

Les modèles logistiques

ou parier futé

Marc Bourdeau

Dans ce chapitre, nous nous intéressons à la modélisation linéaire dans le cas où la variable dépendante, la variable de réponse, est binaire. Une variable binaire observée étant en fait la réalisation d'un essai de Bernoulli, on se voit dans la situation des paris : tel résultat se produira-il ou non ? Est-ce que je parierais sur une des alternatives plutôt que sur l'autre ? Comment choisir ? Et donc comment parier dans le cas où on peut penser que des facteurs ont éventuellement une influence sur la réponse.

C'est l'évaluation de ces facteurs qui nous concerne en modélisation : peut-on améliorer nos estimations des probabilités d'une binomiale en mettant en action les facteurs d'influence potentielle sur cette réponse, en modélisant donc de façon futée leur influence.

Nous abordons ces types de modèles linéaires avec la perspective du parieur. Les concepts de base sont en effet de pratique courante pour tous les parieurs. L'incertitude étant la règle plutôt que l'exception, parier sur des événements qui se produiront ou non dans l'avenir est la façon probabiliste fondamentale de comprendre la réalité. N'est-ce pas là d'ailleurs le sens profond à donner au mot bien plus scientifique d'hypothèse ?

Les applications des ces techniques sont fréquentes : en ingénierie, une variable binaire soumise à toutes sortes d'influence est la conformité d'un produit à certaines normes, son bon fonctionnement après un temps donné d'usure, la présence ou non d'une caractéristique de fabrication ; en sciences sociales, on peut penser à une décision que peut prendre un sujet, adopter ou non tel un certain comportement, avoir ou non une opinion positive sur des questions particulières, *etc.*, toutes questions qui dépendent de nombreux

facteurs potentiels et mesurables ; en sciences médicales un individu a ou non telle ou telle maladie en fonction des ses habitudes de vie, de son âge, de son sexe, de son état général de santé, *etc.*

1 Parier

1.1 La perspective du parieur

Le parieur occasionnel a tendance à raisonner sur des probabilités. Ainsi il dira « j'ai deux chances sur trois de gagner tel ou tel pari ». S'il donne une proportion suérieure à une demie, il aura tendance à parier, sinon il s'abstiendra. À une chance sur deux, il sera en réalité indifférent. La plupart des parieurs voudront prendre des paris seulement dans le cas assez favorables, selon leur estimation évidemment, et ils seront dans ces cas prêts à mettre un certain enjeu dans le pari, enjeu à la mesure de leur conviction sur leurs chances de le gagner.

Beaucoup de gens ne veulent prendre des paris que lorsqu'ils sont assez sûrs de gagner. On le comprend aisément, et ils rejoignent alors le sens courant du terme « je parierais bien que telle ou telle chose va se produire » c'est d'ailleurs souvent une façon, de s'engager avec vigueur envers telle ou telle chose, une façon de parler.

C'est pourquoi le parieur occasionnel refusera, et il s'offusquera même de la situation, de s'engager dans un pari sur ce qui est contraire à ses opinions (on pense aux opinions politiques) même s'il était sûr de gagner l'enjeu, ou à ce qu'il perçoit comme contraire aux opinions présumées de ses interlocuteurs. Et là il sera particulièrement offusqué.

Retenons que le parieur occasionnel pensera naturellement en termes de probabilité.

Le parieur invétéré ou professionnel ne pense pas, lui, du moins pas habituellement, en termes de probabilité puisqu'il pense toujours enjeu : qu'est-il prêt à mettre sur la table ? Il est toujours prêt à engager des sommes pour gagner. Trouver le jobard qui entrera dans son jeu, c'est une forme de travail. Il s'agit aussi pour certains d'épicier la chose, une façon de susciter la discussion, animée de préférence puisqu'il y a un enjeu, de créer un dialogue avec l'autre pour creuser la question.

Le pari peut même devenir une forme de perversion, les casinos, les loteries de diverses nature exploitent bien ces tendances, mais cela est une autre question qui se rapporte à l'enjeu du pari de Pascal, à l'espérance de gains importants ...sur terre comme au ciel!

Puisqu'il pense enjeux avant tout, le parieur invétéré se dira non pas j'ai deux trois chances sur quatre de gagner, et donc une sur quatre de perdre, mais je donnerais du 3 contre 1 sur ce pari, ou du un contre deux lorsqu'il estime qu'il y a environ une chance sur trois de la gagner. Le jeu consistera à trouver la personne qui, elle, est prête à mettre de l'argent au pair dans le premier cas, ou encore à une proportion moindre que du trois pour un, limite qu'en principe le premier parieur ne voudra pas dépasser.

En réalité, le parieur professionnel évaluera naturellement une situation de pari en termes de *cotes*, mais il sera prêt à négocier sur d'autres cotes que celle qu'il estime au meilleur de sa connaissance : il reste en effet toujours une marge d'incertitude au pari, cela en fait tout son sel. Appelons cette cote la cote véritable pour la distinguer de la cote de négociation. Les cotes de négociation sont destinées à tromper l'adversaire... Le pari ne pourra se conclure en effet que lorsque chacun des parieurs aura l'impression que la cote du pari est en sa faveur, lui laissant une espérance de gain positive¹.

Retenons dans ce cas que ce genre de parieur fonctionne avec des cotes en tête².

