Analyse prédictive dans les assurances – possibilités et limites

Colloque ASA, 27 Novembre 2020

Anna SAROVA

Sommaire

- Introduction
- Champs d’application
- Loi et éthique
- Régression linéaire (LM)
- Modèle linéaire généralisé (GLM)
- Réseau de neurones
- Conclusion

Introduction

“If the statistics are boring, you’ve got the wrong numbers.”

—Edward Tufte

Analyse prédictive - une variété de techniques statistiques qui analysent les faits actuels et historiques pour faire des prédictions sur des événements futurs ou inconnus.

IDpol	ClaimNb	Exposure	VehPower	VehAge	DrivAge	BonusMalus	VehBrand	VehGas	Area	Density	Region
1	1	0.10	5	0	55	50	B12	Regular	D	1217	Rhone-Alpes
3	1	0.77	5	0	55	50	B12	Regular	D	1217	Rhone-Alpes
5	1	0.75	6	2	52	50	B12	Diesel	B	54	Picardie
10	1	0.09	7	0	46	50	B12	Diesel	B	76	Aquitaine
11	1	0.84	7	0	46	50	B12	Diesel	B	76	Aquitaine
...											
6114326	0	0.002739726	4	0	54	50	B12	Regular	E	3317	Provence-Alpes-Cotes-D'Azur
6114327	0	0.002739726	4	0	41	95	B12	Regular	E	9850	Ile-de-France
6114328	0	0.002739726	6	2	45	50	B12	Diesel	D	1323	Rhone-Alpes
6114329	0	0.002739726	4	0	60	50	B12	Regular	B	95	Bourgogne
6114330	0	0.002739726	7	6	29	54	B12	Diesel	B	65	Aquitaine

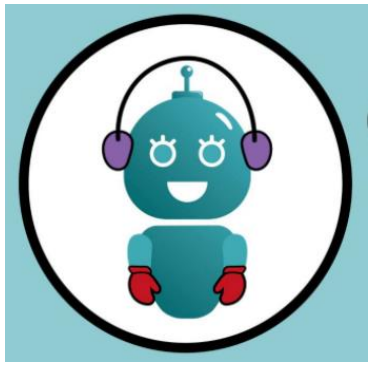
700k+

Champs d’application

- Pricing, reserving et sélection des risques
- Anticipation des fraudes
- Optimisation des campagnes marketing
- Amélioration du service client

Coach

sanitas



Loi et éthique

Cadres légales

- GDPR en Europe, loi LPD en Suisse
- Loi LSA

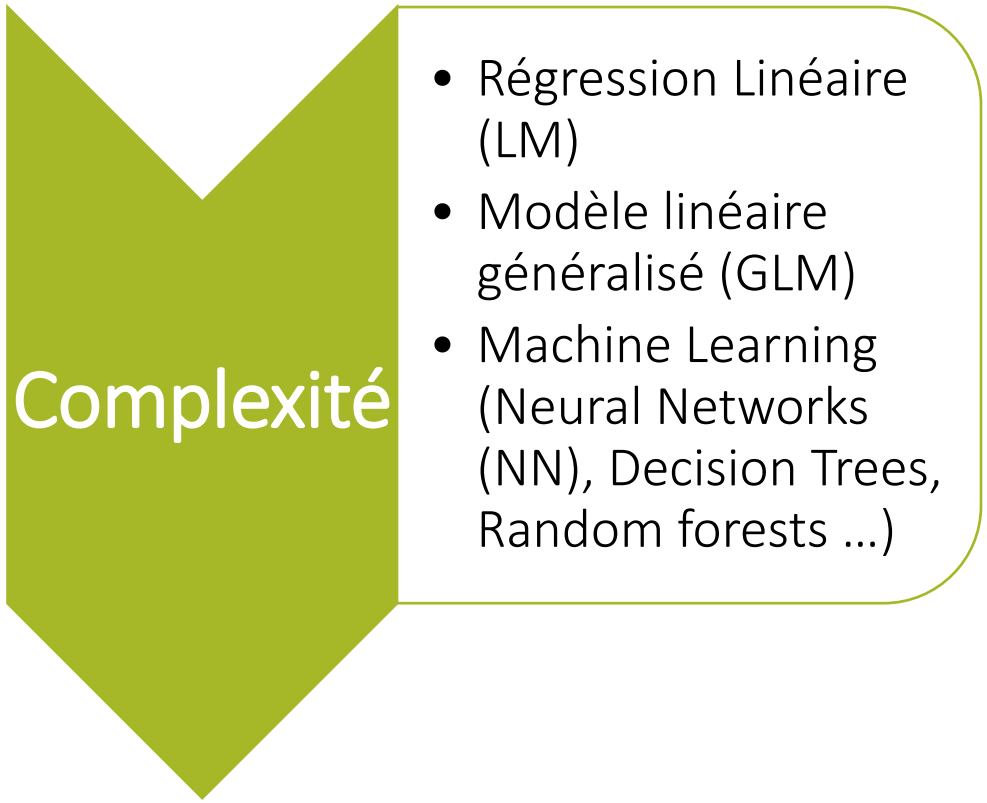
Missions

- Préserver la vie privée, mais permettre la rétractation
- Travailler avec des données limitées
- Flexibilité et l'interprétabilité des données et processus
- Empêcher la discrimination et biais

Possibilités

Les technique utilisées:

- Techniques de régression
- Apprentissage automatique (machine learning)



Régression linéaire (1)

La régression linéaire est utilisée pour prédire la valeur de la variable dépendante Y par la combinaison linéaire des variables explicatives X.

