CHAPITRE 3

Introduction aux modèles linéaires généralisés

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

Jonathan Jalbert – Automne 2019

Les modèles linéaires généralisés constituent une extension 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.
- Effectuer la sélection des variables explicatives.

Dans ce chapitre, nous étudierons la survie des passagers du Titanic. Lors du naufrage du Titanic, entre 1 490 et 1 520 personnes disparaissent parmis les 1316 passagers et 889 membres d'équipage à bord. Nous étudierons l'effet de plusieurs caractéristiques des personnes à bord sur leur probabilité de survie au naufrage. Dans cet exemple, la variable d'intérêt correspond à la survie d'un passager et elle peut être modélisé par la loi de Ber-

noulli :

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

Le modèle de régression linéaire étudié au chapitre précédent ne peut pas modéliser cette variable d'intérêt discrète en fonction de différentes variables explicatives.

3.1 Généralisation des modèles de régression linéaire

L'hypothèse centrale de la régression linéaire consiste à supposer qu'il existe un lien linéaire entre l'espérance de la variable d'intérêt Y et les p variables explicatives $\boldsymbol{X} = (X_1, \ldots, X_p)$:

$$\mathbb{E}(Y|X = \boldsymbol{x}) = \boldsymbol{x}\boldsymbol{\beta}.$$

Cette hypothèse ne peut pas toujours être satisfaite. C'est notamment le cas si la variable réponse est de type Bernoulli. Dans ce cas, l'espérance conditionnelle précédente doit être bornée entre 0 et 1 puisqu'elle correspond à une probabilité de succès. La relation linéaire est inappropriée dans ce cas car elle ne permet pas de borner l'espérance conditionnelle.

Les modèles linéaires généralisés permettent de considérer des relations non linéaires entre l'espérance conditionnelles et les variables explicatives. Les modèles linéaires généralisés utilisent une fonction injective g pour transformer la relation entre l'espérance conditionnelle et les variables explicatives de la façon suivante :

$$\mathbb{E}(Y|X = \boldsymbol{x}) = g^{-1}\left(\boldsymbol{x}\boldsymbol{\beta}\right). \tag{3.1}$$

La fonction $g(\cdot)$ permet de transposer l'espérance conditionnelle de la variable Y aux réels. Le choix de cette fonction $g(\cdot)$ dépend du type de la variable Y. Nous verrons dans les prochaines sections plusieurs choix possibles en fonction du type de la variable d'intérêt. De plus, pour un même type de variable, le choix de la fonction $g(\cdot)$ n'est pas unique, plusieurs fonctions injectives sont possibles.

À l'instar du chapitre précédent, nous supposons que l'incertitude sur les variables explicatives est négligeable par rapport à l'incertitude sur la variable réponse. Par conséquent, les variables $\boldsymbol{X}$ ne sont pas considérées comme des variables aléatoires. Nous supposons également que nous disposons d'un échantillon aléatoire indépendant de taille n, $\{(\boldsymbol{x_i}, Y_i) : 1 \leq i \leq n\}$, pour estimer les paramètres inconnus $\boldsymbol{\beta}$.

3.2 Lorsque la variable réponse est de type Bernoulli

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

$$Y \sim \mathcal{B}ernoulli(\theta); \tag{3.2}$$

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

En ajoutant des variables explicatives $\boldsymbol{x}$ pour Y, l'espérance conditionnelle de $\mathbb{E}(Y|X = \boldsymbol{x})$, qui correspond toujours à une probabilité de succès puisque la variable aléatoire $(Y|\boldsymbol{X} = \boldsymbol{x})$ est une loi de Bernoulli, doit demeurer dans l'intervalle $(0, 1)$. On doit donc trouver une fonction g telle que

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

ou de façon équivalente telle que :

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

Dans la littérature, on retrouve deux choix très populaires de fonction g permettant de satisfaire cette contraite : 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 1

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 avec les données sur les passagers. Dans un premier temps, on souhaite vérifier si la classe du passager a un effet sur sa probabilité de survie. La variable explicative *classe* doit être incorporée dans un modèle de régression logistique à l'aide de deux variables indicatrices (puisqu'il y a trois classes) :

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

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

Pour information, 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. Soit les variables explicatives

3.2.1 Le modèle logit

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

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

De façon réciproque, la fonction inverse s'écrit de la façon suivante :

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

Par conséquent, l'équation (3.1) devient

$$\mathbb{E}(Y|X=\boldsymbol{x}) = \frac{\exp(\boldsymbol{x}\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}\boldsymbol{\beta})}.$$

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, *i.e* $\mathbb{P}(Y=1|\boldsymbol{X}=\boldsymbol{x})$. 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.

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*. C'est une mesure largement utilisée en paris sportifs et en sciences de la santé.