Ces cotes s'appellent *odds* en anglais³ : « *an allowance granted by one making a bet to one accepting the bet and designed to equalize the chances favoring one of the betters* ». Il s'agit bien sûr de la perspective honnête du terme...

¹Le terme d'espérance est à prendre ici dans son sens courant et non pas technique, même s'ils se rejoignent...

²À notre connaissance, aucun dictionnaire français ne donne ce sens au mot cote. On ne rapporte que l'expression « la cote d'un cheval » sans en donner la moindre définition. Mais tous les parieurs sur courses de chevaux pensent en ces termes, en comprenant le sens, au moins intuitivement : « ce cheval paye combien ? ».

³Selon entre autres le dictionnaire Merriam-Webster cité ici, très précis sur la question.

1.2 La perspective du probabiliste

Les probabilités sont nées de la frivolité⁴ : Blaise Pascal [1623-1669], quelque part dans la partie mondaine de sa vie, fut appelé à se prononcer sur une question d'un parieur professionnel et grand libertin, le Chevalier de Méré, et inventa les concepts de base qu'on trouve dans tous les cours de probabilité élémentaire.

On ne mentionne pas souvent les origines des concepts, notamment ce qui a trait aux paris. Et pourtant, tout est là... On cliquera sur l'icone à droite pour un aperçu de la naissance de la discipline dans la frivolité.

Considérons un pari A : ce n'est rien d'autre qu'un essai de Bernoulli où on observera lors de sa réalisation s'il se vérifie ou non, si, en termes d'une *expérience aléatoire*, ce concept de base de la probabilité, A se réalise ou pas. On suppose que A se produit avec une probabilité $P(A) = p$, et donc que la probabilité qu'il ne se produise pas est $P(\bar{A}) = 1 - p$.

Pascal



Définition.

On définit la *cote de probabilité de p* , ou encore la *cote de p* tout simplement, notée $CP(A) \equiv CP(p)$, pour un événement A de probabilité p par le rapport des aléas

$$CP(A) \equiv CP(p) = \frac{p}{1 - p}.$$

C'est ce rapport qui en effet constitue bien la cote d'un pari⁵. On trouve à la FIG. 1, le graphique des CP en fonction de la probabilité estimée de gagner le pari.

⁴On simplifie un peu ici les circonstances d'une naissance plutôt difficile qui fit intervenir plusieurs éminents personnages.

⁵On trouve parfois aussi pour la cote de p le nom *rapport des risques*, ou encore le *rapport des aléas*, et même sans plus de façon, surtout en français hexagonal, le « *odds* » en anglais dans le texte... La cote a l'avantage d'être comprise de tous les parieurs et celui de la concision.

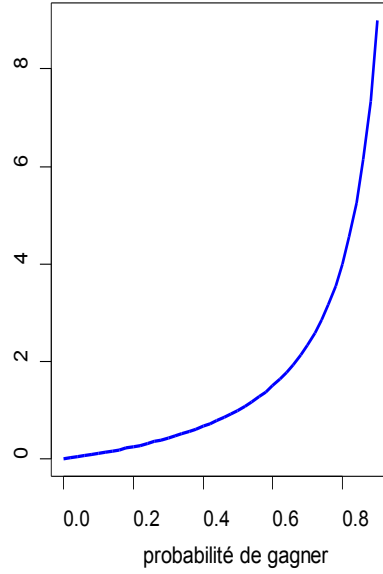


FIG. 1 – En abscisses les probabilités de gagner, en ordonnées les cotes associés.

Exemples. Le passage de la cote du pari à la probabilité de son succès est immédiat : ainsi une cote de 3 contre 1 pour le pari A , correspond à :

$$CP(A) = \frac{3/4}{1/4}.$$

Une cote de 3 contre 2 correspond à une probabilité $P(A) = 3/5$. Une cote de 1 contre 1 correspond à la probabilité $p = 0,5$: les chances de gagner et de perdre sont égales⁶. Des cotes fractionnaires correspondent à des événements de probabilités inférieures à une demie : $p < 0,5$.

Calculons l'espérance de la loi de Bernoulli associée⁷ au pari A qu'un parieur serait prêt à prendre à n contre m : si A se réalise, le parieur gagne un certain nombre de fois m unités dans la monnaie du pari, et perd le même nombre de fois n unités dans le cas contraire. On calcule l'espérance $E(A)$

⁶En anglais on dirait *the odds are even*, ce qui pourrait prêter à quelque confusion...

⁷Par métonymie, on confondra A avec la loi comme on le fait avec le pari qui lui est associé.

de ce pari :

$$p = \frac{n}{n+m} \implies E(A) = mp - n(1-p) = m \frac{n}{n+m} - n \frac{m}{n+m} = 0.$$

Le même calcul fait par l'autre parieur lui donne également la même espérance. Les gains espérés sont nuls, le pari est dit *honnête*⁸. Le pari se prendra pour le sport... On retrouve ici exactement la définition du Merriam-Webster du terme anglais *odds* pour la cote, au sens propre ou véritable, d'un pari.

