Procédure Logistique

Plan de la présentation

- Régression Logistique
  - Qu'est-ce que c'est?
  - Exemples
- Proc Logistic
  - Interprétation
  - Options disponibles
  - Problèmes courants

Régression

- Expliquer une variable (variable expliquée) à l'aide de variables explicatives.
- Ex: Vente de crème glacée en fonction de la température moyenne
- Objectifs:
  - Prédiction
  - Étude des Liens

Régression Logistique

- Variable expliquée binaire
  - oui/non
  - malade/pas malade
  - etc.
- Variables explicatives
  - Continue (Dette, poids, etc.)
  - Catégorielle (Sexe, groupe sanguin, etc.)
- Concrètement, la régression logistique permet de modéliser la probabilité que la variable expliquée soit un succès à l'aide de variables explicatives.

Modèle

- On pose $x_1, x_2, \ldots$ comme étant nos variables explicatives et y comme étant notre variable expliquée.
- $\mathrm{P}(y \text{ soit un succès}) = \pi$
- $\pi = \frac{1}{1+e^{-\alpha-\beta_1 x_1-\beta_2 x_2-\cdots}}$
- $0 < \pi < 1$
- $\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots$

Exemple

- https://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_logistic_sect060.htm
- Variable expliqué: *Pain* (Yes/No)
- Variables explicatives:
  - *Treatment* (A/B/P);
  - *Sex* (M/F);
  - *Age*;
  - *Duration*;

Exemple

	Treatment	Sex	Age	Duration	Pain
1	P	F	68	1	No
2	B	M	74	16	No
3	P	F	67	30	No
4	P	M	66	26	Yes
5	B	F	67	28	No
6	B	F	77	16	No
7	A	F	71	12	No
8	B	F	72	50	No
9	B	F	76	9	Yes
10	A	M	71	17	Yes
11	A	F	63	27	No
12	A	F	69	18	Yes
13	B	F	66	12	No
14	A	M	62	42	No

```
proc logistic data=Neuralgia;
    class Treatment Sex;
    model Pain (event = 'YES') = Treatment Sex Age Duration;
run;
```

Exemple

- $\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta_1 T_A + \beta_2 T_B + \beta_3 S_M + \beta_4 Age + \beta_5 Duration$
- $T_A = \begin{cases} 1 \text{ lorsque } Treatment = A \\ 0 \text{ lorsque } Treatment \neq A \end{cases}$
- *Remarque:* Pour une variable catégorielle ayant n catégories, on a besoin de n-1 variables dichotomiques.

Variables Catégorielles

```
proc logistic data=Neuralgia;
   class Treatment Sex;
   model Pain (event = 'YES') = Treatment Sex Age Duration;
run;
```

Class Level Information			
Class	Value	Design Variables	
Treatment	A	1	0
	B	0	1
	P	-1	-1
Sex	F	1	
	M	-1	

$$T_A = \begin{cases} 1 \text{ lorsque } Treatment = A \\ 0 \text{ lorsque } Treatment = B \\ -1 \text{ lorsque } Treatment = P \end{cases}$$

```
proc logistic data=Neuralgia;
   class Treatment Sex(DESC) /PARAM = REF;
   model Pain (event = 'YES') = Treatment Sex Age Duration;
run;
```

Class Level Information			
Class	Value	Design Variables	
Treatment	A	1	0
	B	0	1
	P	0	0
Sex	M	1	
	F	0	

$$T_A = \begin{cases} 1 \text{ lorsque } Treatment = A \\ 0 \text{ lorsque } Treatment \neq A \end{cases}$$

Coefficients

- $\pi = \frac{1}{1+e^{-\alpha-\beta_1 x_1-\beta_2 x_2-\cdots}}$
- Propriétés:
  - $\beta_1 = 0 \rightarrow x_1$ n'a pas d'effet sur la probabilité estimée
  - $\beta_1 > 0 \rightarrow$plus la valeur de x_1 est élevée, plus la probabilité estimée est élevée
  - $\beta_1 < 0 \rightarrow$plus la valeur de x_1 est élevée, moins la probabilité estimée est élevée