Dans le cas univarié, la régression linéaire peut être exprimée comme suit:

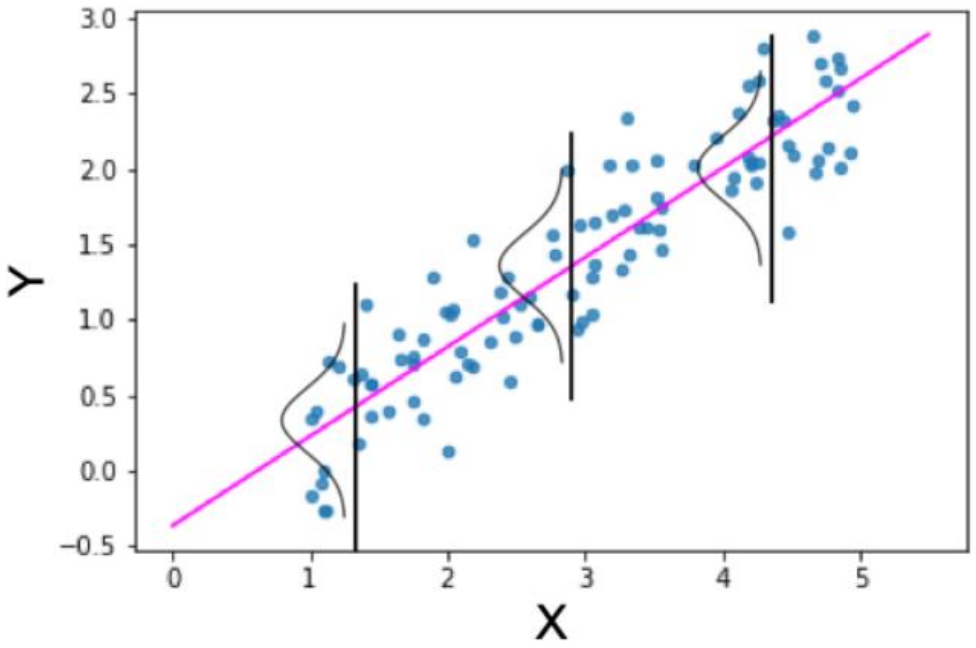
$$\mathbb{E}(Y) = \beta_0 + \beta_1 X$$

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

$$\varepsilon \sim \mathcal{N}(0, \sigma^2)\ idd$$

$$Y \sim \mathcal{N}(\beta_0 + \beta_1 x\,, \sigma^2) \quad pour\ un\ x\ fixe$$

Quelles sont les limitations?

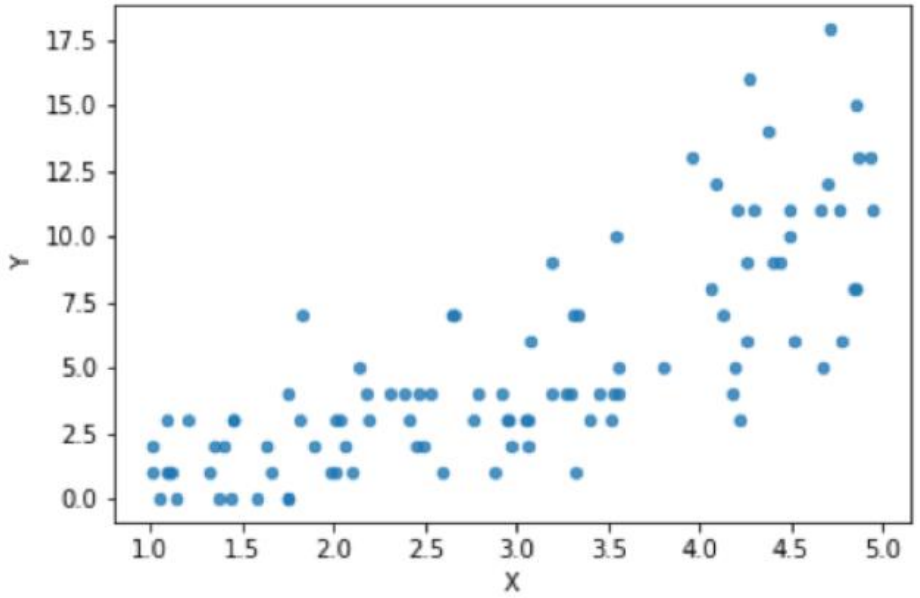


Régression linéaire (2)