1.3 Retour au parieur

Voilà maintenant un petit ajout à la confusion sémantique ! Dans les systèmes de paris organisés, par exemple les courses de chevaux, on trouve un concept de *cote*, et c'est la seule acception en ce sens du terme en français. Mais ce n'est pas exactement le concept défini plus haut, c'en est plutôt l'inverse ! Et ce n'est pas celui non plus dans le sens courant du terme en langue anglaise tel que rapporté plus haut.

Supposons qu'on ait une course de deux chevaux et que, pour simplifier, tous les paris encourus sont remis aux parieurs. En réalité, l'organisation prend un pourcentage des paris pour son bon office, mais tout le reste de l'argent parié est remis aux parieurs.

Supposons maintenant que le cheval *A* reçoive la faveur populaire et que, des trois milliers de parieurs ayant misé chacun une unité de monnaie, deux mille ont parié sur lui, le troisième millier ayant parié sur *B*.

En termes probabilistes, le cheval *A* a une cote de 2 (contre 1), puisque l'opinion populaire, la masse ne peut se tromper c'est entendu..., donne une probabilité de 2/3 au cheval *A* et de 1/3 au cheval *B* : $CP(A) = 2$.

Si *A* gagne, les parieurs sur lui se partagent la mise : ils retrouvent chacun le montant de leurs paris, soit une unité de monnaie, plus une demie unité de monnaie chacun provenant de ceux qui ont parié sur *B* : ils gagnent une unité de monnaie pour chaque deux unités engagées. Le cheval *A* paye du 1 pour 2, c'est sa cote au sens des parieurs du terme : *A* est favori à 1 contre 2. Inversement, si *B* gagne, chaque parieur sur lui aura gagné deux unités de monnaie provenant des parieurs sur *A* qui ont perdu : le cheval *B* a une cote de 2 contre 1.

⁸Ce que le parieur professionnel cherche à éviter !

Ces cotes sont le fait de l'organisation du système des paris, ceux qui ont parié sur A restent, quant à eux, certains à 2 contre 1 de gagner leurs paris au sens probabiliste du terme... L'organisation, elle, paye bien une unité contre deux engagé si A gagne, et inversement pour B .

Cette confusion entre le sens probabiliste et celui de l'organisation n'est en fait qu'une question de point de vue, celui du parieur et celui du payeur l'un étant l'inverse de l'autre, comme si ce n'étaient pas les mêmes. Ça se comprend tout de même sans problème.

Pour terminer, quelle est dans ce cas l'espérance de ce jeu ? Admettons que le cheval A ait effectivement 2 chances sur 3 de gagner ainsi que l'a estimé la faveur populaire.

Pour un parieur sur A : il gagne une demie unité avec probabilité $2/3$, et perd une unité avec probabilité $1/3$. Son espérance est donc :

$$E(A) = \frac{1}{2} \times \frac{2}{3} - 1 \times \frac{1}{3} = 0.$$

Pour un parieur sur B , on trouve aussi $E(B) = 0$ par un raisonnement analogue. C'est un jeu honnête, et la cote annoncée sur les chevaux satisfait encore à la définition du Merriam-Webster !

Il est à noter que les paris sur les chevaux, les loteries gouvernementales ou autres, ne sont jamais honnêtes : leur espérance pour le parieur est forcément négative puisque les organisateurs prennent toujours des frais d'administration. Mais pourquoi alors, et c'est surtout le cas pour les loteries, trouve-t-on autant de parieurs ?

C'est le concept d'espérance qui est la clé de la réponse, et on le voit illustré de façon saisissante dans le pari de Pascal : la probabilité du gain peut bien être quasi nulle, un epsilon non négatif aussi petit qu'on veut, mais lui est associée un gain infini, soit le paradis, alors que le coût de l'alternative, à savoir que Dieu et son paradis n'existent pas, est une bagatelle. Le calcul de l'espérance de ce pari est infini ! Et en plus, parier sur Dieu fait faire une bonne vie...

Dans toutes les loteries, une somme ridicule donne une toute petite chance de gagner au moins toute une vie de salaire, sinon plusieurs, voire une infinité de vie de bonheur extatique comme dans celle de Pascal. Alors, un petit pari quelqu'un ?

2 Modéliser : le modèle logistique

Le parieur n'a que son intuition à sa disposition : les événements sur lesquels il parie sont uniques. Le statisticien au contraire peut compter sur des données : la variable binaire Y a été observée disons n fois, qu'on suppose indépendantes : $Y_i = y_i = 0$ ou 1 , $i = 1, \dots, n$. On pourrait simplement utiliser les estimateurs habituels de proportion pour caractériser complètement la binômiale, et prévoir si on en a besoin la valeur de la prochaine réalisation.

Mais, tout comme en modélisation linéaire habituelle, on aimerait améliorer la qualité des paramètres et des prédictions, et peut-être que les informations supplémentaires sur les conditions de réalisation des Y_i permettent de mieux expliquer l'incidence des valeurs réalisées : ainsi la probabilité de réalisation de l'événement serait mieux estimée avec la connaissance des valeurs des variables explicatives.

Exemple. Le cas aujourd'hui devenu classique pour l'illustration de la méthodologie qu'on présente dans ce qui suit est celui des sceaux d'étanchéité (*O-rings*) utilisés dans les propulseurs auxiliaires de la navette spatiale états-unienne *Challenger* qui ont fait défaut et causé son explosion le 28 janvier 1986, entraînant la mort des cosmonautes présents à bord et causé un onde de choc considérable dans la population⁹. La Nasa a failli ne pas s'en relever¹⁰ !

