- (M_0) : $\text{logit}[\pi_\theta(\mathbf{x}_i)] = \theta_0 + \theta_1 \mathbb{1}_{x_i^{(1)}=1}$

$$k_0 = 2$$

- (M_1) : $\text{logit}[\pi_\theta(\mathbf{x}_i)] = \theta_0 + \theta_1 \mathbb{1}_{x_i^{(1)}=1} + \theta_2 x_i^{(2)} + \theta_3 x_i^{(3)}$

$$k = 4$$

$$T = \mathcal{D}(M_0) - \mathcal{D}(M_1) \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \chi^2(4 - 2)$$

```
anova(glm.student,glm.additif,test="Chisq")
```

Analysis of Deviance Table

Model 1: default ~ student

Model 2: default ~ student + balance + income

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	9998	2908.7			
2	9996	1571.5	2	1337.1	< 2.2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Test the nullity of θ_2 and θ_3 simul.

- We want to test the nullity of θ_2 and θ_3 simultaneously
- Method 2: Wald's test

$$\mathcal{H}_0 : C\theta = 0_2 \text{ against } \mathcal{H}_1 : C\theta \neq 0_2 \text{ with } C = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Under \mathcal{H}_0 ,

$$T = (C\hat{\theta}_{ML})' \left[C\mathcal{I}_n(\hat{\theta}_{ML})^{-1} C' \right]^{-1} (C\hat{\theta}_{ML}) \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \chi^2(2).$$

```
hattheta <- glm.additif$coefficients
hatpi <- glm.additif$fitted.values
W <- diag(hatpi*(1-hatpi))
X <- cbind(rep(1,10000),student,balance,income)
In <- t(X) %*% W %*% X
C <- matrix(c(0,0,1,0,0,0,0,1),nrow=4)
t(t(C)%*%hattheta) %*% solve(t(C)%*%solve(In)%*%C) %*% (t(C)%*%hattheta) > qchisq(0.95,df=2)
```

```
 [,1]
[1,] TRUE
```

Variable selection

- More generally, a variable selection procedure can be implemented
- For instance, we can implement a backward selection procedure based on the AIC criterion

```
step.backward <- step(glm.additif)
```

```
Start:  AIC=1579.54  
default ~ 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 ~ student + balance
```

		Df	Deviance	AIC
<none>		1571.7	1577.7	
- student	1	1596.5	1600.5	
- balance	1	2908.7	2912.7	

Variable selection

- We can use *stepAIC* (*MASS* library) with AIC (option “ $p=2$ ”) or BIC (option “ $p=\log(n)$ ”)

```
library(MASS)
stepAIC(glm.additif, direction=c("backward"),p=2,trace=0) # AIC
```

```
Call: glm(formula = default ~ student + balance, family = binomial(link = "logit"),
  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

```
stepAIC(glm.additif, direction=c("backward"),p=log(nrow(Default))) # BIC
```

Start: AIC=1579.54

default ~ 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 ~ student + balance

Model with interaction

- Full model with all interactions (order 2):

$$\begin{aligned}\text{logit}[\pi_{\theta}(\mathbf{x}_i)] &= \theta_0 + \theta_2 x_i^{(2)} + \theta_3 x_i^{(3)} + \theta_{23} x_i^{(2)} x_i^{(3)} \\ &\quad + (\beta_1 + \beta_2 x_i^{(2)} + \beta_3 x_i^{(3)}) \mathbb{1}_{x_i^{(1)}=1}\end{aligned}$$

- Then, use a variable selection procedure to simplify the model and valid with a testing procedure

Model with interaction

```
glm.full<-glm(default~.^2,data=Default,family="binomial")
summary(glm.full)
```

Call:

```
glm(formula = default ~ .^2, family = "binomial", data = Default)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.4848	-0.1417	-0.0554	-0.0202	3.7579

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.104e+01	1.866e+00	-5.914	3.33e-09 ***
studentYes	-5.201e-01	1.344e+00	-0.387	0.699
balance	5.882e-03	1.180e-03	4.983	6.27e-07 ***
income	4.050e-06	4.459e-05	0.091	0.928
studentYes:balance	-2.551e-04	7.905e-04	-0.323	0.747
studentYes:income	1.447e-05	2.779e-05	0.521	0.602
balance:income	-1.579e-09	2.815e-08	-0.056	0.955