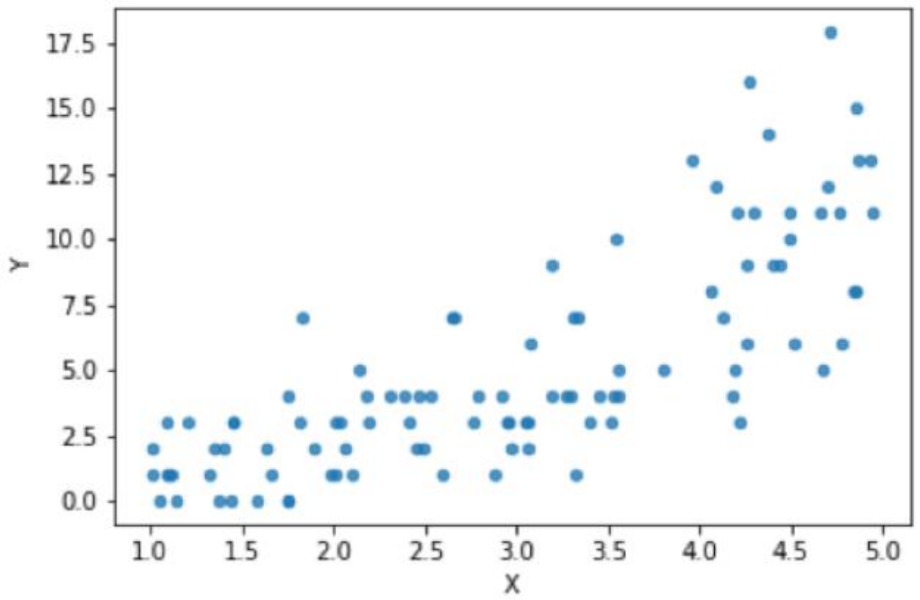
Supposons qu'on veut prévoir le nombre de produits défectueux (Y) avec une valeur de capteur (X) comme variable explicative.

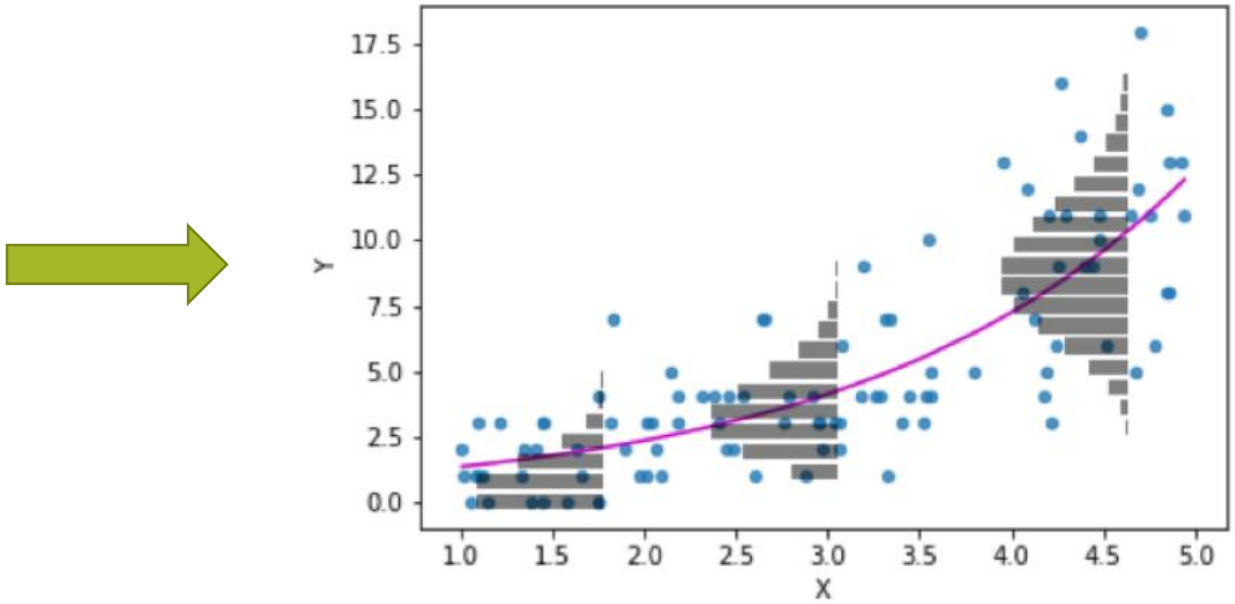
Limitations:

- La relation entre X et Y n'a pas l'air linéaire. Il est plus probable qu'elle soit exponentielle.
- La variance de Y ne semble pas constante par rapport à X. Ici, la variance de Y semble augmenter lorsque X augmente.
- Comme Y représente le nombre de produits, il doit toujours être un nombre entier positif.