La Commission présidentielle chargée de l'affaire avait invité le physicien **Richard P. Feynman**¹¹ à y siéger. Cet esprit libre et inquisiteur découvrit la raison du mauvais fonctionnement des *O-rings* de la navette. Les ingénieurs de la compagnie qui avait produit les sceaux (Morton Thiokol) soupçonnaient

⁹Entre autres sites qui rapportent les détails, on pourra consulter les suivants : <http://www.nth-degree.com/regeg.html>, ou encore <http://www.math.yorku.ca/SCS/Gallery/missed.html>, et, pour ce qui est des difficultés d'aller à contre-courant, <http://www.onlineethics.com/index.html>.

¹⁰On peut même se demander si cette affaire ne fut pas un argument massue contre la « *Big Science* » et la technologie en général qui commençait à prendre de l'ampleur à l'époque : pour faire bref, quand on craint l'épidémie du Sida, cette nouvelle peste (on est en 1986), « à quoi sert de conquérir l'Univers ? »

¹¹[New York 1918- Los Angeles 1988], Prix Nobel de physique en 1965, et personnalité médiatique bien connue à cause de ses talents de vulgarisateur. On lira dans ce site certains documents relatifs à cette affaire, notamment l'annexe F, particulièrement cinglante, de la main de Feynman lui-même.

bien le problème : à l'examen des sceaux d'étanchéité recueillis sur la propulseurs auxiliaires des lancements précédent, ils avaient fini par croire que les lancements pouvaient être dangereux à basse température. Mais c'est Feynman qui prouva hors de tout doute la chose : la matière utilisée dans les sceaux devient trop rigide à basse température pour jouer son rôle. La veille du départ, les ingénieurs avaient tout fait pour empêcher le lancement. Peine perdue, onze heures de discussion avec les gestionnaires de la Nasa — oui, onze, du moins selon le rapport de la Commission —, n'ont pas suffi à les convaincre. Le lancement eut lieu avec les conséquences que l'on sait.

Une des raisons de ce manque de conviction de la part des ingénieurs fut probablement qu'ils n'avaient pas d'arguments chiffrés bien percutants : on a discuté de bien des choses à partir d'une présentation qui ne montrait jamais graphiquement le lien entre la température et les défauts observés des *O-rings* des lancements précédents. Le manque de connaissance des ingénieurs (de l'époque) en modélisation statistique pourrait aussi être invoquée...

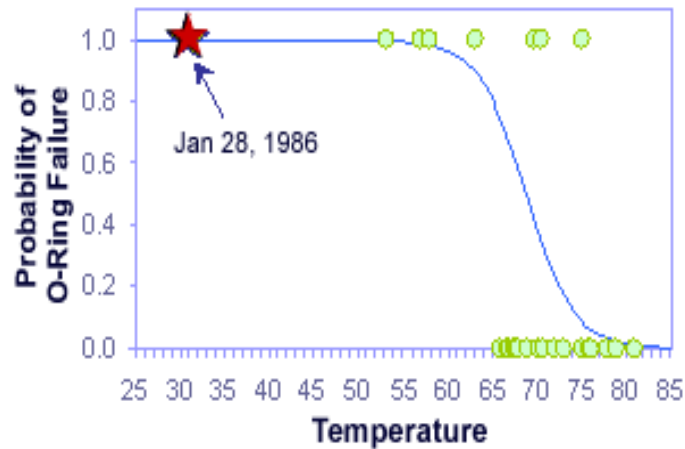


FIG. 2 – Les données et le modèle logistique ajusté pour les *O-rings* des navettes spatiales américaines avant la date fatidique. Le code 1 est pour la défaillance, 0 pour le bon fonctionnement. La température le jour du lancement fatidique était de 31 °F.

Eussent-ils eu la modélisation logistique dans leur coffre à outils ainsi qu'un simple graphique¹² tel celui de la FIG. 2, les gestionnaires eussent-ils

¹²Tiré du premier site Internet cité plus haut. L'ignorance n'est pas toujours le bonheur !

été moins portés sur la productivité avec l'œil rivé sur la ligne budgétaire rétrécissante de la Nasa, que « la face du monde eût pu en être changée... »

Exercices. On trouvera en cliquant l'icône à droite le fichier des données du bon ou mauvais fonctionnement des sceaux d'étanchéité, tels qu'observées sur les propulseurs auxiliaires retrouvés lors des lancements précédents des navettes spatiales.



Data

1. Importer les données dans un logiciel de traitement statistique, et obtenir un graphique semblable à celui de la FIG. 2.
2. Sur la base de ces observations, uniquement de ces observations sans y superposer mentalement la courbe logistique observée la FIG. 2, seriez-vous porté aisément à croire que l'effet de la température est important, et qu'à moins de 40 ° F, on risquait un problème grave ?

Note : On ne comprendra le sens à accorder au graphique de la logistique que plus loin. Contentons-nous pour l'instant de dire que les points sur la courbe sigmoïde (inverse ici) sont des probabilités estimées de défaillance en fonction de la température. Mais oublions tout cela pour répondre à la question précédente.