Coefficients

- $\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta_1 T_A + \beta_2 T_B + \beta_3 S_F + \beta_4 Age + \beta_5 Duration$

	Analysis of Maximum Likelihood Estimates						
	Parameter		DDL	Estimate	Standard Error	Wald Chi-Square	Pr > Khi-2
α	Intercept		1	-17.4066	6.6914	6.7669	0.0093
β_1	Treatment	A	1	-3.1817	1.0161	9.8049	0.0017
β_2	Treatment	B	1	-3.7085	1.1407	10.5700	0.0011
β_3	Sex	M	1	1.8322	0.7963	5.2946	0.0214
β_4	Age		1	0.2621	0.0970	7.2977	0.0069
β_5	Duration		1	-0.00586	0.0330	0.0315	0.8591

Matrice de confusion

		Malade	Santé
Estimation	Malade	Vrai Positif	Faux Positif
	Santé	Faux Négatif	Vrai Négatif
		Groupe Positif	Groupe Négatif

- Sensibilité = Vrai Positif / Groupe Positif
  - La proportion de gens identifiés comme étant malade parmi les malades.
  - Les modèles très sensible sont surtout utiles pour s'assurer que la maladie n'est pas présente (peu de faux négatifs)
- Spécificité = Vrai Négatif / Groupe Négatif
  - La proportion de gens identifiés comme étant en santé parmi les gens en santé.
  - Les modèles très spécifiques sont utiles pour s'assurer qu'une maladie est bien présente (peu de faux positifs).

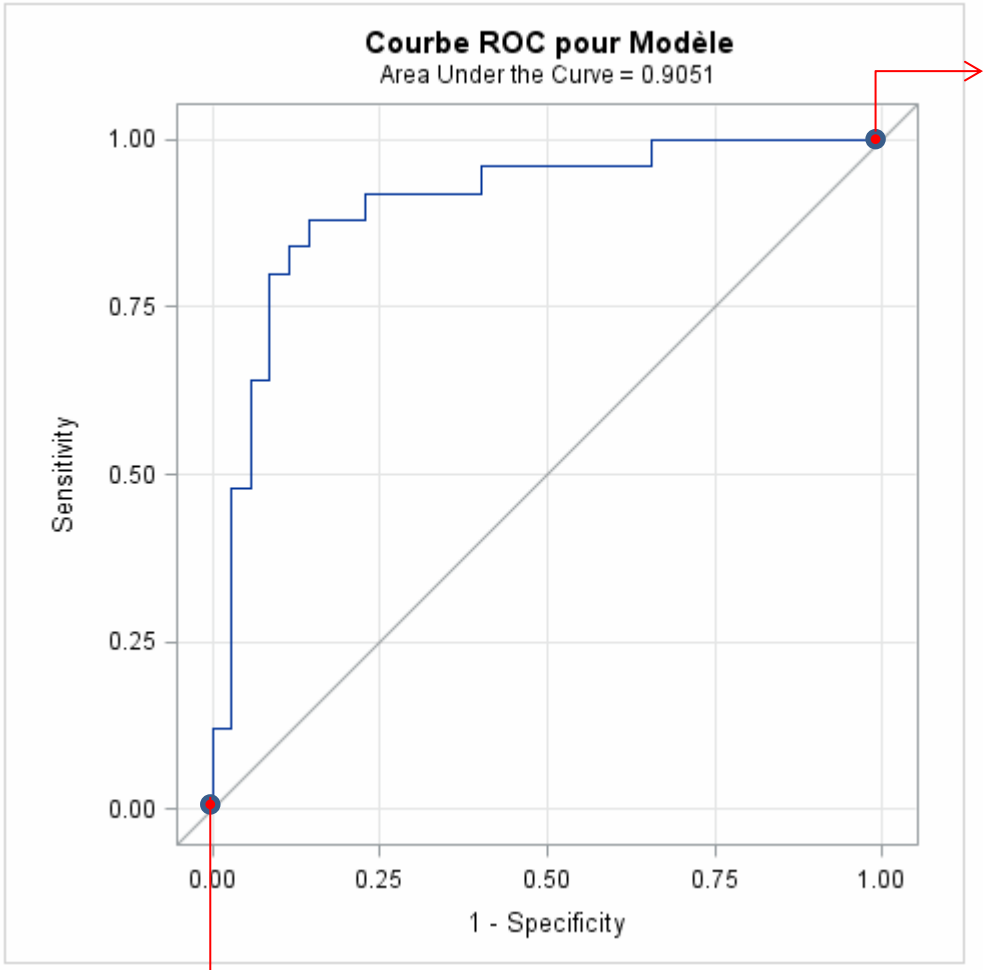
Table de classification

```
proc logistic data=Neuralgia;
   class Treatment Sex(DESC) /PARAM = REF;
   model Pain (event = 'YES') = Treatment Sex Age Duration / Ctable pprob= (0 to 1 by 0.1);
run;
```