GLM de Poisson





$$Y \sim \text{Poisson}(\lambda)$$
$$g\big(\mathbb{E}(Y)\big) = ln(\lambda) = \beta_0 + \beta_1 X \Leftrightarrow \mathbb{E}(Y) = \lambda = \exp(\beta_0 + \beta_1 X) > 0$$

La ligne rose trace l'espérance $\mathbb{E}(Y)$

Théorie: LM vs GLM

Régression linéaire (LM)	Modèle linéaire généralisé (GLM)
$\mathbb{E}(Y) = \beta_0 + \beta_1 X$	$\mathrm{g}\big(\mathbb{E}(Y)\big) = \beta_0 + \beta_1 X$
$Y = \beta_0 + \beta_1 X + \varepsilon$	$Y = g^{-1}(\beta_0 + \beta_1 X) + \varepsilon$
$\varepsilon \sim \mathcal{N}(0, \sigma^2)\ idd$	$\varepsilon \sim famille\ exponentielle\ idd$
$Y \sim \mathcal{N}(\beta_0 + \beta_1 x\,, \sigma^2)\ \ pour\ un\ x\ fixe$	$Y \sim famille\ exponentielle\ idd$

Famille exponentielle: loi Normale, Binomiale, Poisson, Gamma …

GLM exemple: assurance automobile

But: prédire la fréquence de sinistres pour le portefeuille assuré.

Le nombre de sinistre par année pour chaque police dans le portefeuille:
$Y \sim \mathbf{Pois}(\lambda(X))$

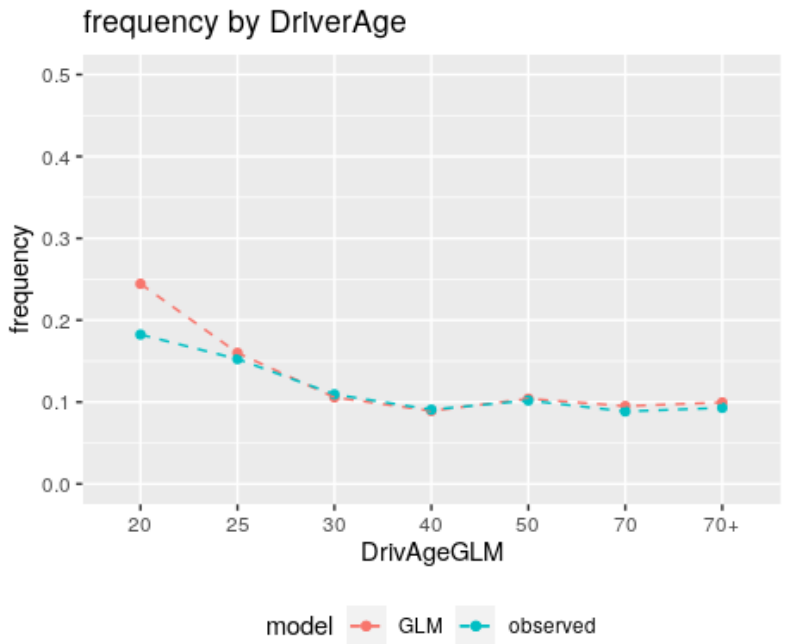
Les étapes à suivre:

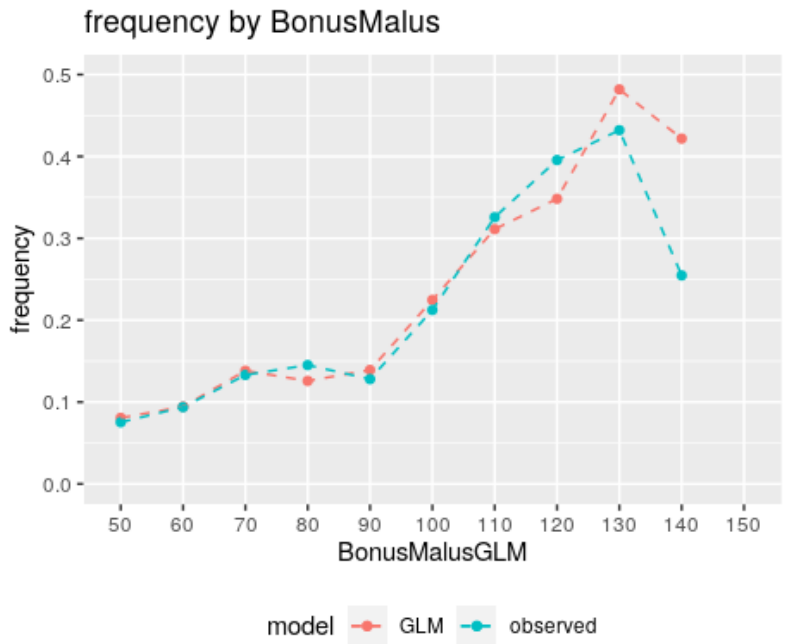
1. Choisir le modèle approprié
2. Diviser les données: apprentissage et essai
3. Décider les variables explicatives X
4. Appliquer GLM pour estimer la fréquence $\lambda(X)$: minimiser la f. de déviance β
5. Tester plusieurs modèles

GLM exemple: résultats

Les variables explicatives possibles (X):

- Area
- VehPower
- VehAge
- DriverAge
- BonusMalus
- ….