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2920.6 on 9999 degrees of freedom
Residual deviance: 1571.1 on 9993 degrees of freedom
AIC: 1585.1

Number of Fisher Scoring iterations: 8

Model with interaction

```
stepAIC(glm.full, direction=c("backward"),p=log(nrow(Default)),trace=0)
```

```
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

```
anova(glm.sansincome,glm.full)
```

Analysis of Deviance Table

Model 1: default ~ student + balance

Model 2: default ~ (student + balance + income)^2

	Resid. Df	Resid. Dev	Df	Deviance
1	9997	1571.7		
2	9993	1571.1	4	0.61588

0.61588
↑

on ne rejette pas H_0 au risque 5%

⇒ on ne conserve que les variables Student & Bal.

$$T = \mathcal{D}(n_1) - \mathcal{D}(n_2) \xrightarrow[n \rightarrow \infty]{L} \chi^2(4)$$

Focus on “default ~ student+balance”

- Model: $Y_i \sim \mathcal{B}(\pi_{\theta}(\mathbf{x}_i))$ indep. with

$$\text{logit}[\pi_{\theta}(\mathbf{x}_i)] = \theta_0 + \theta_1 \mathbb{1}_{x_i^{(1)}=1} + \theta_2 x_i^{(2)} = X_i \theta \text{ where } X_i = (1, \mathbb{1}_{x_i^{(1)}=1}, x_i^{(2)})$$

```
glm.final = glm(default ~ student + balance, data=Default, family=binomial)
summary(glm.final)
```

Call:

```
glm(formula = default ~ student + balance, family = binomial,
    data = Default)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.4578	-0.1422	-0.0559	-0.0203	3.7435

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.075e+01	3.692e-01	-29.116	< 2e-16 ***
studentYes	-7.149e-01	1.475e-01	-4.846	1.26e-06 ***
balance	5.738e-03	2.318e-04	24.750	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2920.6 on 9999 degrees of freedom
Residual deviance: 1571.7 on 9997 degrees of freedom
AIC: 1577.7

Number of Fisher Scoring iterations: 8



Focus on “default ~ student+balance”

- Probability prediction: $= F(\hat{\eta}_i)$

$$\hat{\pi}_i = \frac{e^{\hat{\eta}_i}}{1 + e^{\hat{\eta}_i}} \text{ with } \hat{\eta}_i = X_i \hat{\theta}_{ML}$$

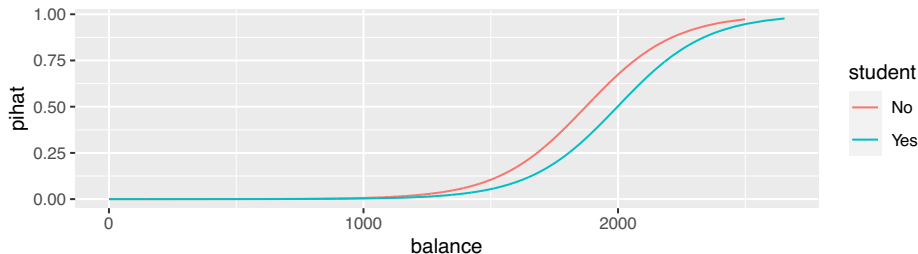
- Fitted values $\hat{Y}_i = \mathbb{1}_{\hat{\pi}_i > 0.5} = \hat{y}_i$

```
hatpi <- glm.final$fitted.values  
table(default, hatpi > 0.5)
```

\hat{y}_i

y_i

	default	FALSE	TRUE
No	9628	39	
Yes	228	105	



Parameter interpretation

- 2 individuals with the same balance value (bal), one student and one non-student $\mathbf{x} = (1, 1, bal)$ and $\tilde{\mathbf{x}} = (1, 0, bal)$ then

$$OR(\mathbf{x}, \tilde{\mathbf{x}}) = \frac{e^{\theta_0 + \theta_1 + \theta_2 bal}}{e^{\theta_0 + \theta_2 bal}} = e^{\theta_1}$$

thus a student is $e^{-0.7149} = 0.489$ times likely to be in default than a non-student for the same value of *balance*.