Classification Table

Prob Level	Correct		Incorrect		Percentages				
	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.000	25	0	35	0	41.7	100.0	0.0	58.3	.
0.100	24	15	20	1	65.0	96.0	42.9	45.5	6.3
0.200	23	20	15	2	71.7	92.0	57.1	39.5	9.1
0.300	22	23	12	3	75.0	88.0	65.7	35.3	11.5
0.400	20	25	10	5	75.0	80.0	71.4	33.3	16.7
0.500	16	30	5	9	76.7	64.0	85.7	23.8	23.1
0.600	16	32	3	9	80.0	64.0	91.4	15.8	22.0
0.700	12	32	3	13	73.3	48.0	91.4	20.0	28.9
0.800	10	33	2	15	71.7	40.0	94.3	16.7	31.3
0.900	5	34	1	20	65.0	20.0	97.1	16.7	37.0
1.000	0	35	0	25	58.3	0.0	100.0	.	41.7

Courbe ROC


Courbe ROC pour Modèle
Area Under the Curve = 0.9051
Sensitivity
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
1 - Specificity
Seuil = 0%
Sensibilité = 100%
Spécificité = 0%
Seuil = 100%
Sensibilité = 0%
Spécificité = 100%

Courbe ROC

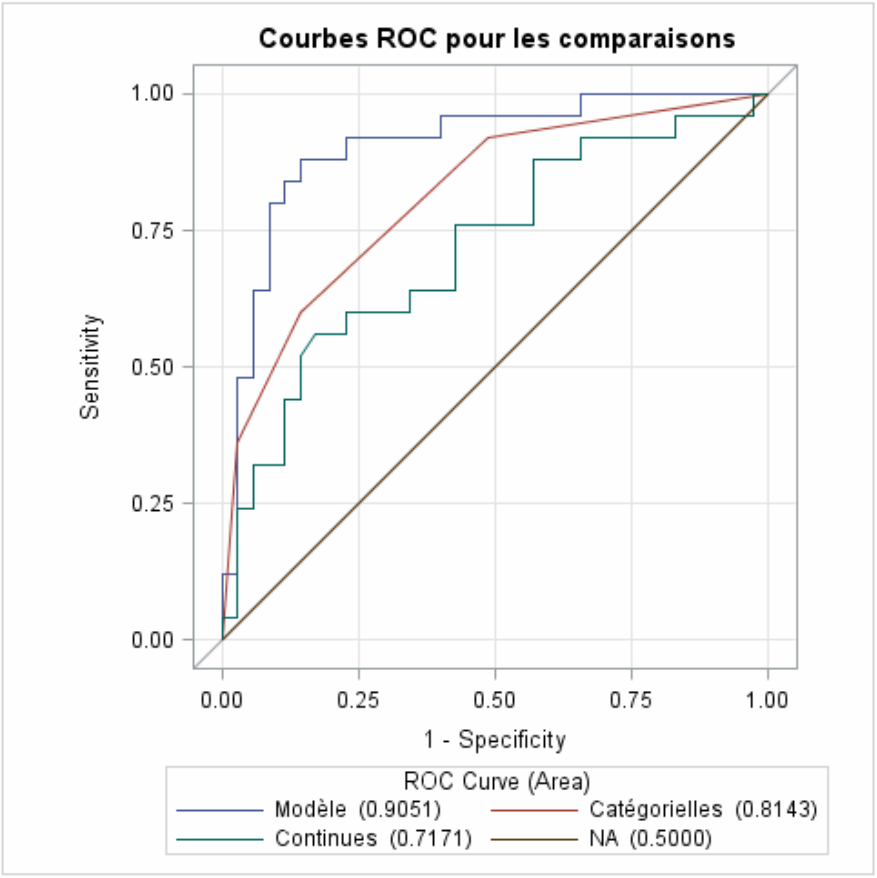
```
ods graphics on;
proc logistic data=Neuralgia;
   class Treatment Sex(DESC) /PARAM = REF;
   model Pain (event = 'YES') = Treatment Sex Age Duration/OUTROC = TableROC;
run;
```

