CHAPITRE 3

Introduction aux modèles linéaires généralisés

MTH3302 - Méthodes probabilistes et statistiques pour l'I.A.

Jonathan Jalbert – Hiver 2023

Les modèles linéaires généralisés constituent des extensions du modèle de régression linéaire étudié au chapitre précédent. Ils permettent de modéliser des variables d'intérêts discrètes, telle que des variables de type Bernoulli, binomiale et Poisson, ainsi que d'assouplir les hypothèses de linéarité et de normalité des erreurs. La régression logistique, qui est un type de modèle linéaire généralisé pour les variables aléatoires de type Bernoulli, est abondamment utilisée en apprentissage machine pour classer les observations en deux catégories, également appelé *clustering*.

Ce chapitre présente la base de la théorie des modèles linéaires généralisés. À la fin du chapitre, vous devriez être en mesure de

- Écrire le modèle linéaire généralisé pour une variable d'intérêt distribuée selon la loi de Bernoulli, binomiale et la loi de Poisson.
- Estimer les paramètres de ces modèles par la méthode du maximum de la vraisemblance à l'aide d'un logiciel tel que Julia.
- Évaluer la qualité d'un modèle linéaire généralisé.
- Effectuer la sélection des variables explicatives.

Dans ce chapitre, les modèles linéaires généralisés seront implémentés pour étudier l'effet des caractéristiques des passagers du Titanic sur leur probabilité de survie. Rappelons que lors du naufrage du Titanic en 1912, entre 1490 et 1520 personnes sont disparues parmi les 1316 passagers et 889 membres d'équipage à bord. Nous nous intéresserons qu'aux 1309 passagers pour lesquels nous avons des informations. Ces passagers ont été scindés en deux

groupes : 872 passagers dans l'ensemble d'entraînement et 437 dans l'ensemble de test. Le but sera d'*apprendre* des 872 passagers de l'ensemble d'entraînement pour correctement prédire le sort des 437 passagers de l'ensemble de test.

3.1 Lorsque la variable réponse est de type Bernoulli

Lorsque la variable d'intérêt Y_i ne prend que les valeurs dans l'ensemble $\{0, 1\}$, la distribution naturelle pour Y_i est la loi de Bernoulli :

$$Y_i \sim \mathcal{B}ernoulli(\theta_i); \tag{3.1}$$

où θ_i correspond à la probabilité de succès. L'espérance de la loi de Bernoulli est égale à la probabilité de succès : $E(Y_i) = \theta_i$. L'espérance se situe donc dans l'intervalle $(0, 1)$.

Exemple 1

Dans l'exemple des passagers du Titanic, posons Y_i la variable aléatoire modélisant la survie du passager i :

$$Y_i = \begin{cases} 1 & \text{si le passager } i \text{ a survécu,} \\ 0 & \text{si le passager } i \text{ n'a pas survécu.} \end{cases}$$

La variable d'intérêt correspond à la loi de Bernoulli avec la probabilité de succès $0 \leq \theta_i \leq 1$ inconnue :

$$Y_i \sim \mathcal{B}ernoulli(\theta_i).$$

Exemple 2

Si on suppose que tous les passagers ont la même probabilité de survie, *i.e.* $\theta_i = \theta$ pour $1 \leq i \leq n$, l'estimation de θ correspond à la proportion de passagers ayant survécu au naufrage :

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

Avec l'échantillon de d'entraînement, on trouve que $\hat{\theta} = 0.39$. Par conséquent, si on sélectionne un passager au hasard de la liste des passagers, la probabilité que ce passager ait survécu au naufrage est de 0.39.
Si ce modèle est utilisé pour prédire la survie des 437 passagers de l'échantillon de test, on prédit que tous les passagers de l'échantillon de test n'ont pas survécu. Le pourcentage de bonnes prédictions est de 62.7%.

À l'instar de la régression linéaire, on souhaite expliquer la variabilité de la variable d'intérêt à l'aide de p variables explicatives $X_1, \ldots, X_p$. L'hypothèse 1 de la régression

linéaire

$$\mathbb{E}(Y_i|X_i = \boldsymbol{x}_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$$

ne peut pas s'appliquer dans le cas où la variable d'intérêt Y est une variable de Bernoulli. En effet, l'espérance conditionnelle à gauche doit être bornée entre 0 et 1 puisqu'elle correspond à une probabilité de succès. Or la relation linéaire à droite est définie sur l'ensemble des réels.

