METHODOLOGIE : REGRESSION LOGISTIQUE

Pour vouloir comprendre notre tambouille, vous devez être soit courageux soit irresponsables. Dans les deux cas, vous aimez souffrir. Nous, on les bichonne nos masos. Alors vous allez en avoir pour votre argent. Comment ça, vous n'avez pas payé ? Mince, il faut remédier à ça…

I- Principes généraux

1. Usages des régressions logistiques

Prenons un exemple concret (il n'y en aura pas beaucoup). 38% des personnages féminins décèdent au cours de la série contre 52% des personnages masculins.

« Chouette ! se dit Catelyn. J'ai moins de risque de mourir que mon mari. » Quand tout à coup *patatra*. Elle meurt (*Fig. 1*).

Alors déjà, si 38% c'est mieux que 52%, ce n'est pas la joie pour autant. D'autant plus que ces risques globaux cachent bien des choses ! Les personnages féminins et masculins diffèrent en de nombreux points. Par exemple : les personnages masculins sont plus souvent des combattants et leurs homologues féminines se prostituent plus. Or, un combattant a plus de risque de mourir qu'un non combattant (logique, non ?) et un personnage qui se prostitue à moins de risque de mourir qu'un autre qui ne se prostitue pas[1] ! Dès lors, comment savoir si les personnages féminins meurent moins *parce qu*'ils sont féminins (on parle alors d'*effet propre* du sexe sur la mortalité) ou si ce sont leurs caractéristiques particulières qui les protègent (on parle d'*effet de composition* : par exemple, si les prostitué.e.s meurent moins, et que les personnages féminins se prostituent plus, alors les personnages féminins sont protégés par cet effet de composition).

Pour neutraliser les effets de composition, nous recourons aux régressions. Celles-ci nous permettent de déterminer l'effet propre du sexe sur les risques de mortalité *toutes choses égales par ailleurs*. Concrètement, on peut alors étudier l'impact du sexe sur la mortalité en faisant *comme si* les personnages féminins et masculins combattaient et se prostituaient autant. Génial, non ?

[1] 53% des personnages non prostitués décèdent au cours de la série contre 13% des personnages qui se prostituent. De même, 68% des personnages qui combattent décèdent contre 37% des personnages qui ne combattent pas.

Fig. 1 : La désillusion de Catelyn Stark

Bref, on met toutes les caractéristiques qu'on veut dans le modèle. On n'oublie pas de dire au modèle si le personnage est mort ou non dans la série, sinon il galère… A partir de ça, le modèle peut calculer le risque théorique de mourir d'un personnage compte-tenu de ses caractéristiques. Pas mal, non ?

Les résultats des régressions peuvent être utiles pour faire des prédictions : vous disposez de risques calculés sur les sept premières saisons et vous les appliquez à la prochaine saison. C'est un peu osé parce que les scénaristes peuvent faire ce qu'ils veulent dans l'ultime saison. Jon Snow peut donc (re)mourir car l'intrigue n'aura plus besoin de lui. On peut aussi s'en servir pour démontrer des causalités : par exemple, Catelyn Stark est morte *parce qu*'elle était un grand seigneur et Oberyn Martell *parce qu*'il combattait. Encore une fois, c'est osé. Ils sont surtout morts *parce que* les scénaristes l'ont décidé. Nous utilisons simplement les régressions pour observer des liens entre le risque de mourir et certaines caractéristiques : il existe ainsi un lien entre le fait de combattre et celui de mourir !

2. Un exemple complet de lecture

Pour ceux qui ne se sentent pas encore très à l'aise avec la lecture des régressions (on peut comprendre), nous proposons un exemple complet de lecture sur notre site internet (*Fig. 2*). Pour d'autres exemples concrets, n'hésitez pas à vous reporter aux régressions présentées dans nos articles (« *GoT :* série féministe ou misogyne ? », « *GoT :* des corps « masculins et normaux » ? »…) qui sont toujours accompagnées de notes de lecture. C'est normalement compréhensible.

	Modèle 1		Modèle 2	
	Ods Ratio	**Significat.**	**Ods Ratio**	**Significat.**
Intercept	-	****	-	41%
Sexe				
Homme	*(Réf.)*		*(Réf.)*	
Femme	2 fois moins	***	-	38%
Prostitution				
N'est pas un.e prostitué.e	X		*(Réf.)*	
Est un.e prostitué.e	X		4 fois moins	°
Combattant.e				
N'est pas combattant.e	X		*(Réf.)*	
Garde royal.e	X		-	79%
Chevalier.e	X		-	31%
Autre combattant.e	X		3 fois plus	****
Nombre de personnes de même allégeance				
Nombre de personnes de même allégeance	-2% par unité	***	-2% par unité	**
**** p < 0,01% ; *** p < 0,1% ; ** p < 1% ; * p < 5% ; ° p < 10%				
X : caractéristique non mise dans le modèle				

Fig. 2 : Un exemple simple de lecture disponible sur notre site internet !