Données: freMTPL2freq, R package CASdatasets

GLM: possibilités et limites

+ Y ne suit plus juste la loi Normale

+ Robustes, transparents et « assez » faciles à comprendre

+ Reconnus dans l'industrie

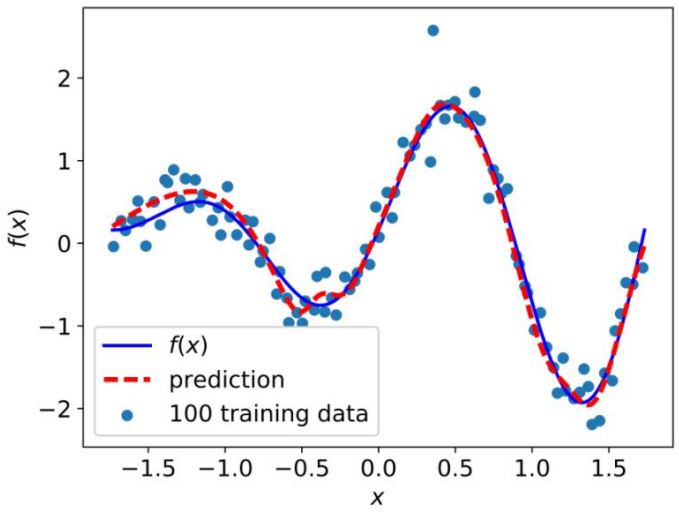
- Hypothèses concernant la distribution de Y existe

- Résultats moins interprétable que LM

- Nécessité de grandes quantités de données

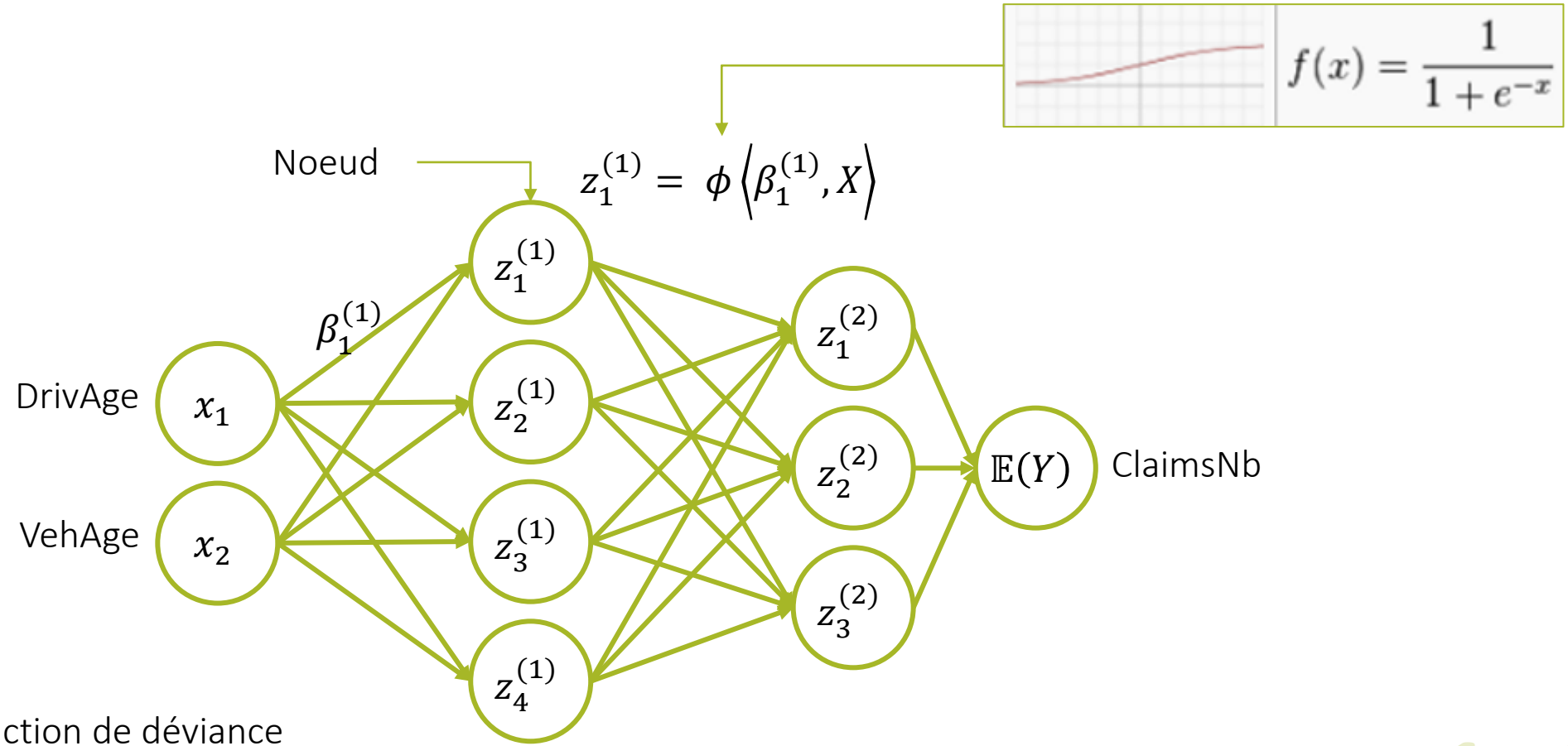
Réseau de neurones (1)