Les modèles linéaires généralisés utilisent une fonction injective $g(\cdot)$ pour transformer l'image de l'espérance conditionnelle à l'ensemble des réels. Dans le cas de la loi de Bernoulli, cette fonction transpose les valeurs dans l'intervalle $(0, 1)$ à l'ensemble des réels :

$$g : (0, 1) \to \mathbb{R}.$$

Avec cette fonction, la relation linéaire avec les variables explicatives peut être utilisée de la façon suivante :

$$g\left\{\mathbb{E}(Y_i|\boldsymbol{X}_i = \boldsymbol{x}_i)\right\} = \boldsymbol{x}_i^\top \boldsymbol{\beta}. \tag{3.2}$$

Dans la littérature, on retrouve deux choix très populaires de fonction $g(\cdot)$ permettant de satisfaire cette contrainte : la fonction *logit* et la fonction *probit*. Les modèles résultants du choix de l'une de ces fonctions sont décrits dans les deux prochaines sections. Les modèles linéaires généralisés pour une variable de type Bernoulli sont appelés modèles de **régression logistique**.

Exemple 3

Les passagers voyageant à bord du Titanic étaient séparés en trois classes. La première classe accueille les passagers les plus fortunés du navire. La deuxième classe, plus hétéroclite, comprend des entrepreneurs, des enseignants, des ecclésiastiques, etc. La troisième classe est composé surtout d'immigrants qui voyagent en famille.
Dans le film Titanic de 1997, une image qui est véhiculée est que les passagers de première classe avaient une meilleure chance de survie que les passagers des deux autres classes. Cette hypothèse sera testée en implémentant le modèle de régression logistique utilisant la classe des passagers comme variable explicative qualitative. Elle est incorporée dans le modèle à l'aide des deux variables indicatrices suivantes :

$$x_1 = \begin{cases} 1 & \text{si le passager voyage en première classe} \\ 0 & \text{si le passager ne voyage pas en première classe} \end{cases}$$

$$x_2 = \begin{cases} 1 & \text{si le passager voyage en deuxième classe} \\ 0 & \text{si le passager ne voyage pas en deuxième classe} \end{cases}$$

3.1.1 Le modèle logit

Le modèle logit utilise la fonction de lien logit g définie de la façon suivante :

$$\begin{array}{rccc} g & : & (0,1) & \to & \mathbb{R} \\ & & z & \mapsto & \ln\left(\frac{z}{1-z}\right). \end{array}$$

Exercice 1

Montrez que la fonction réciproque s'exprime ainsi :

$$\begin{array}{rccc} g^{-1} & : & \mathbb{R} & \to & (0,1) \\ & & z & \mapsto & \frac{e^z}{1+e^z}, \end{array}$$

et que par conséquent, l'équation (3.2) peut s'écrire de la façon suivante :

$$\mathbb{E}(Y_i|\boldsymbol{X}_i = \boldsymbol{x}_i) = \frac{\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})}.$$

En examinant l'expression de l'espérance conditionnelle précédente, on remarque que si $\beta_j > 0$, alors une hausse de X_j, alors que toutes les autres variables explicatives restent inchangées, augmente la probabilité d'observer un succès. Si $\beta_j < 0$, alors une hausse de X_j, alors que toutes les autres variables explicatives restent inchangées, diminue la probabilité d'observer un succès. Si $\beta_j = 0$, alors la variable X_j n'influence pas la probabilité de succès.

Remarque. *Si $Y \sim \mathcal{B}ernoulli(\theta)$, le ratio $\theta/(1-\theta) \in (0, \infty)$ est appelé* cote *en probabilités. Par exemple, une cote de 2 signifie que l'événement* succès *est deux fois plus probable que l'événement* échec. *Cette mesure est utilisée en paris sportifs et en sciences de la santé.*

Exercice 2

Pour le modèle logit, montrez que le logarithme de la cote de Y_i s'exprime ainsi :

$$\ln\left(\frac{\theta_i}{1-\theta_i}\right) = \boldsymbol{x}_i^\top \boldsymbol{\beta} = \beta_0 + x_1\beta_1 + \ldots x_p\beta_p.$$