Supposons qu'on a des facteurs explicatifs d'une réponse binaire. On ne peut donc utiliser les ajustements aux moindres carrés des modèles linéaires usuels puisque la réponse n'est pas continue. Mais on aimerait bien pouvoir montrer que certains facteurs explicatifs hypothétiques jouent bien leur rôle de facteurs explicatifs, alors que d'autres sont sans effet.

On recourt alors à une transformation. Et on pense justement aux paris, et aux rapports des aléas, aux cotes de probabilité. La variable à expliquer étant binaire, elle correspond à un essai de Bernoulli $Y \sim \mathcal{B}(1; p)$. Plutôt que d'ajuster cette variable discrète, on ajuste le logarithme du rapport des aléas de Y . Ainsi, dans le cas d'une seule variable explicative X , on ajuste le modèle

$$\log(\text{CP}(p)) \equiv \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X. \quad (1)$$

Le logarithme d'une cote de p s'appelle son *logit*, ou simplement le logit. Ce qui donne immédiatement $\text{CP}(p)$:

$$\text{CP}(p) \equiv \frac{p}{1-p} = e^{\beta_0 + \beta_1 X}. \quad (2)$$

Enfin, un peu d'algèbre donne l'expression suivante pour p :

$$p = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}. \quad (3)$$

Et p s'exprime en termes d'un *modèle logistique* en fonction de X . D'où le nom éponyme pour le modèle dans le cas d'une réponse binaire. En fait, le passage de l'équation (1) à l'équation (3) n'est autre que celui d'une fonction à son inverse (bien comprendre cela). La généralisation à plus d'un facteur explicatif est immédiate. La fonction qui lie X à p est dite le *lien*. D'autres sont possibles.

On comprend bien sur la FIG. 3 pourquoi le modèle logistique peut être utilisé pour modéliser des réponses binaires. En fonction des valeurs de β_1 les valeurs estimées de p passent plus ou moins rapidement de 0 à 1 ($\beta_1 > 0$), ou de 1 à 0 ($\beta_1 < 0$) selon le sens de l'effet de la variable explicative.

À l'issue de l'étape de l'estimation de la procédure, on obtient les valeurs estimées de p sur chaque observation par le modèle, les \hat{p} , en fonction des variables explicatives. Les valeurs observées sont les valeurs de la binomiale Y , donc soit des 0 soit des 1 ; les valeurs estimées sont, elles, entre 0 et 1.

Si on est intéressé à la prévision dans l'application des résultats de la modélisation logistique, on aura à déterminer, le modèle ayant été validé, une règle de décision : en deçà de quelle valeur \hat{p} conviendra-t-on de dire que la réalisation estimée est $\hat{Y} = 0$, au delà de laquelle on dira que $\hat{Y} = 1$? En d'autres termes : on devra, toujours dans le cas où on veut faire des prédictions, fixer un seuil \hat{p}_0 et la règle de décision qui s'y rattache.

Remarque. On remarque que lorsqu'on dispose de plusieurs observations pour chaque valeur des variables explicatives — on peut même regrouper leurs valeurs voisines à cet effet au besoin —, il est possible d'estimer pour la variable explicative binaire une proportion échantillonnale \hat{p} pour chacune d'entre elles, et observer le graphique des (x_i, \hat{p}_i) . Celui-ci devrait avoir une forme sigmoïdale comme l'indique l'équation (3). Voir les exercices.

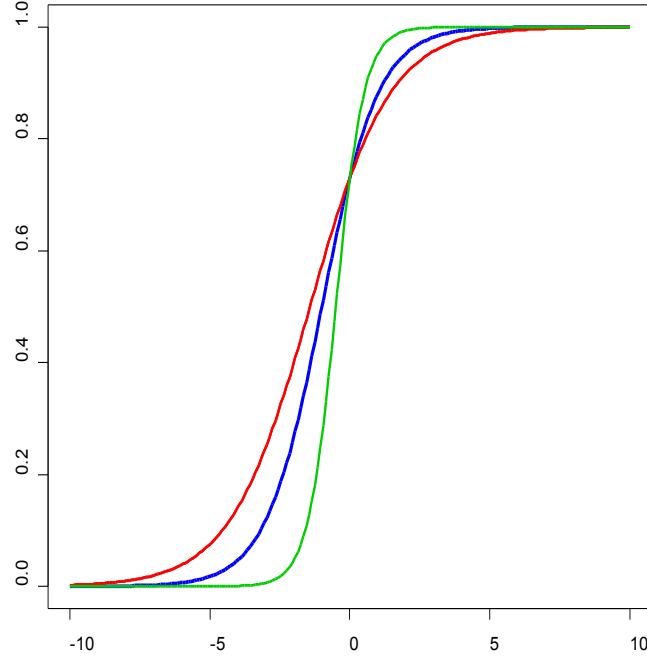


FIG. 3 – La fonction logistique donnant le *lien* pour p , en ordonnées, en fonction de X dont les valeurs, purement à titre indicatif, sont en abscisses. Avec $\beta_0 = 1$, et en rouge $\beta_1 = 0,7$; en bleu : $\beta_1 = 1$; en vert $\beta_1 = 2$.

Enfin, on tire de l'équation 2 l'interprétation de l'augmentation marginale de la valeur de X : l'effet n'est pas linéaire avec β_1 , mais multiplicatif dont la valeur est e^{β_1} . Ainsi donc, c'est le $CP(p)$ qui change par le facteur e^{β_1} .