- Généralisation de GLM
- Fonction de déviance comme fonction à minimiser
- La principale différence avec les GLM:
  le prédicteur linéaire est remplacé par un non linéaire



Modèle linéaire généralisé (GLM)	Neural networks
$\mathrm{g}\big(\mathbb{E}(Y)\big) = \langle \beta, X \rangle$	$\mathrm{g}\big(\mathbb{E}(Y)\big) = \langle \beta, z^{(d:1)}(X) \rangle$ $z_j^{(m)}(X) = \phi \left\langle \beta_j^{(m)}, X \right\rangle$ ϕ fonction d'activation, non linéaire

Réseau de neurones (2)



But: minimiser la fonction de déviance
Dans cet exemple résoudre pour $\beta \in \mathbb{R}^{31}$

Réseau de neurones: possibilités et limites

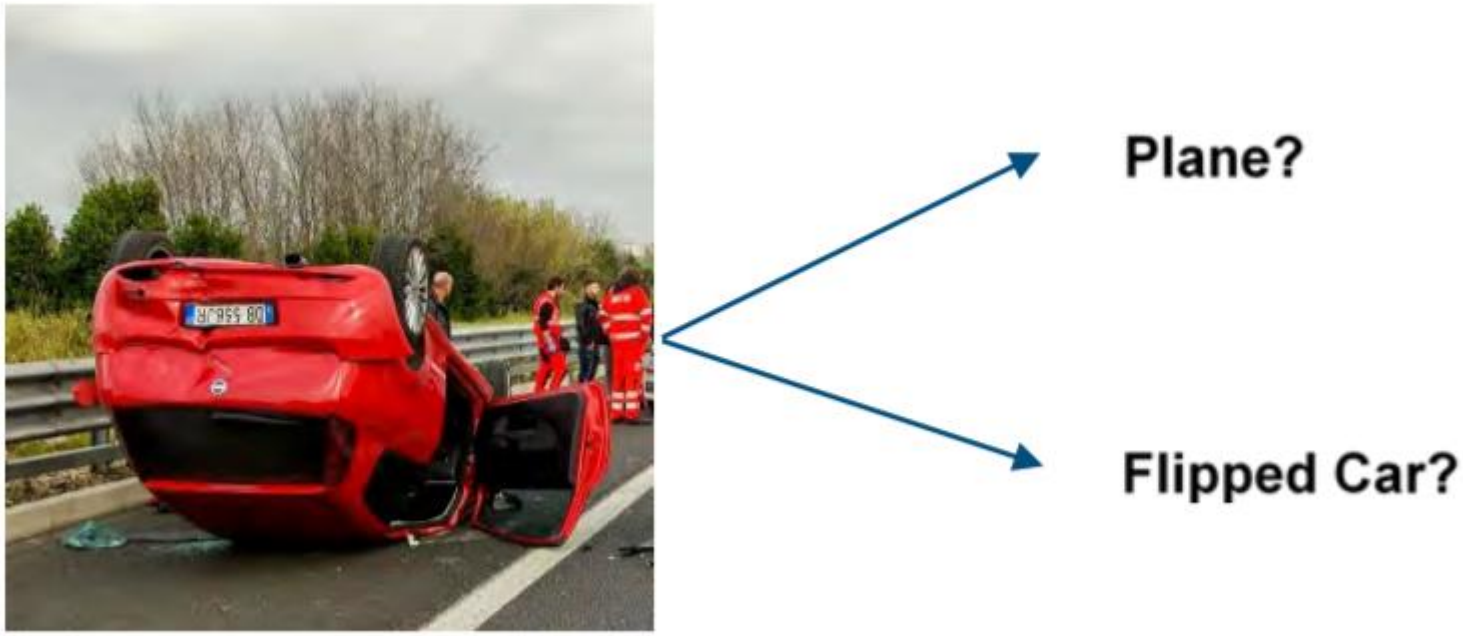
+ Possible de modéliser de relations non – linéaires

+ Pas de restriction sur la distribution de Y

- Pas de solution unique

- Interprétation difficile

… et encore une limite



Conclusion

- Analyses prédictives offrent d’innombrables possibilités
- Cadre éthique et légal doivent être respecté
- Connaissance et compréhension des modèles utilisés est importante
- Garder les modèles à jour et produire la documentation nécessaire

Merci!