- 2 individuals (both student or non-student) and $balance(\mathbf{x}) = balance(\tilde{\mathbf{x}}) + 1$

$$OR(\mathbf{x}, \tilde{\mathbf{x}}) = e^{\theta_2 \overbrace{(balance(\mathbf{x}) - balance(\tilde{\mathbf{x}}))}^{= 1}}$$

thus when we increase the balance variable by one unit, the chance of being at default is multiplied by $e^{0.005738} = 1.005$.

1 Logistic regression

2 Log regression

2 Log regression

- Description of the example
- Log regression with one regressor
- Multiple log regression

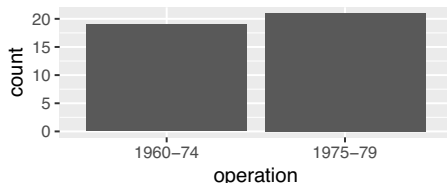
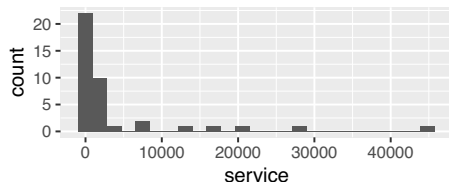
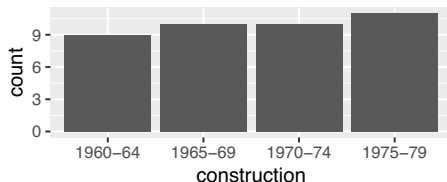
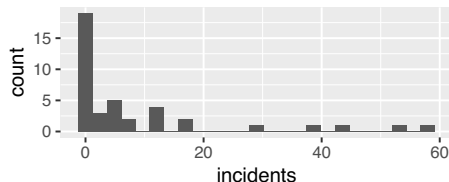
Example : Number of maritime accidents

- We are interested in the variable *incidents* = number of damage incidents per month of commissioning of a ship
- The data frame contains 40 observations on 5 ship types in 4 vintages and 2 service periods:
 - *type*: factor with levels “A” to “E” for the different ship types
 - *construction*: factor with levels “1960-64”, “1965-69”, “1970-74”, “1975-79” for the periods of construction
 - *operation*: factor with levels “1960-74”, “1975-79” for the periods of operation
 - *service*: aggregate months of service

Example

```
data("ShipAccidents")  
#str(ShipAccidents)  
summary(ShipAccidents)
```

	construction	operation	service	incidents
A:8	1960-64: 9	1960-74:19	Min. : 0.0	Min. : 0.0
B:8	1965-69:10	1975-79:21	1st Qu.: 175.8	1st Qu.: 0.0
C:8	1970-74:10		Median : 782.0	Median : 2.0
D:8	1975-79:11		Mean : 4089.3	Mean : 8.9
E:8			3rd Qu.: 2078.5	3rd Qu.:11.0
			Max. :44882.0	Max. :58.0



2 Log regression

- Description of the example
- Log regression with one regressor
- Multiple log regression

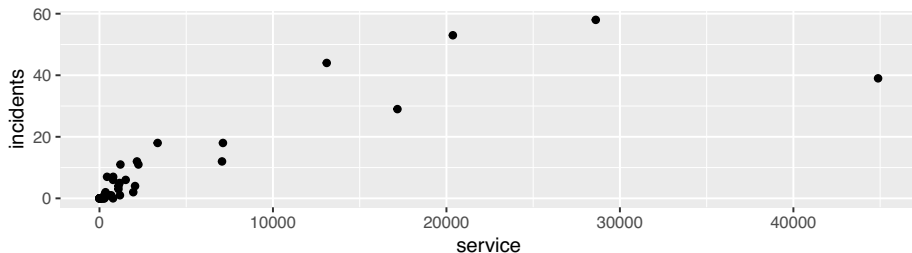
With a quantitative explanatory variable

- Goal: explain the response variable *incidents* (Y) with the quantitative variable *service* (x)

- Model:

$$\begin{cases} Y_i \sim \mathcal{P}(\lambda(\mathbf{x}_i)), \forall i = 1, \dots, n \\ \ln[\lambda(\mathbf{x}_i)] = \theta_0 + \theta_1 x_i \\ Y_1, \dots, Y_n \text{ independent} \end{cases}$$