Si malgré cela vous ne saisissez toujours pas… Dites-vous qu'il n'y a pas de mauvais élèves mais que des mauvais professeurs.

II- Méthodologie complète : pour aller encore plus loin !

1. Les paramètres du modèle

a) Coefficients

$$Ln(\frac{p}{1-p}) = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + ... + Bi * Xi$$

Avec *p* la probabilité de mourir au cours de la série, *Xi* les variables dépendantes, *Bi* les coefficients de régression calculés.

- Si ce coefficient est négatif : la caractéristique protège de la mort.
- Si ce coefficient est positif : la caractéristique expose à la mort.

Toutefois, nous n'utilisons pas ces coefficients dans nos articles. En effet, il est très difficile d'interpréter ce coefficient et d'en faire des phrases concrètes. Dommage la formule est assez esthétique d'après Romane !

D'ailleurs, notre statisticienne peste : les mathématiques peuvent s'expliquer ! « Cette formule n'est pas qu'esthétique ! Elle relie un nombre *p* (qui est un score ; plus le score est grand, plus le personnage a de risque de décéder) à des caractéristiques individuelles (*Xi*). » Elle calcule l'intensité des liens (les *Bi*) entre les caractéristiques et les risques de décès. Pour cela elle dispose d'une vraie liste de personnages morts ou non. Elle va faire en sorte que ceux qui sont morts aient effectivement un « score » plus élevé que ceux qui sont vivants. Une fois que les *Bi* sont déterminés (il s'agit pour ceux qui sont copains-copains avec les mathématiques d'un problème de maximisation), *p* est une fonction des *Xi*.
Oui, mais pourquoi cette mocheté de logarithme népérien ? Simplement pour que le score (*p*) soit compris entre 0 et 1 et soit associable à une probabilité !

b) Odds Ratio

Quoiqu'il en soit, nous préférons dans nos articles l'utilisation des *Odds Ratios* (OR).

- Si l'OR est plus petit que 1, la caractéristique protège de la mort.
- Si l'OR est plus grand que 1, la caractéristique expose à la mort.

Mais quel est l'intérêt par rapport aux coefficients ? L'interprétation ! Si l'OR est de 2 pour le fait de combattre on peut alors dire que « par rapport à un personnage qui ne combat pas, un personnage qui combat a deux fois plus de risque de mourir ». Si l'OR est de 0,5 pour la prostitution : « par rapport à un personnage qui ne se prostitue pas, un personnage qui se prostitue a deux fois moins (1/0,5=2) de risque de mourir ».

C'est plus commode, vous en conviendrez.

c) Significativité

La significativité statistique d'un coefficient permet d'évaluer le niveau de certitude selon lequel on peut affirmer qu'il existe un lien entre la caractéristique et le risque de décéder. Par exemple, pour une significativité de 5% : « il existe un risque de 5% pour que la sous ou la surmortalité mise au jour soit en fait inexistante ».

En sciences humaines et sociales, nous prenons en général 5% de risque au maximum. Au-delà le risque de se planter est trop grand : le résultat est dit « non significatif ». Mais ce n'est qu'une convention et ça dépend de l'objectif ! Imaginez : vous avez une conjonctivite. Un nouveau médicament fait des merveilles. Seulement, il cause la mort dans 4% des cas. Vous prenez le risque ? En sciences humaines et sociales, 4% de risque c'est peu. En médecine, ça peut être dramatique…

Nous, on aime tellement le risque que l'on vous indique même jusqu'à 10% (°) de risque ! Les lecteurs se souviendront qu'il faut prendre ces résultats avec des pincettes.

2. Qualité du modèle

Notre modèle est créé ! Il reste à en évaluer la qualité. Dans nos articles, nous n'en parlons pas. On vous demande un acte de foi. On vous en donne toutefois un aperçu (*Fig. 3*).

Nous pouvons distinguer deux types d'indicateurs :

- *Type 1 :* Les indicateurs qui comparent les risques calculés par le modèle à la réalité

a) Paires concordantes

Une première manière d'évaluer la qualité du modèle est de compter le nombre de paires concordantes. D'accord, mais qu'est-ce que c'est ?