- Variables:
  - *_SENSIT_* : Sensibilité;
  - *_1MSPEC_* : 1 – spécificité;
  - *_PROB_* : Le seuil de probabilité à partir duquel une observation est prédite comme étant un succès.

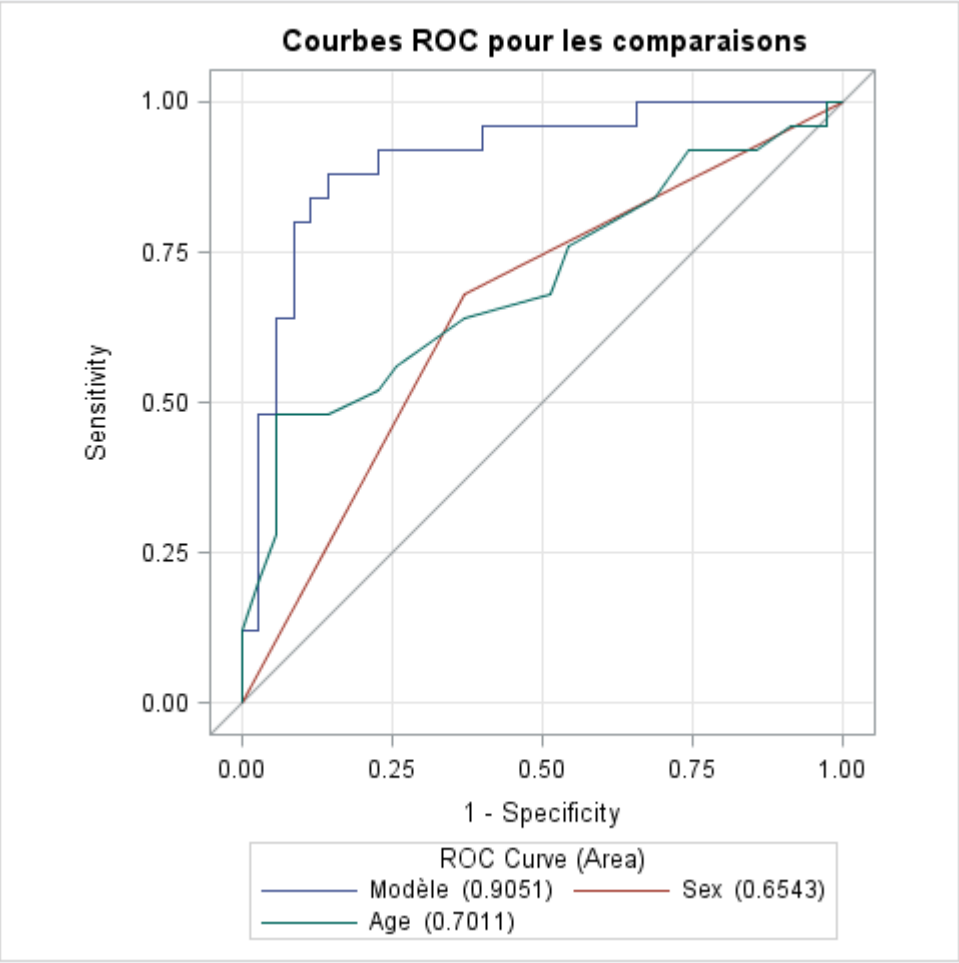
Courbe ROC

```
proc logistic data=Neuralgia;
   class Treatment Sex(DESC) /PARAM = REF;
   model Pain (event = 'YES') = Treatment Sex Age Duration/outroc = TableROC;
   ROC 'Catégorielles' Treatment Sex;
   ROC 'Continues' Age Duration;
   ROC 'NA';
run;
```