Le coefficient de régression β_j s'interprète comme la variation du logarithme de la cote lorsque x_j augmente d'une unité et que toutes les autres variables demeurent inchangées. Si x_j augmente d'une unité et que toutes les autres variables explicatives restent inchangées, alors la cote $\theta_i/(1-\theta_i)$ est multipliée par la facteur $\exp(\beta_j)$. Ce facteur $\exp(\beta_j)$ est communément appelé *rapport de cotes*, puisque cette valeur correspond à la cote de l'événement

$$\{Y_i = 1|\boldsymbol{X}_i = (1, x_{i1}, \ldots, x_{ij} + 1, \ldots, x_{ip})\}$$

divisée par la cote de l'événement

$$\{Y_i = 1|\boldsymbol{X}_i = (1, x_{i1}, \ldots, x_{ij}, \ldots, x_{ip})\}.$$

Estimation des paramètres

L'estimation des paramètres dans le cas de la régression logistique et plus généralement dans le cas des modèles linéaires généralisés se fait en maximisant la vraisemblance. On suppose que les observations de l'échantillon d'entraînement sont indépendantes. Cette supposition correspond à l'hypothèse 3 du chapitre précédent. De plus, à l'instar du chapitre précédent, on considère que les variables explicatives sont des constantes.

Exercice 3

Pour le modèle de régression logistique avec la fonction de lien logit, montrez que la vraisemblance de l'échantillon aléatoire $\{(\boldsymbol{x}_i, Y_i) : i = 1, \ldots, n\}$ s'exprime sous la forme suivante :

$$f_{(\boldsymbol{Y}|\boldsymbol{\beta})}(\boldsymbol{y}) = \prod_{i=1}^{n} \frac{\exp\left(y_i \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)}{1 + \exp\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}\right)}$$

Contrairement à la régression linéaire, il n'existe pas de forme analytique pour les estimations des coefficients de régression. La vraisemblance doit être maximisée numériquement à l'aide d'un langage de programmation tel que Julia.

Exemple 4: (suite)

En utilisant la classe pour prédire la survie des passagers du Titanic de l'ensemble d'entraînement, la commande `glm()` de la librairie GLM de Julia peut être utilisée. La sortie correspondante est la suivante :

```
Y ~ 1 + x₁ + x₂

Coefficients:
```

	Coef.	Std. Error	z	Pr(>|z|)	Lower 95%	Upper 95%
(Intercept)	-1.03532	0.10412	-9.94	<1e-22	-1.23939	-0.831247
x₁	1.65013	0.178737	9.23	<1e-19	1.29981	2.00045
x₂	0.678642	0.181435	3.74	0.0002	0.323036	1.03425

La colonne *Coef.* correspond à l'estimation du coefficient de régression, *Std. Error* correspond à l'erreur standard de l'estimation du coefficient de régression, z correspond à la statistique du test vérifier si la vraie valeur du coefficient est égale à 0, $Pr(>|z|)$ correspond à la valeur-p de ce test, *Lower 95%* et *Upper 95%* correspondent respectivement aux bornes inférieures et supérieures de l'intervalle de confiance pour le coefficient de régression. On obtient donc que

$$\hat{\beta}_0 = -1.03, \quad \hat{\beta}_1 = 1.65 \quad \text{et} \quad \hat{\beta}_2 = 0.68$$

Exercice 4

Avec les estimations obtenues à l'exemple précédent, montrez que les estimations des probabilités de survie d'un passager de première, de deuxième et de troisième classe sont respectivement de 65%, 41% et 26%.

Exercice 5

Avec les estimations obtenues à l'exemple précédent, montrez que les cotes correspondante à la survie d'un passager de première, deuxième et troisième classe sont respectivement de 1.86, 0.70 et 0.36.

3.1.2 Le modèle probit

Le modèle probit s'applique dans les mêmes circonstances que dans celles du modèle logit. Dans les deux cas, les modèles sont développés pour une variable d'intérêt de type Bernoulli. Il n'y a que la fonction de lien g qui change. Le modèle probit utilise la fonction de répartition inverse de la loi normale centrée réduite Φ comme fonction $g(\cdot)$:

$$\begin{array}{rccc} g & : & (0,1) & \to & \mathbb{R} \\ & & z & \mapsto & \Phi^{-1}(z). \end{array}$$

où Φ correspond à la fonction de répartition de la loi normale centrée réduite

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) \, du.$$

Par conséquent, l'équation (3.2) devient :

$$\mathbb{E}(Y_i | \boldsymbol{X}_i = \boldsymbol{x}_i) = \Phi\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}\right).$$