Diagram annotations: A blue arrow labeled λ_i points to the $\lambda(\mathbf{x}_i)$ term in the first equation. A blue double arrow points from $\lambda(\mathbf{x}_i)$ to the $\ln[\lambda(\mathbf{x}_i)]$ term in the second equation. A pink circle highlights the $\ln[\lambda(\mathbf{x}_i)]$ term, and a green circle highlights the $\theta_0 + \theta_1 x_i$ term.



```
fit.service <- glm(incidents ~ service, data=ShipAccidents, family=poisson)
summary(fit.service)
```

Call:

```
glm(formula = incidents ~ service, family = poisson, data = ShipAccidents)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-6.0040	-3.1674	-2.0055	0.9155	7.2372

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	1.613e+00	7.150e-02	22.55	<2e-16 ***
service	6.417e-05	2.870e-06	22.36	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 730.25 on 39 degrees of freedom
 Residual deviance: 374.55 on 38 degrees of freedom
 AIC: 476.41

Number of Fisher Scoring iterations: 6

$$\begin{aligned} & \text{---} \mathcal{D}(n) \approx n_0 : \ln(x) = \mu_0 \\ & \text{---} \mathcal{D}(n) \end{aligned}$$

Example



```
import pandas as pd
import numpy as np
import statsmodels.api as sm
from statsmodels.formula.api import glm

Accidpy=r.ShipAccidents
fitervicepy=glm('incidents~service',data=Accidpy,family=sm.families.Poisson()).fit()
print(fitervicepy.summary())
```

Generalized Linear Model Regression Results

```
=====
Dep. Variable:          incidents   No. Observations:          40
Model:                  GLM        Df Residuals:              38
Model Family:           Poisson    Df Model:                  1
Link Function:           Log       Scale:                    1.0000
Method:                 IRLS       Log-Likelihood:         -236.21
Date:                   Mar, 22 aoû 2023   Deviance:              374.55
Time:                   09:45:58    Pearson chi2:          368.
No. Iterations:         6           Pseudo R-squ. (CS):     0.9999
Covariance Type:        nonrobust
=====
```

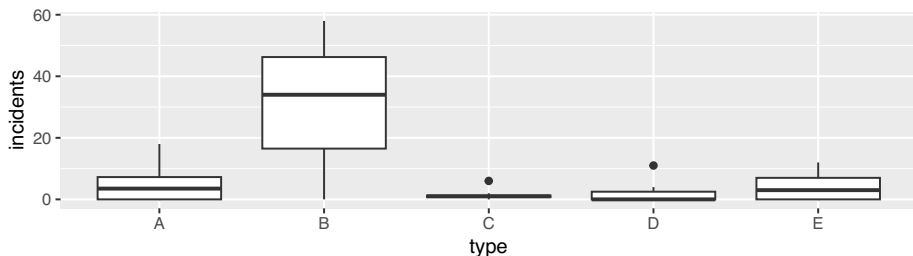
	coef	std err	z	P> z	[0.025	0.975]
Intercept	1.6127	0.072	22.555	0.000	1.473	1.753
service	6.417e-05	2.87e-06	22.356	0.000	5.85e-05	6.98e-05

```
=====
```

With a **qualitative** explanatory variable

- Goal: Explain *incidents* with the qualitative variable *type* having 5 levels.
- To make the model identifiable, we must choose a reference level (here, *type* = A).
- Model :

$$\begin{cases} Y_i \sim \mathcal{P}(\lambda(\mathbf{x}_i)), \forall i = 1, \dots, n \\ \ln[\lambda(\mathbf{x}_i)] = \theta_0 + \theta_1 \mathbb{1}_{\text{type}_i=B} + \theta_2 \mathbb{1}_{\text{type}_i=C} + \theta_3 \mathbb{1}_{\text{type}_i=D} + \theta_4 \mathbb{1}_{\text{type}_i=E} \\ Y_1, \dots, Y_n \text{ independent} \end{cases}$$



```
fit.type <- glm(incidents ~ type, data=ShipAccidents, family=poisson)
summary(fit.type)
```

Call:

```
glm(formula = incidents ~ type, family = poisson, data = ShipAccidents)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-7.9530	-2.0616	-0.4541	1.2873	4.3425