- *_SOURCE_* : Permet de distinguer les différents modèles considérés



Courbe ROC


Courbes ROC pour les comparaisons
Sensitivity
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
1 - Specificity
ROC Curve (Area)
Modèle (0.9051)
Sex (0.6543)
Age (0.7011)

Exportation Modèle

```
proc logistic data=Neuralgia OUTMODEL = Model_Logistic;
   class Treatment Sex(DESC) /PARAM = REF;
   model Pain (event = 'YES') = Treatment Sex Age Duration;
run;
```

```
proc logistic  inmodel=Model_Logistic;
    score data= Neuralgia_test
    fitstat out= Eval outroc = ROC;
run;
```

Procédure LOGISTIC

Fit Statistics for SCORE Data

Data Set	Total Frequency	Log Likelihood	Error Rate	AIC	AICC	BIC	SC	R-Square	Max-Rescaled R-Square	AUC	Brier Score
WORK.NEURALGIA_TEST	37	-23.4187	0.2973	58.83737	61.38283	68.97065	68.50288	0.080946	0.109263	0.884848	0.215134

EVAL

Filtrer et trier | Générateur de requêtes | Données | Description | Graphique | Analyser | Exporter | Envoyer à

	Treatment	Sex	Age	Duration	Pain	F_Pain	I_Pain	P_No	P_Yes
1	P	F	69	11	"		Yes	0.3512838928	0.6487161072
2	B	M	64	6	No	No	No	0.9271769152	0.0728230848
3	P	M	67	16	"		Yes	0.1310027059	0.8689972941
4	B	F	77	18	No	No	No	0.7387352566	0.2612647434

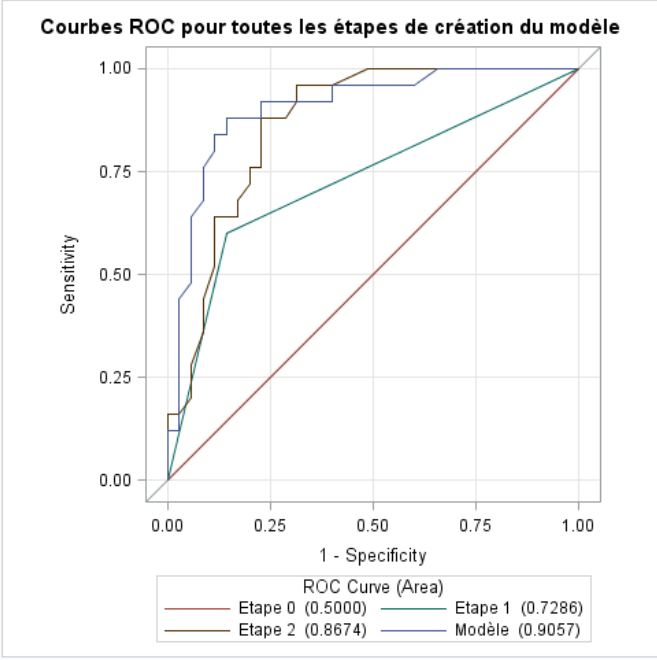
Sélection de variables

```
proc logistic data=Neuralgia OUTMODEL = Model_Logistic;
     class Treatment Sex(DESC) /PARAM = REF;
     model Pain (event = 'YES') = Treatment Sex Age Duration/
     SELECTION = stepwise
     SLENTRY = 0.05
     outroc = TableROC;
run;
```

- Plusieurs options de sélection:
  - Backward
  - Foward
  - Stepwise
  - Etc.
- SLENTRY ou SLSTAY: Le seuil déterminant si une variable est significative.