Exercice 6

Pour le modèle de régression logistique avec la fonction de lien probit, montrez que la vraisemblance de l'échantillon aléatoire $\{(\boldsymbol{x}_i, Y_i) : i = 1, \ldots, n\}$ s'exprime sous la forme suivante :

$$f_{(\boldsymbol{Y}|\boldsymbol{\beta})}(\boldsymbol{y}) = \prod_{i=1}^{n} \left\{\Phi(\boldsymbol{x}_i^\top \boldsymbol{\beta})\right\}^{y_i} \left\{1 - \Phi(\boldsymbol{x}_i^\top \boldsymbol{\beta})\right\}^{1-y_i}.$$

À l'instar du modèle logit, les estimateurs du maximum de la vraisemblance des coefficients de régression ne s'expriment pas sous une forme analytique. La vraisemblance doit donc être maximisée numériquement avec un langage de programmation.

La fonction de lien probit est surtout utilisée dans le cadre de la régression bayésienne. Sa formulation permet des simplifications non négligeables pour l'implémentation du modèle dans le cadre bayésien. En maximum de la vraisemblance, la fonction de lien probit n'est que très peu utilisée.

Exemple 5

Avec les données sur les passagers du Titanic, les estimations des paramètres du modèle probit obtenues numériquement avec la librairie GLM de Julia sont les suivantes :

$$\hat{\beta}_0 = -0.64;$$
$$\hat{\beta}_1 = 1.02;$$
$$\hat{\beta}_2 = 0.41.$$

Exercice 7

Avec les estimations obtenues à l'exemple précédent, montrez que les estimations des probabilités de survie d'un passager de première, de deuxième et de troisième classe sont respectivement de 65%, 41% et 26%. Elles sont essentiellement les mêmes qu'avec le modèle logit.

3.1.3 Indice de qualité du modèle

La qualité du modèle de régression logistique peut être évaluée autant en estimation sur l'ensemble d'entraînement qu'en prédiction sur l'ensemble de test. Supposons en premier lieu que la qualité du modèle soit évaluée en estimation sur l'ensemble d'entraînement. Pour chacune des observations, on obtient une estimation de la probabilité de succès de Y_i :

$$\hat{\theta}_i = g^{-1}\left(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}\right).$$

Généralement, on estime Y_i de la façon suivante :

$$\hat{Y}_i = \begin{cases} 0 & \text{si } \hat{\theta}_i < u; \\ 1 & \text{si } \hat{\theta}_i \geq u. \end{cases}$$

Remarque. *Le seuil u optimal pour prédire la classe de Y_i peut être différent de $1/2$ si les observations sont débalancées en faveur de l'une ou l'autre des classes succès ou échec.*

Quatre cas de figure sont alors possibles en régression logistique :

(i) On estime $\hat{Y}_i = 1$ lorsque $Y_i = 1$. C'est un **vrai positif** (VP).

(ii) On estime $\hat{Y}_i = 1$ lorsque $Y_i = 0$. C'est un **faux positif** (FP).

(iii) On estime $\hat{Y}_i = 0$ lorsque $Y_i = 1$. C'est un **faux négatif** (FN).

(iv) On estime $\hat{Y}_i = 0$ lorsque $Y_i = 0$. C'est un **vrai négatif** (VN).

On dénombre ces cas de la façon suivante :

$$\text{VP} = \text{Card}\left\{i \in \{1, \ldots, n\} : Y_i = 1 \text{ et } \hat{Y}_i = 1\right\}$$
$$\text{FP} = \text{Card}\left\{i \in \{1, \ldots, n\} : Y_i = 0 \text{ et } \hat{Y}_i = 1\right\}$$
$$\text{FN} = \text{Card}\left\{i \in \{1, \ldots, n\} : Y_i = 1 \text{ et } \hat{Y}_i = 0\right\}$$
$$\text{VN} = \text{Card}\left\{i \in \{1, \ldots, n\} : Y_i = 0 \text{ et } \hat{Y}_i = 0\right\}$$

De ces quantités, deux mesures peuvent être définies : la sensibilité et la spécificité. La sensibilité correspond à la probabilité d'avoir un positif correctement identifié et la spécificité correspond à la probabilité d'avoir un vrai négatif correctement identifié :

$$\text{sensibilité} = \frac{VP}{VP + FN} \qquad \text{et} \qquad \text{spécificité} = \frac{VN}{VN + FP}.$$

En général, un bon modèle est un compromis entre sensibilité et spécificité. Par exemple, si un modèle indiquait que tous les passagers ont survécu au naufrage du Titanic, alors sa sensibilité serait égale à 1 puisque tous les survivants seraient bien identifiés. Par contre, sa spécificité serait nulle puisque toutes les victimes seraient identifiés comme survivants. Plusieurs mesures de performance utilisent un mélange de la sensibilité et de la spécificité comme mesure de performance, tel le score F1.