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	1.6582	0.1543	10.747	< 2e-16 ***
typeB	1.7957	0.1666	10.777	< 2e-16 ***
typeC	-1.2528	0.3273	-3.827	0.00013 ***
typeD	-0.9045	0.2875	-3.146	0.00165 **
typeE	-0.2719	0.2346	-1.159	0.24650

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 730.25 on 39 degrees of freedom
 Residual deviance: 275.65 on 35 degrees of freedom
 AIC: 383.52

Number of Fisher Scoring iterations: 6



```
fittypepy=glm('incidents~C(type)',data=Accidpy,family=sm.families.Poisson()).fit()
print(fittypepy.summary())
```

Generalized Linear Model Regression Results

Dep. Variable:	incidents	No. Observations:	40
Model:	GLM	Df Residuals:	35
Model Family:	Poisson	Df Model:	4
Link Function:	Log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-186.76
Date:	Mar, 22 août 2023	Deviance:	275.65
Time:	09:45:58	Pearson chi2:	249.
No. Iterations:	5	Pseudo R-squ. (CS):	1.000
Covariance Type:	nonrobust		

	coef	std err	z	P> z	[0.025	0.975]
Intercept	1.6582	0.154	10.747	0.000	1.356	1.961
C(type) [T.B]	1.7957	0.167	10.777	0.000	1.469	2.122
C(type) [T.C]	-1.2528	0.327	-3.827	0.000	-1.894	-0.611
C(type) [T.D]	-0.9045	0.287	-3.146	0.002	-1.468	-0.341
C(type) [T.E]	-0.2719	0.235	-1.159	0.246	-0.732	0.188

Effect of variable *type*

- Since the variable *type* has 5 levels, a sub-model test is used to test the effect of this variable:
 - (M_0) : $\ln[\lambda(\mathbf{x}_i)] = \theta_0$
 - (M_1) : $\ln[\lambda(\mathbf{x}_i)] = \theta_0 + \theta_1 \mathbb{1}_{\text{type}_i=\text{B}} + \theta_2 \mathbb{1}_{\text{type}_i=\text{C}} + \theta_3 \mathbb{1}_{\text{type}_i=\text{D}} + \theta_4 \mathbb{1}_{\text{type}_i=\text{E}}$
- Test's statistics:

Sous H_0 $T = \mathcal{D}(M_0) - \mathcal{D}(M_1) \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \chi^2(5-1)$

k k-1

- Reject zone: $\mathcal{R}_\alpha = \{T > v_{1-\alpha,4}\}$ where $v_{1-\alpha,4}$ is the $(1 - \alpha)$ quantile of $\chi^2(4)$.
 $\mathbb{P}_{H_0}(T > v) \xrightarrow[n \rightarrow +\infty]{} \mathbb{P}(\chi^2(4) > v) = \alpha$
- P-value: $pval = \mathbb{P}_{H_0}(T > T^{obs}) \xrightarrow[n \rightarrow +\infty]{} \mathbb{P}(\chi^2(4) > T^{obs})$

Effect of variable *type*



No

Yes

```
anova(glm(incidents ~ 1, data=ShipAccidents, family=poisson), fit.type, test="Chisq")
```

Analysis of Deviance Table

Model 1: incidents ~ 1

Model 2: incidents ~ type

	Resid.	Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	39		730.25			
2	35		275.65	4	454.6	< 2.2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

*→ on rejette H_0 au risque 5%
⇒ effet de la variable type.*

```
from scipy.stats import chi2
LR_stat=(-2)*(fittypepy.llnull - fittypepy.llf);
pvalue=1-chi2(4).cdf(LR_stat);
print(LR_stat)
```

454.60255359036773

```
print(pvalue)
```

0.0

2 Log regression

- Description of the example
- Log regression with one regressor
- Multiple log regression