Le logiciel trie nos personnages en deux groupes : ceux qui sont morts et ceux, plus chanceux, qui ne le sont pas. Il va ensuite créer des paires en prenant à chaque fois un personnage mort et un vivant. Il compare ensuite les risques de décéder (calculés par le modèle) de la paire :

- Si le risque de mourir calculé par le modèle est plus fort pour le personnage effectivement mort que pour celui qui est en vie, c'est bien ! Cela veut dire que le modèle a marché : c'est une paire

Indicateurs de qualité du modèle	Part paire concordante	Aire sous la courbe de ROC	AIC divisé par AIC_L0	AIC brute (AIC modèle L0)
Femme (base)	74%	74%	4%	514 (553)
Femme (prostitution)	74%	75%	5%	509 (553)
Femme (combattant)	79%	79%	8%	491 (553)
Corps	85%	85%	19%	435 (553)
Profil	86%	86%	21%	420 (553)
Allégeance (base)	77%	77%	8%	492 (535)
Allégeance (comb.)	80%	80%	10%	479 (535)
Allégeance (richesse)	82%	82%	14%	462 (535)
Allégeance (CR*)	83%	83%	14%	460 (535)

* CR : Conseil restreint

Fig. 3 : La qualité de nos modèles (voir notre site internet pour plus d'explications…)

concordante.

- Sinon, c'est une paire discordante. Le modèle n'explique pas le fait qu'un personnage soit mort dans la série et l'autre non.

Nous avons 193 personnages vivants et 205 morts. Nous avons donc 39 565 paires à vérifier (193*205) ! Plus la part de paires concordantes parmi le nombre total de paires est élevée, plus le modèle est explicatif.

b) Aire sous la courbe de ROC

Voilà un autre indicateur de qualité du modèle (nommé c). Il s'agit de l'aire sous la courbe de *ROC* (proposée par Hosmer et Lemeshow en 1941) qui varie en 0 et 1. Alors, autant on peut expliquer concrètement ce que c'est qu'une paire concordante, autant là… on ne peut pas. Toujours est-il que l'on considère que la discrimination des deux populations (ici les personnages décédés et les personnages vivants) est :

- nulle si l'aire sous la courbe vaut 50% ;
- acceptable si elle appartient à [70% ; 80%[;
- excellente si elle appartient à [80% ; 90%[;
- exceptionnelle si elle est supérieure ou égale à 90%.

Concrètement, le modèle comprend exceptionnellement bien ce qui différencie personnages vivants et personnages morts si l'indicateur est supérieur ou égal à 90%.

- *Type 2 :* Les indicateurs s'appuyant sur la vraisemblance du modèle

c) AIC et SC

La part de paires concordantes est cependant influencée par le nombre de variables introduites dans le modèle : plus j'injecte de variables dans mon modèle, plus celui-ci sera jugé « bon » par les indicateurs vus ci-dessus (paire concordante et aire sous la courbe de *ROC*). C'est assez facile à comprendre : n'importe quelle caractéristique peut expliquer une infime partie d'un phénomène. Vraiment n'importe quelle caractéristique ! Même la couleur de la voiture de l'acteur... C'est du hasard, mais il n'empêche que ça ne peut qu'améliorer l'indicateur des paires concordantes ! Alors comment pallier ce biais ?

Deux autres indicateurs nous renseignent sur la qualité du modèle tout en prenant en compte le nombre de variables introduites : le critère d'Alsaïke (*AIC*) et celui de Schwartz (*SC*) qui valent respectivement :

$$AIC = 2 * (K + 1) - 2\ln(L)$$
$$SC = (K + 1) * \ln(n) - 2\ln(L)$$

Avec *L*, la valeur maximale de la vraisemblance (calculée avec les valeurs estimées des paramètres) et *K* le nombre de variables introduites dans le modèle.

Plus ces critères sont petits, meilleure est la vraisemblance. Mais attention ! Si l'on ajoute une caractéristique supplémentaire, *hop*, les critères augmentent. Il faut donc vérifier, lorsque l'on ajoute une variable, que le gain en vraisemblance compense l'ajout de la variable : idéalement le critère doit diminuer. On peut se contenter d'une stagnation ou d'une légère augmentation si dans le même temps les paires concordantes et autres indicateurs augmentent fortement.

Vous l'aurez compris : tout est question d'équilibre. Il faut mettre assez de variables pour expliquer au mieux un phénomène tout en restant parcimonieux pour ne pas perdre en vraisemblance.

Pour interpréter ces deux critères, il est utile de connaître leur valeur pour un modèle sans caractéristique explicative (L_0). Si nos indicateurs deviennent supérieurs à ce modèle qui par définition n'explique rien… on a du souci à se faire.

Bibliographie

Asfa, C. (2016) « Le modèle Logit Théorie et application ». *In : Document de travail*, INSEE. [En ligne] : https://www.INSEE.fr/fr/statistiques/fichier/2022139/Le-modele-Logit-CB.pdf [Consulté le 6 avril 2018]

Gillaizeau, F. et Grabar, S. (2011). « Modèles de régression multiple ». *In : Sang Thrombose Vaisseaux*. N°7. [En ligne] : http://docplayer.fr/49631088-Modeles-de-regression-multiple.html [Consulté le 6 avril 2018]

Janvier, B. (2001). « La significativité statistique ». [En ligne] : http://baptiste.janvier.free.fr/stats/pdf/proba.pdf [Consulté le 13 avril 2018]