3.1.4 Courbe ROC

La qualité d'ajustement d'un modèle de régression logistique peut être évaluée à l'aide des proportions observées de vrais positifs et de vrais négatifs pour différents seuils de décision $u \in (0, 1)$. Ce seuil correspond à la valeur de la probabilité pour laquelle un succès est prédit. Autrement dit, on a que $\hat{Y}_i = 1$ si $\theta_i > u$.

En fonction du seuil u, les proportions de vrais positifs p_u (la sensibilité) et de faux positifs q_u (1 - spécificité) peuvent être calculées de la façon suivante :

$$p_u = \frac{\text{VP}}{\text{VP} + \text{FN}}$$

$$q_u = \frac{\text{FP}}{\text{FP} + \text{VN}}.$$

La courbe traçant la proportion de vrais positifs p_u en fonction de la proportion de faux positifs q_u pour plusieurs valeurs de u s'appelle la courbe ROC (pour *Receiver Operating Characteristic curve*). Plus la courbe longe le segment formé par les points (0,0), (0,1) et

Aire sous la courbe ROC (A)	Indice de la qualité du modèle
$0.9 \leq A \leq 1$	Excellent
$0.8 \leq A < 0.9$	Bon
$0.7 \leq A < 0.8$	Moyen
$0.6 \leq A < 0.7$	Faible
$0.5 \leq A < 0.6$	Mauvais

TABLE 3.1 – Indice de la qualité d'un modèle de régression logistique en fonction de l'aire sous la courbe ROC.

(1,1), plus le modèle de régression logistique est bon pour classifier les observations. Le pire scénario consisterait en une droite à 45 degrés qui relie les points (0,0) et (1,1). La courbe ROC permet également d'évaluer la sensibilité à la règle de décision en faisait varier le seuil $0 \leq u \leq 1$.

L'aire sous la courbe ROC donne un indice de la qualité du modèle. Plus l'aire est grande, meilleur est le modèle. Le tableau 3.1 compile une gradation arbitraire quoique utile des modèles de régression logistique.

La courbe ROC et l'aire sous celle-ci est très utile pour mesurer de façon absolue la qualité d'un modèle de régression logistique. L'aire sous la courbe ne possède cependant pas une interprétation aussi intéressante que le coefficient de détermination dans le cas de la régression linéaire. Elle permet néanmoins d'évaluer la qualité d'un modèle de régression logistique et d'effectuer une comparaison de modèle.

Exemple 6

La figure 3.1 illustre la courbe ROC pour le modèle de régression logistique utilisant uniquement la classe des passagers comme variable explicative. L'aire sous la courbe est de 0.67, ce qui correspond à un modèle ayant un faible pouvoir de discrimination selon le tableau 3.1. Pour augmenter la qualité du modèle, il faudra ajouter des variables explicatives. C'est d'ailleurs l'objet du TD concernant ce chapitre.

3.2 Lorsque la variable d'intérêt est une variable aléatoire distribuée selon la loi binomiale

Soit la variable $Y \sim \mathcal{B}inomiale(m, \theta)$, où le nombre d'essais $m \geq 1$ est connu et la probabilité $0 \leq \theta \leq 1$ est inconnue. On souhaite intégrer des variables explicatives $\boldsymbol{x}$ pour modéliser la probabilité de succès θ. Si $m = 1$, alors on retombe sur le cas de la régression logistique présenté à la section précédente.

Plutôt que d'utiliser directement la variable Y qui prend des valeurs dans l'ensemble $\{0, 1, \ldots, m\}$, on pose la variable $Z = Y/m$ qui prend des valeurs dans l'ensemble $\{0, 1/m,$

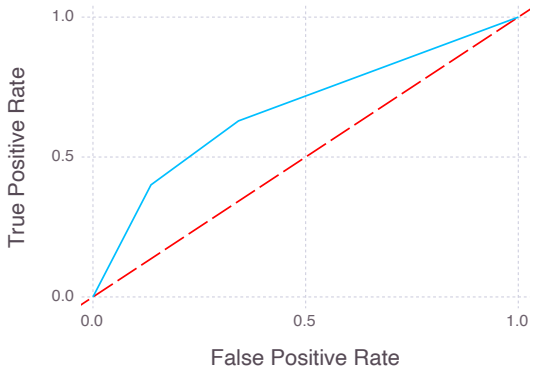


FIGURE 3.1 – Courbe ROC du modèle de prédiction n'utilisant que la classe des passagers du Titanic pour prédire leur survie.