Multiple log regression

- Goal: Explain the response variable *incidents* with all the available explanatory variables.
- Since a model with interactions (2nd order) has 37 parameters and the sample size is $n = 40$, we only consider an additive log regression model here.
- Model: $Y_i \sim \mathcal{P}(\lambda(\mathbf{x})_i)$ with

les Y_i et

$$\ln[\lambda(\mathbf{x}_i)] = \theta_0 + \alpha_1 \mathbb{1}_{\text{type}_i=\text{B}} + \alpha_2 \mathbb{1}_{\text{type}_i=\text{C}} + \alpha_3 \mathbb{1}_{\text{type}_i=\text{D}} + \alpha_4 \mathbb{1}_{\text{type}_i=\text{E}} \\ + \beta_1 \mathbb{1}_{\text{const}_i="65-69"} + \beta_2 \mathbb{1}_{\text{const}_i="70-74"} + \beta_3 \mathbb{1}_{\text{const}_i="75-79"} \\ + \gamma_1 \mathbb{1}_{\text{op}_i="75-79"} + \theta_1 \text{service}_i$$

= $X_i \theta$

Example

```
fit.add <- glm(incidents ~ . , data=ShipAccidents, family=poisson)
summary(fit.add)
```

Call:

```
glm(formula = incidents ~ . , family = poisson, data = ShipAccidents)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.5810	-1.4773	-0.8972	0.5952	3.2154

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	5.492e-04	2.787e-01	0.002	0.998427
typeB	5.933e-01	2.163e-01	2.743	0.006092 **
typeC	-1.190e+00	3.275e-01	-3.635	0.000278 ***
typeD	-8.210e-01	2.877e-01	-2.854	0.004321 **
typeE	-2.900e-01	2.351e-01	-1.233	0.217466
construction1965-69	1.148e+00	1.793e-01	6.403	1.53e-10 ***
construction1970-74	1.596e+00	2.242e-01	7.122	1.06e-12 ***
construction1975-79	5.670e-01	2.809e-01	2.018	0.043557 *
operation1975-79	8.619e-01	1.317e-01	6.546	5.92e-11 ***
service	7.270e-05	8.488e-06	8.565	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 730.253 on 39 degrees of freedom
Residual deviance: 99.793 on 30 degrees of freedom
AIC: 217.66

Number of Fisher Scoring iterations: 5



```
fitaddpy =glm('incidents~C(type)+C(construction)+C(operation)+service',
              data=Accidpy,family=sm.families.Poisson()).fit()
print(fitaddpy.summary())
```

Generalized Linear Model Regression Results

```
=====
Dep. Variable:          incidents    No. Observations:          40
Model:                  GLM          Df Residuals:              30
Model Family:           Poisson     Df Model:                  9
Link Function:          Log         Scale:                   1.0000
Method:                 IRLS        Log-Likelihood:         -98.830
Date:                   Mar, 22 aoû 2023    Deviance:              99.793
Time:                   09:45:59    Pearson chi2:          90.0
No. Iterations:         6           Pseudo R-squ. (CS):     1.000
Covariance Type:        nonrobust
=====
```

	coef	std err	z	P> z	[0.025	0.975]
Intercept	0.0005	0.279	0.002	0.998	-0.546	0.547
C(type) [T.B]	0.5933	0.216	2.743	0.006	0.169	1.017
C(type) [T.C]	-1.1903	0.327	-3.635	0.000	-1.832	-0.548
C(type) [T.D]	-0.8210	0.288	-2.854	0.004	-1.385	-0.257
C(type) [T.E]	-0.2900	0.235	-1.233	0.217	-0.751	0.171
C(construction) [T.1965-69]	1.1479	0.179	6.403	0.000	0.796	1.499
C(construction) [T.1970-74]	1.5965	0.224	7.122	0.000	1.157	2.036
C(construction) [T.1975-79]	0.5670	0.281	2.018	0.044	0.016	1.118
C(operation) [T.1975-79]	0.8619	0.132	6.546	0.000	0.604	1.120
service	7.27e-05	8.49e-06	8.565	0.000	5.61e-05	8.93e-05

Variable selection

- It is possible to implement a variable selection procedure using `step(fit.add)` (backward procedure with AIC criterion)

```
step(fit.add, trace=1)
```

Start: AIC=217.66

incidents ~ type + construction + operation + service

	Df	Deviance	AIC
<none>		99.793	217.66
- type	4	148.053	257.92
- operation	1	147.687	263.55
- service	1	182.605	298.47
- construction	3	191.419	303.29

Call: `glm(formula = incidents ~ type + construction + operation + service, family = poisson, data = ShipAccidents)`

Coefficients:

(Intercept)	typeB	typeC
0.0005492	0.5932730	-1.1903189
typeD	typeE	construction1965-69
-0.8210370	-0.2899922	1.1478796
construction1970-74	construction1975-79	operation1975-79
1.5964752	0.5669790	0.8618750
service		
0.0000727		

Degrees of Freedom: 39 Total (i.e. Null); 30 Residual

Null Deviance: 730.3

Test of sub-models

- We can for instance test the sub-model without the variables *contructions* and *operation*