En effet, notons $CP(1)$ la valeur du rapport des aléas pour une certaine valeur de $X = x_i$, notons $CP(2)$ la quantité analogue pour la valeur $X = x_i + 1$. L'équation (1) nous donne

$$\log \left(\frac{CP(2)}{CP(1)} \right) = \log (CP(2)) - \log (CP(1)) = \beta_1 ,$$

ce qui implique donc :

$$CP(2) = CP(1) e^{\beta_1} .$$

L'effet de l'augmentation *marginale* de X est de facteur e^{β_1} est bien multiplicatif sur les cotes des probabilités estimées, les rapport des aléas estimés.

Définition.

Le quotient de deux cotes, $CP(1)$ et $CP(2)$, s'appelle le *rapport des cotes*. On le note RCP :

$$RCP = \frac{CP(1)}{CP(2)}.$$

En anglais on utilise communément le terme *odds-ratio* pour ce rapport des cotes, et, rappelons-le, le terme *odds* pour la cote de p ou le rapport des aléas¹³. On a rarement besoin de préciser notationnellement les valeurs sous-jacentes des probabilités.

Exercices.

1. Supposons qu'on a obtenu les estimations suivantes $\hat{\beta}_0 = -1,705$ et $\hat{\beta}_1 = 4,007$, $\hat{p}(x_i) = 0,154$ pour une certaine valeur x_i de X , quelle est le changement dans $CP(0,154)$ lorsque X augmente de 1 ? Quelle est la nouvelle valeur estimée $\hat{p}(x_i + 1)$?
2. Obtenez l'effet sur la cote de probabilité de l'accroissement de X_i à un accroissement quelconque *i.e.* le changement de $X_i = x_i$ à $X_i = x_i + a$. L'accroissement marginal qui a été explicité plus haut et qui donne l'interprétation du coefficient β_1 est le cas particulier où $a = 1$.

2.1 Le modèle logistique multivarié

La généralisation au modèle multivarié se fait aisément. Il suffit essentiellement d'un changement de notation pour bien faire voir les variables en jeu. Supposons qu'on a une suite X_1, X_2, \dots, X_k de variables explicatives de

¹³Si on préfère le terme *rapport des aléas* à *cote*, pour des raisons de précision de langage, le *odds-ratio* devient alors le bi-rapport des aléas, puisque deux rapports sont mis en rapport...

la réponse binaire $Y \sim \mathcal{B}(1; p)$. Le modèle logistique à plusieurs variables est donc :

$$\log(\text{CP}(Y | X_1, \dots, X_k)) = \beta_0 + \beta_a X_1 + \dots + X_k, \quad (4)$$

où on a noté bien sûr $\text{CP}(Y | X_1, \dots, X_k) = p/(1 - p)$.

L'interprétation des coefficients β_i est la même que dans le cas univarié, en ajoutant la provision : « toutes les autres variables étant égales par ailleurs ».

Ainsi donc, le rapport des cotes estimées, le RCP, lors de l'augmentation marginale de la valeur de X_i , de $X_i = x_i$ à $X_i = x_i + 1$, toutes les autres variables explicatives étant égales par ailleurs, vaut e^{β_i} : les deux cotes sont proportionnelles.

2.2 Estimer le modèle

Remarquons qu'on n'a aucune justification théorique permettant d'utiliser dans ce cas une régression aux moindres carrés. Il faut procéder par la méthodes de la maximisation de la vraisemblance de l'échantillon des $Y_i \sim \mathcal{B}(1; p)$, $i = 1, \dots, n$.

Les Y_i étant binaires, on a :

$$P(Y_i = 1) = p_i \quad P(Y_i = 0) = 1 - p_i,$$

et on peut écrire pour les densités des Y_i :

$$f_i(Y_i) = p_i^{Y_i} (1 - p_i)^{(1-Y_i)} \quad i = 1, \dots, n.$$

La vraisemblance de l'échantillon est donc, par indépendance des observations :

$$L(\beta_0, \beta_1) \equiv g(Y_1, \dots, Y_n) = \prod_{i=1}^n f_i(Y_i) = \prod_{i=1}^n p_i^{Y_i} (1 - p_i)^{(1-Y_i)},$$

et son logarithme est :

$$\log(g(Y_1, \dots, Y_n)) = \sum_{i=1}^n Y_i \log\left(\frac{p_i}{1 - p_i}\right) + \sum_{i=1}^n \log(1 - p_i). \quad (5)$$

Maintenant on tire de l'équation (3), dans le cas d'une seule variable explicative, que

$$1 - p_i = [1 + e^{(\beta_0 + \beta_1 X_i)}]^{-1}, \quad (6)$$

et, finalement on substitue cette dernière équation (6) ainsi que (2) dans (5), et on change la notation de la log-vraisemblance pour faire ressortir les paramètres :

$$\log(L(\beta_0, \beta_1)) = \sum_{i=1}^n Y_i(\beta_0 + \beta_1 X_i) - \sum_{i=1}^n \log(1 + e^{(\beta_0 + \beta_1 X_i)}). \quad (7)$$

Les observations sont données : (x_i, y_i) , $i = 1, \dots, n$, on fait appel aux méthodes numériques pour déterminer les valeurs (b_0, b_1) qui maximisent la vraisemblance de l'échantillon.