$2/m, \ldots, 1\}$. La variable Z correspond à la proportion de succès parmi les m essais. On a que

$$\mathbb{E}(Z) = \mathbb{E}\left(\frac{Y}{m}\right) = \frac{1}{m}\,\mathbb{E}(Y) = \frac{1}{m}\,m\theta = \theta.$$

L'espérance de la variable aléatoire Z est donc égale à la probabilité de succès. Les fonctions de lien logit et probit présentées à la section précédente sont par conséquent adaptées.

Soit un échantillon aléatoire composé de n variables $Y_i \sim \mathcal{B}inomiale(m_i, \theta_i)$ pour $i \in \{1, \ldots, n\}$ où les nombres d'essais m_i sont tous connus. Pour chacun des Y_i, un vecteur de $(p+1)$ variables explicatives (incluant l'ordonnée à l'origine) $\boldsymbol{x}_i$ est disponible pour expliquer la proportion de succès.

Exercice 8

Avec la fonction de lien logit, montrez que la vraisemblance de l'échantillon aléatoire $\{(\boldsymbol{x}_i, Y_i) : i = 1, \ldots, n\}$ s'exprime sous la forme suivante :

$$f_{(\boldsymbol{Y}|\boldsymbol{\beta})}(\boldsymbol{y}) = \prod_{i=1}^{n} \binom{m_i}{y_i} \theta_i^{y_i} (1-\theta_i)^{(m_i - y_i)};$$

où

$$\theta_i = \frac{\exp(\boldsymbol{x}_i^\top \beta)}{1 + \exp(\boldsymbol{x}_i^\top \beta)}.$$

À l'instar de la régression logistique, les estimateurs du maximum de la vraisemblance des coefficients de régression ne s'expriment pas sous une forme analytique. La vraisemblance doit donc être maximisée numériquement.

Contrairement à la régression linéaire et à la régression logistique, il n'existe pas d'indice absolu pour mesurer la qualité d'un modèle de régression binomiale. Pour comparer plusieurs modèles entre eux, on pourra se servir de la validation croisée ou du *Bayesian Information Criterion (BIC)*. Le BIC est une mesure d'adéquation du modèle pénalisée en fonction du nombre de paramètres. Nous étudierons cette mesure au Chapitre 5.

3.3 Lorsque la variable d'intérêt est une variable aléatoire distribuée selon la loi de Poisson

La loi de Poisson apparaît naturellement en probabilités lorsque l'on dénombre des événements, par exemple le nombre d'accidents à une certaine intersection. Si $Y \sim \mathcal{P}oisson(\theta)$ avec $\theta > 0$, on a que l'espérance de Y est strictement positive :

$$\mathbb{E}(Y) = \theta > 0.$$

La fonction g généralement utilisée dans ce cas est la fonction $g(z) = \log(z)$. Par conséquent, l'équation (3.2) devient

$$E(Y_i \mid \boldsymbol{X}_i = \boldsymbol{x}_i) = \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}).$$

Le modèle résultant est souvent appelé *régression de Poisson*.

Exercice 9

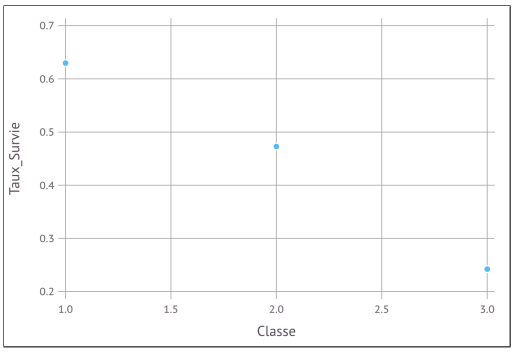
Pour la régression de Poisson, montrez que la vraisemblance des paramètres pour l'échantillon aléatoire $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ s'exprime sous la forme suivante :

$$f_{(\boldsymbol{Y}|\boldsymbol{\beta})}(\boldsymbol{y}) = \prod_{i=1}^{n} \frac{\exp\left(y_i \boldsymbol{x}_i^\top \boldsymbol{\beta} - e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}\right)}{y_i!}$$

À l'instar de la régression logistique, les estimateurs du maximum de la vraisemblance des coefficients de régression ne s'expriment pas sous une forme analytique. La vraisemblance doit donc être maximisée numériquement.