Sources:

- Wuthrich, Mario V., From Generalized Linear Models to Neural Networks, and Back (December 11, 2019). Available at SSRN: https://ssrn.com/abstract=3491790 or http://dx.doi.org/10.2139/ssrn.3491790
- Ferrario, Andrea and Noll, Alexander and Wuthrich, Mario V., Insights from Inside Neural Networks (April 23, 2020). Available at SSRN: https://ssrn.com/abstract=3226852 or http://dx.doi.org/10.2139/ssrn.3226852
- Yuho Kida, Generalized linear models Introduction to advanced statistical modeling: https://towardsdatascience.com/generalized-linear-models-9cbf848bb8ab
- https://www.dlapiperdataprotection.com/?t=law&c=CH
- https://www.jipitec.eu/issues/jipitec-10-2-2019/4916
- https://www.soa.org/globalassets/assets/files/resources/research-report/2018/applying-image-recognition.pdf

Annexe (1): régression linéaire

Les coefficients β_0, β_1 sont déterminés par la méthode des moindres carrés qui minimise la somme des carrés des erreurs:

$$L(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 + \beta_1 X_i)^2$$

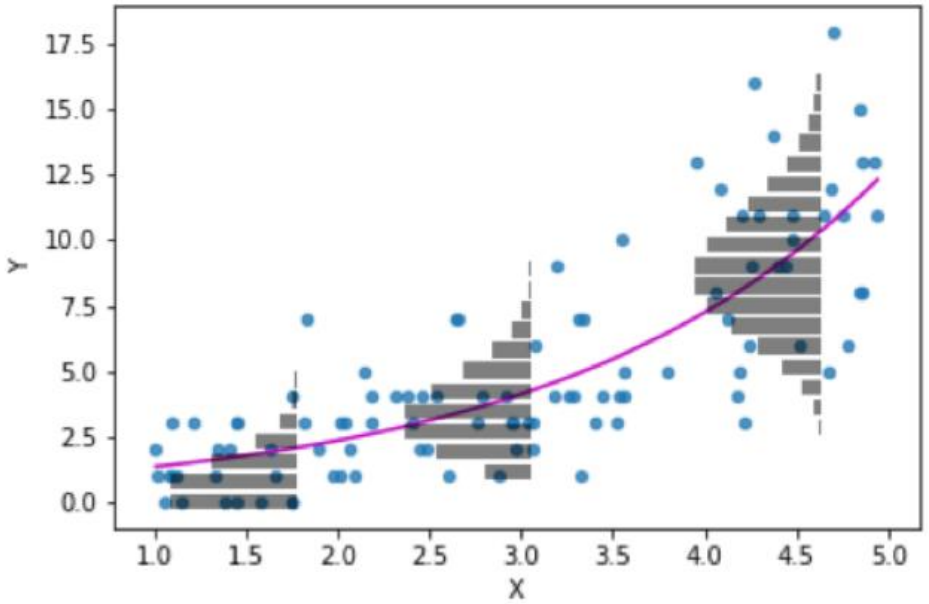
On résout le système des deux équations à deux inconnus:

$$\nabla L(\hat{\beta}_0, \hat{\beta}_1) = 0$$

Annexe (2): fonction de déviance

Pour un modèle M_0 et les observations y la déviance est définie comme suit:

$$\mathrm{D}(\mathrm{y}, \hat{\lambda}) = 2 \times (\log\left(p(y|\hat{\theta}_s)\right) - \log\left(p(y|\hat{\theta}_0)\right))$$



$\hat{\lambda} = \mathbb{E}(Y|\hat{\theta}_0)$: les estimations du modèle M_0

$\hat{\theta}_0$: les paramètres du modèle M_0

$\hat{\theta}_s$: les paramètres du modèle saturé M_s

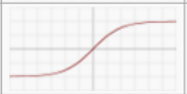
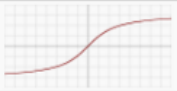
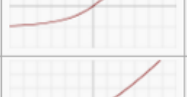
Un modèle saturé est un modèle dans lequel il existe autant de paramètres estimés que de points de données: chaque observation $y_i \sim \mathrm{Pois}(y_i)$

Annexe (3): fonction de déviance de Poisson

$$D(y, f(\boldsymbol{x})) = 2\ln\prod_{i=1}^{n}\exp(-y_i)\frac{y_i^{y_i}}{y_i!} - 2\ln\prod_{i=1}^{n}\exp\{-f(\boldsymbol{x}_i)\}\frac{f(\boldsymbol{x}_i)^{y_i}}{y_i!}$$