Ces valeurs estimées des paramètres permettent de déterminer les \hat{p}_i pour chaque observation par l'équation (3).

Le passage au cas de multiples variables explicatives se fait sans problème.

Remarques.

1. Notons pour en terminer avec l'estimation que les variables explicatives peuvent être discrètes tout autant que continues.
2. L'interprétation de l'effet des coefficients b_i suit la ligne déjà explicitée pour une seule variable explicative avec une provision supplémentaire. Soit β_i le coefficient de la variable X_i , estimé par b_i : lorsque toutes les autres variables sont fixes, une croissance d'une unité de la variable X_i donne $\exp(b_i)$ pour le rapport des cotes *estimées*, les deux cotes estimées sont proportionnelles.

Exercices. On ne sait pas encore exactement tester la validité et la pertinence des modèles logistiques. On admettra dans ce qui suit qu'on se trouve dans la situation où les variables explicatives le sont bel et bien, et que les modèles eux-mêmes sont valides. On explorera le détail plus loin.

1. On se rapporte aux données des sceaux d'étanchéité des navettes spatiales états-uniennes. Écrivez les équations du modèle logistique simple approprié (référez-vous aux équations (1), (2), (3)).
 - (a) Utilisez un logiciel de traitement statistique pour estimer les paramètres d'un modèle logistique utilisant la température pour expliquer les défaillances des *O-rings*. Le logiciel devrait vous annoncer que le modèle est valide et que la valeur estimée du coefficient

β_a , soit b_1 , est significativement non nulle. Quel sens donnez-vous à la valeur estimée de β_1 ?

- (b) Recodez maintenant à l'inverse la variable *Défectuosité* : 0 devient 1 et inversement. Recommencez l'analyse de la question précédente. Utilisez des codes quelconques et examinez encore les résultats. Quels changements observez-vous ? Que concluez-vous de ces changements sur le sens que donne le logiciel à la probabilité p , et à la définition de la binomiale modélisée. L'interprétation des rapports des cotes est-elle identique ?
- (c) Obtenez les probabilités estimées pour chaque observation dans les premier et deuxième codages mentionnés à la question précédente. Comparez les deux suites d'estimations.
- (d) En vous reportant au modèle estimé dans l'un ou l'autre codage — qu'on suppose valide —, estimez la probabilité de défectuosité d'un sceau alors que la température est celle du vol fatal, soit 31 ° F. Estimez, toujours sous hypothèse de validité du modèle, en deçà de quelle température la probabilité de défectuosité d'un vol dépasse-t-elle 0,95, 0,99.

On ne sait pas encore vraiment comment choisir (décider) si une température conduit à une défectuosité, en d'autres termes quelle est la précision sur les valeurs estimées \hat{p} . Tout de même, les réponse à l'item précédent donne une certaine idée, pour ne pas dire une idée certaine..., suffisante en tout cas pour avoir de bons arguments pour vouloir faire stopper une mission dont le départ se déroule à moins de 40 ° F !

Mais on pense ici en termes probabilistes, et donc on se demande quelle précision attribuer aux estimations précédentes. Pour cela, on remarque que le logarithme est une fonction croissante, de même que $CP(p)$ en fonction de p , ce qui entraîne que le logit est une fonction croissante. On détermine, par dérivation si nécessaire, ou simplement par l'examen des termes du modèle logistique, c'est bien plus simple car le modèle est linéaire dans les coefficients β_i , quelles sont les valeurs maximales et minimales du logit, en fonction d'un jeu de valeurs données pour les variables explicatives, selon les valeurs maximales et minimales des coefficients du modèle.

On peut calculer ensuite, à l'aide des intervalles de confiance des coefficients qui vous sont donnés par tout bon logiciel, un intervalles de

même confiance pour la valeur du logit de p estimée, et donc par transformation un intervalles de confiance pour la valeur estimée de $CP(p)$ pour n'importe quelles valeurs des variables explicatives.

- (e) À l'aide de la procédure qu'on vient d'expliquer, déterminez un intervalle de confiance à droite de confiance 95% pour $CP(p)$.
- (f) Auriez-vous été prêt à parier, sur la base des calculs précédents, qu'il n'y aurait pas eu de défectuosité (avec les conséquences prévisibles qu'on connaissait bien...) sur le bon fonctionnement des sceaux d'étanchéité à 31 ° F. Au besoin passez de la cote à la probabilité de ce pari¹⁴.

2. Dans une expérience destinée à tester l'effet de la pression sur la qualité de la fixation de la teinture lors de la fabrication de billes de plastique, on a testé 250 billes à chacun de six niveaux de pression sur une échelle logarithmique à base deux comptée à partir d'une pression de référence. On a donc des pressions de 1 à 6, ces chiffres sont sans unités. Chaque bille a été examinée et on a noté 0 : une bille non conforme ; 1 : une bille conforme. On a pris soin d'aléatoriser les billes affectées aux groupes, cela vous semble-t-il une précaution importante ?

En cliquant sur l'icone à droite on importe le fichier Excel correspondant comprenant 3 feuilles : (1) les données regroupées, *i.e.* comprenant les données synthétisées à partir de (2) les données originales, et (3) le descriptif des variables.