```
fit.ssmod <- glm(incidents ~ type +service,data=ShipAccidents,family=poisson)
anova(fit.ssmod, fit.add, test="Chisq")
```

Analysis of Deviance Table

Model 1: incidents ~ type + service

Model 2: incidents ~ type + construction + operation + service

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	34	230.832			
2	30	99.793	4	131.04	< 2.2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```
fitssmodpy = glm('incidents~C(type)+service',data=Accidpy,family=sm.families.Poisson()).fit()
LR_stat=(fitssmodpy.deviance - fitaddpy.deviance)
print(LR_stat)
```

131.0387423849605

```
print(1-chi2(4).cdf(LR_stat))
```

0.0

Prediction

- If we want to predict the average number of incidents for a ship with $\text{type}="A"$, $\text{construction}="65-69"$, $\text{operation}="60-74"$, $\text{service}=1000$
 $x_0 \hat{\theta} \xrightarrow{P_1} \hat{\lambda}_0$ $\hat{y}_0 = \text{avg}_{k \in \mathcal{K}} \mathbb{P}(U=k) \bar{u} \propto \mathbb{P}(\hat{\lambda}_0^{\text{obs}})$.
 $\hat{\lambda}_0 = e^{X_0 \hat{\theta}_{ML}}$ with $X_0 = (1, \underbrace{0, 0, 0, 0}_{\text{type}}, \underbrace{1, 0, 0}_{\text{construction}}, 0, 1000)$

```
new.data = data.frame(type=factor("A"), construction=factor("1965-69"),operation=factor("1960-74"), service = 1000)
lambda_hat = exp(predict(fit.add,new.data))
lambda_hat
```

```
1
3.391016
```

- Prediction of some probabilities: Let $A \sim \mathcal{P}(\hat{\lambda}_0)$. For instance,
 - ship has no incident: $\mathbb{P}(A = 0) = e^{-\hat{\lambda}_0}$
 - ship has at most one incident: $\mathbb{P}(A \leq 1) = (1 + \hat{\lambda}_0)e^{-\hat{\lambda}_0}$

```
c(exp(-lambda_hat), (1+lambda_hat) * exp(-lambda_hat))
```

```
1      1
0.03367446 0.14786507
```

$$Y_0 \sim P(\lambda_0)$$

$$P_n(\lambda_0) = X_0 \Theta.$$

$$(Y_1, \dots, Y_n) \longrightarrow \hat{\Theta}$$

$$\hat{\lambda}_0 = \exp[X_0 \hat{\Theta}]$$

$$\hat{\lambda}_0^{\text{obs}} = \exp[X_0 \hat{\Theta}^{\text{obs}}]$$

$$\hat{Y}_0^{\text{obs}} = \underset{k \in \mathbb{N}}{\text{argmax}} \mathbb{P}(U=k) \text{ s.t. } U \sim P(\hat{\lambda}_0^{\text{obs}})$$

$$\mathbb{P}(U=k) = e^{-\hat{\lambda}_0^{\text{obs}}} \frac{(\hat{\lambda}_0^{\text{obs}})^k}{k!}$$

- [1] Jean-Marc Azais and Jean-Marc Bardet. *Le modèle linéaire par l'exemple-2e éd.: Régression, analyse de la variance et plans d'expérience illustrés avec R et SAS*. Dunod, 2012.
- [2] Jean-Jacques Daudin. *Le modèle linéaire et ses extensions-Modèle linéaire général, modèle linéaire généralisé, modèle mixte, plans d'expériences (Niveau C)*. 2015.
- [3] Peter McCullagh. *Generalized linear models*. Routledge, 2018.
- [4] Nalini Ravishanker, Zhiyi Chi, and Dipak K Dey. *A first course in linear model theory*. CRC Press, 2021.
- [5] Alvin C Rencher and G Bruce Schaalje. *Linear models in statistics*. John Wiley & Sons, 2008.