Sélection de variables

Summary of Stepwise Selection							
	Effect			Number	Score	Wald	
Step	Entered	Removed	DDL	In	Chi-Square	Chi-Square	Pr > Khi-2
1	Treatment		2	1	13.7143		0.0011
2	Age		1	2	10.6038		0.0011
3	Sex		1	3	5.9959		0.0143



Séparation complète ou quasi-complète des données

- WARNING: There is possibly a quasi-complete separation of data points. The maximum likelihood estimate may not exist.

Complete Separation

	Mutation=NO	Mutation=YES	Total
Drug=NON-EXPOSURE	0	14	14
Drug=EXPOSURE	37	0	37
Total	37	14	51

Quasi-Complete Separation

	Mutation=NO	Mutation=YES	Total
Drug=NON-EXPOSURE	12	0	12
Drug=EXPOSURE	25	14	39
Total	37	14	51

Séparation complète ou quasi-complète des données

- Solutions:
  - Vérifier le code
    - Variable expliquée présente dans les variables explicatives
  - Vérifier les fréquences croisées
    - Y-a-t-il une catégorie qui contient uniquement des échecs ou uniquement des réussites?
  - Trop de variables ou insuffisamment d'observations
    - Utiliser ou changer d'algorithme de sélection
    - Enlever les variables ou obtenir plus d'observations

Colinéarité

Note: The following parameters have been set to 0, since the variables are a linear combination of other variables as shown.

PlaceboOui =	Intercept - TreatmentA - TreatmentB

Analysis of Maximum Likelihood Estimates						
Parameter		**DDL**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > Khi-2**
Intercept		1	-17.4066	6.6914	6.7669	0.0093
Treatment	**A**	1	-3.1817	1.0161	9.8049	0.0017
Treatment	**B**	1	-3.7085	1.1407	10.5700	0.0011
Sex	**M**	1	1.8322	0.7963	5.2946	0.0214
Placebo	**Oui**	0	0	.	.	.
Age		1	0.2621	0.0970	7.2977	0.0069
Duration		1	-0.00586	0.0330	0.0315	0.8591

Remarque

- La régression logistique est généralisable pour une variable expliquée ayant plus de 2 catégories. La procédure SAS à utiliser lorsque cela se produit reste proc logistic.
- Lorsque la population utilisée pour l'apprentissage provient d'un sondage, il est mieux d'utiliser proc surveylogistic.

Bibliographie

- Support de SAS : https://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#logistic_toc.htm
- Pratique de la Régression Logistique, Régression Logistique Binaire et Polytomique, Ricco Rakotomalala.
- Microsoft Developer Network : http://msdn.microsoft.com/fr-fr/library/ms174828(v=sql.90).aspx.
- Nico J. D. Nagelkerke : A note on a general definition of the coefficient of determination. *Biometrika*, 78(3):691–692, 1991.
- William F. Mccarthy : The existence of maximum likelihood estimates for the binary response logistic regression model. *COBRA Preprint Series*, 2007.
- Institute for digital research and Education: What is complete or quasi-complete separation in logistic/probit regression and how do we deal with them? http://www.ats.ucla.edu/stat/mult_pkg/faq/general/complete_separation_logit_models.htm

R^2

```
proc logistic data=Neuralgia;
   class Treatment Sex(DESC) /PARAM = REF;
   model Pain (event = 'YES') = Treatment Sex Age Duration / rsquare;
run;
```

R carré	0.4208	R carré remis à l'échelle max.	0.5664

- R^2de Cox et Snell
- R^2de Nagelkerke