Data

- (a) Considérez les données regroupées. Pouvez-vous croire que les probabilités de défectueux selon la pression suivent une forme logistique. Par ailleurs, examinez les graphiques, toujours en fonction des 6 niveaux de pression, des $CP(i)$ ainsi que des $\log(CP(i))$. L'ajustement du modèle logistique semble-t-il valable ?
- (b) Obtenez l'estimateur sans biais du p de la binômiale associée la variable NC. Cet estimateurs ne tient nullement compte de la variable explicative des données.

¹⁴Il vas de soi qu'aucune compagnie un tant soit peu responsable ne serait prête à courir des risques qui entraînent des conséquences graves, — et la morts d'astronautes, qui est un des enjeux ici et pas le moindre, en est sûrement une ! — à moins que leurs probabilités estimées ne soient inférieures de plusieurs ordres de grandeur à celles qu'on calcule ici !

- (c) À l'aide d'un logiciel statistique, faites l'ajustement du modèle, et interprétez l'effet (naturellement statistiquement significatif) de la pression sur la probabilité d'obtenir une bille conforme lorsqu'on change la pression de la valeur 1 à la valeur 2, puis de 1 à la valeur 4.

- 3. On trouvera en cliquant l'icone à droite les résultats d'une expérience non planifiée dans une nouvelle société de production artisanale de fromage.

L'affinement du fromage se fait dans deux caves, notés 1 et 2, où les conditions d'humidité et de température sont très stables, mais différentes de l'une à l'autre. On dispose aussi du lait de trois producteurs distincts, notés de 1 à 3. Les fabrications ont commencé pour les trois laits dans la même semaine. On a mesuré la qualité du fromage des meules choisies au hasard selon la durée en mois, ou *âge*, du fromage, du passage dans la cave. On admet que le vieillissement est très homogène dans chacune des caves, et la qualité est par deux indices : le premier ('Conforme') est un indice général de conformité aux normes sanitaires; le second est un indice de conformité dû à la couleur du fromage, un critère important (hélas!) pour bien des consommateurs.

L'expérience ne fut pas réellement planifiée, c'est une première étape dans l'analyse du procédé de fabrication, et on se demande si la cave, l'origine du lait, de même que la durée de l'affinage, ont une influence sur la qualité. Les producteurs ont enregistré plusieurs autres mesures dont nous ne parlerons pas ici.

- (a) Obtenez par types de lait et selon la cave d'affinage, les nombres de fromages conformes et non conformes. Les caves vous semblent-elles équivalentes ? Mêmes questions pour les types de lait.

Pour des raisons qu'on explique sur la page 'Codes' du fichier Excel, créez les nouvelles variables qu'on utilisera dans les analyses subséquentes.

- (b) Obtenez les paramètres du modèle logistique pour expliquer la conformité des fromages par les valeurs des variables 'Âge', 'Type de lait' (recodé) ainsi que la 'Cave' (recodée elle-aussi). Vous
- (c) Utilisez le critère des déviations pour confirmer que les valeurs des paramètres associés aux divers laits ne jouent pas de rôle.



Data

Mais malgré ces constatations de fait, le producteur est certain que les laits font une différence, et il est décidé de garder les paramètres associés aux types de lait dans le modèle.

- (d) Calculez les changements estimés des cotes associées à la qualité lorsqu'on passe d'une cave à l'autre. De même pour les types de lait (les trois RCP associés estimés par le modèle). Vous exprimerez ces nombres en pourcentage.
- (e) Utilisez l'intervalle de confiance calculé par le logiciel pour le paramètre correspondant qui décrit le changement dans les RCP pour en déterminer un intervalle de confiance, lorsqu'on passe de la première cave à la seconde.

Afin de sécuriser vos conclusions sur le modèle logistique obtenu, créez une variable 'Rnd(1)' et recommencez quelques fois vos analyses avec seulement aux environs d'une moitié des observations¹⁵.

- (f) Énoncez vos conclusions ainsi que les raisonnements pour y parvenir.

Maintenant créez une nouvelle variable pour l'indice global de qualité par la multiplication des deux indices partiels. Vous établirez les paramètres du modèle sur les 98 premières observations seulement. Celles-ci constituent un bon échantillon test pour le modèle, puisqu'on a pris la précaution d'aléatoriser la cueillette des données. Vous utiliserez le reste des observations pour valider votre choix '0 ou 1' pour la qualité globale des fromages selon que la probabilité estimée $\hat{\pi}_h$ est en-deça ou au delà du seuil π_0 que vous aurez déterminé.

- (g) Donnez une justification pour la définition de cet indice global de qualité.
- (h) Tirez des calculs des paramètres du modèle logistique sur les variables issues du recodage pour les 'Types de lait', ainsi que pour les caves (recodées) pour décider quelle est la combinaison qui maximise la probabilité d'un fromage conforme selon cet indice global de qualité.

¹⁵Pour ce faire, vous générez quelques fois à neuf les nombres aléatoires...

Il est à noter que même si la prédiction pour les nouvelles observations est un objectif qu'on trouve parfois, il est rare que le succès soit très bon dans ce genre de prédiction...

Mais l'objectif était avant tout d'identifier des facteurs d'influence sur une réponse binaire. Toujours en modélisation, ces deux objectifs : décrire et prédire. Une bonne description n'est pas un gage de grand pouvoir de prédiction...