3.4 Exercices

1. Dans les exemples de ce chapitre, le modèle de régression logistique utilisant la fonction de lien logit a été ajusté sur les 872 passagers de l'échantillon d'entraînement afin de prédire la survie des 437 passagers de l'échantillon de test en fonction de la classe des passagers.
   (a) Le graphique suivant illustre le taux de survie des 891 passagers de l'échantillon d'entraînement en fonction de leur classe. Est-ce que ce graphique suggère que la probabilité de survie varie en fonction de la classe ?



(b) Écrivez le modèle de régression logistique correspondant en utilisant la fonction de lien logit.

(c) En utilisant les variables indicatrices suivantes pour la classe :

$$x_1 = \begin{cases} 1 & \text{si le passager voyage en première classe} \\ 0 & \text{si le passager ne voyage pas en première classe} \end{cases}$$

$$x_2 = \begin{cases} 1 & \text{si le passager voyage en deuxième classe} \\ 0 & \text{si le passager ne voyage pas en deuxième classe} \end{cases}$$

les estimations des paramètres sont données dans la sortie Julia suivante :

```
Formula: Y ~ 1 + X₁ + X₂

Coefficients:
             Estimate Std.Error  z value Pr(>|z|)
(Intercept)  -1.13977  0.105316 -10.8224   <1e-26
X₁             1.6704   0.17591  9.49574   <1e-20
X₂            1.03097   0.18137  5.68434    <1e-7
```

Est-ce que la classe des passagers explique la probabilité de survie des passagers ? Justifiez.

(d) Selon les estimations des paramètres obtenues par Julia, quelle est la probabilité de survie d'un passager de première classe ?

(e) À quoi correspond l'ordonnée à l'origine (*intercept*) dans ce modèle ?

2. On utilise le modèle de régression logistique avec la fonction de lien *logit* pour déterminer si la personne i empruntera le transport en commun pour son prochain déplacement :

$$Y_i = \begin{cases} 0 & \text{si la personne } i \text{ n'emprunte pas le transport en commun ;} \\ 1 & \text{si la personne } i \text{ emprunte le transport en commun.} \end{cases}$$

La variable explicative x_i correspond à la distance (en mètres) entre le lieu de résidence de la personne i et l'arrêt de transport en commun le plus près.

Avec un échantillon aléatoire de taille n, on obtient les estimations suivantes des paramètres de régression :

$$\hat{\beta}_0 = 1,4 \quad \text{et} \quad \hat{\beta}_1 = -0,02.$$

a) Quelle est la probabilité qu'une personne habitant à 100 m d'un arrêt de transport en commun emprunte le transport en commun pour son prochain déplacement ?

b) Que représente l'ordonnée à l'origine β_0 dans ce modèle ?

c) On souhaite incorporer le statut de la personne (étudiant, travailleur, retraité, autre) dans le modèle de régression logistique. Détaillez toutes les variables explicatives qui seront nécessaires et écrivez l'équation du nouveau modèle.

d) Selon votre modèle défini à la question (c), que représente maintenant l'ordonnée à l'origine β_0 de votre modèle ?

3. On modélise le nombre d'accidents par année Y à une intersection par la loi de Poisson de paramètre $\theta > 0$ inconnu :

$$Y \sim \mathcal{P}oisson(\theta).$$

On recense le nombre d'accidents à cette intersection depuis les n dernières années. On a donc un échantillon aléatoire de taille n : $(Y_1, \ldots, Y_n)$.

a) Quel est l'estimateur du maximum de la vraisemblance de θ ?

b) Quelle est l'interprétation de θ pour le présent problème ?

c) Supposons que l'on souhaite déterminer s'il existe une tendance en fonction des années du nombres d'accidents, quelle serait la variable explicative appropriée pour vérifier cette affirmation avec un modèle de régression ?

d) Pour une valeur de la variable explicative x donnée, comment s'exprime l'espérance de la variable Y ? Autrement dit, comment s'exprime $\mathbb{E}(Y|X = x)$ pour le modèle de régression du numéro (c) ?