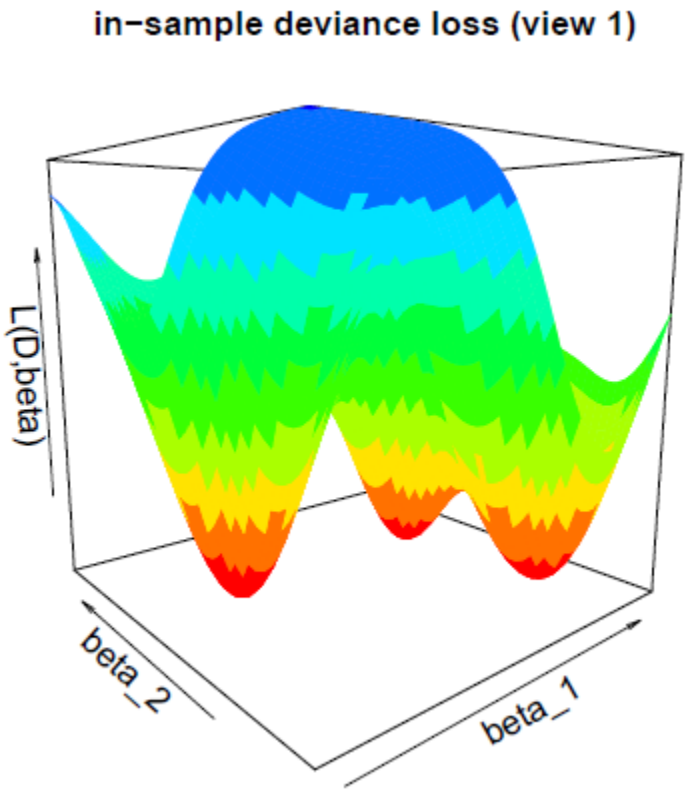
$$= 2\sum_{i=1}^{n}\left[y_i\ln\frac{y_i}{f(\boldsymbol{x}_i)} - \{y_i - f(\boldsymbol{x}_i)\}\right].$$

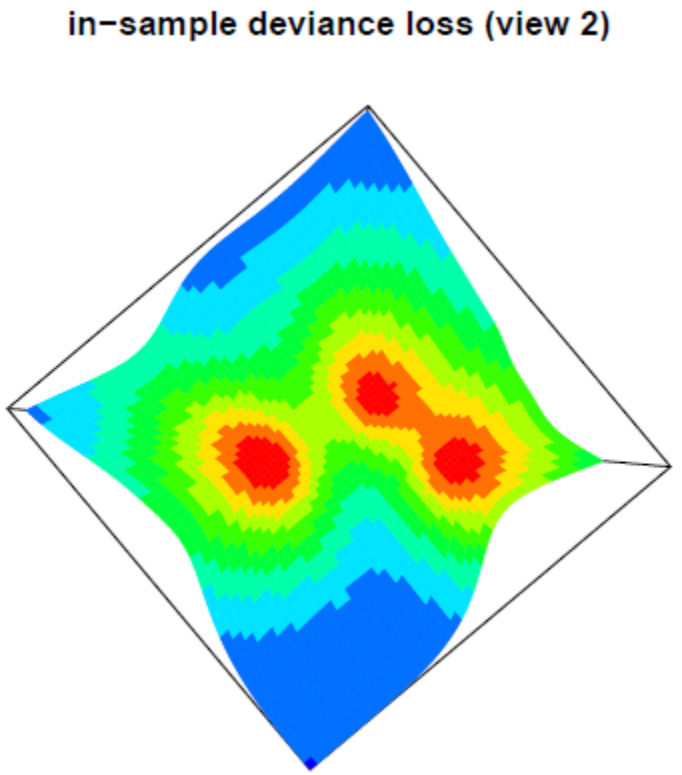
Annexe (4): Fonctions d'activation

Name	Plot	Equation
Identity		$f(x) = x$
Binary step		$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$
Logistic (a.k.a Soft step)		$f(x) = \frac{1}{1 + e^{-x}}$
TanH		$f(x) = \tanh(x) = \frac{2}{1 + e^{-2x}} - 1$
ArcTan		$f(x) = \tan^{-1}(x)$
Rectified Linear Unit (ReLU)		$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$
Parameteric Rectified Linear Unit (PReLU) [2]		$f(x) = \begin{cases} \alpha x & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$
Exponential Linear Unit (ELU) [3]		$f(x) = \begin{cases} \alpha(e^x - 1) & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$
SoftPlus		$f(x) = \log_e(1 + e^x)$

Annexe (5): Fonction de deviance non convexe

La fonction de déviance n'est pas convexe en β => pas de solution unique!





Annexe (6): Réassurance facultative vie

- Le réassureur qui donne en premier la réponse positive, gagne le contrat
- La demande pour la réassurance facultative a beaucoup augmenté
- Modèle GLM: pour optimiser le temps et aider les souscripteurs