Exercice 1

Dénotons la probabilité de succès conditionnelle à $(\boldsymbol{X}=\boldsymbol{x})$ par

$$\theta_{\boldsymbol{x}} = \mathbb{E}(Y|X=\boldsymbol{x}).$$

Pour le modèle logit, montrez que le logarithme de la cote $\theta_{\boldsymbol{x}}/(1-\theta_{\boldsymbol{x}})$ s'exprime sous la forme suivante :

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

Montrez aussi que la cote s'exprime sous la forme suivante :

$$\frac{\theta_{\boldsymbol{x}}}{1-\theta_{\boldsymbol{x}}} = \exp\left(\boldsymbol{x}\boldsymbol{\beta}\right).$$

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 $\mu_{\boldsymbol{x}}/(1-\mu_{\boldsymbol{x}})$ 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

$$\{Y = 1 | \boldsymbol{X} = (x_1, \ldots, x_j + 1, \ldots, x_p)\}$$

divisée par la cote de l'événement

$$\{Y = 1 | \boldsymbol{X} = (x_1, \ldots, x_j, \ldots, x_p)\}.$$

Basé sur un échantillon aléatoire, les $p+1$ coefficients de régression $(\beta_j : 0 \leq j \leq p)$ sont généralement estimés avec la méthode du maximum de la vraisemblance.

Exercice 2

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 :

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

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 2: (suite)

Avec les données sur les passagers du Titanic, les estimations obtenues numériquement avec le package GLM de Julia sont les suivantes :

$$\hat{\beta}_0 = -1,03;$$
$$\hat{\beta}_1 = 1,65;$$
$$\hat{\beta}_2 = 0,68.$$

Alors, la probabilité de survie d'un passager de première classe est de 65%, d'un passager de deuxième classe est de 41% et celle d'un passager de troisième classe est de 26%.

Exercice 3

Avec les estimations obtenues à l'exemple précédent, quelle est la cote correspondante à la survie d'un passager de troisième classe ? Rép. $\exp(-\hat{\beta}_0)$.

3.2.2 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 :

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

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 (3.1) devient :

$$\mathbb{E}(Y|X = \boldsymbol{x}) = \Phi\left(\boldsymbol{x}\boldsymbol{\beta}\right).$$

Exercice 4

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 :

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

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.

3.2.3 Indice de la qualité de l'ajustement

Si un modèle de régression logistique est utilisé pour l'échantillon aléatoire $\{(\boldsymbol{x}_i, Y_i) : i = 1, \ldots, n\}$, alors pour chacune des observations Y_i, on obtient une estimation de la probabilité que Y_i soit un succès :

$$\hat{\theta}_{\boldsymbol{x}_i} = \frac{\exp(\boldsymbol{x}_i\hat{\boldsymbol{\beta}})}{1 + \exp(\boldsymbol{x}_i\hat{\boldsymbol{\beta}})}.$$

Avec une règle de décision appropriée, on peut estimer Y_i par $\hat{Y}_i$. Il y a alors quatre cas possibles :

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

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

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

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

La *sensibilité* d'un test correspond à la probabilité d'avoir un vrai positif et la *spécificité* correspond à la probabilité d'avoir un vrai négatif. En général, un bon test est un compromis entre la sensibilité et la spécificité. Par exemple, si un test indiquait pour tous les patients qu'ils sont porteur du VIH, alors le test aurait une sensibilité de 1 puisque tous les porteurs du VIH seraient bien identifiés. Par contre, sa sensibilité serait nulle puisque tous les non porteurs seraient identifiés comme porteurs.

La qualité d'ajustement d'un modèle de régression logistique peut être estimée à l'aide des proportions observées de vrais positifs et de vrais négatifs. Soit le réel $u \in (0, 1)$ et soit la règle de décision suivante pour prédire Y_i :

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

On peut alors calculer la proportion de vrais positifs p_u et la proportion de faux positifs q_u de la façon suivante :

$$p_u = \frac{\text{Card}\left\{i \in \{1, \ldots, n\} : Y_i = 1 \text{ et } \hat{Y}_i = 1\right\}}{\text{Card}\left\{i \in \{1, \ldots, n\} : Y_i = 1\right\}};$$

$$q_u = \frac{\text{Card}\left\{i \in \{1, \ldots, n\} : Y_i = 0 \text{ et } \hat{Y}_i = 1\right\}}{\text{Card}\left\{i \in \{1, \ldots, n\} : Y_i = 0\right\}}.$$

Ces calculs peuvent être répétés pour plusieurs u. La courbe traçant la proportion de vrais positifs en fonction de la proportion de faux positifs, c'est-à-dire les points (q_u, p_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 (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.

Aire sous la courbe ROC (A)	Indice de la qualité du modèle
$0,9 \leq A < 1$	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.

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.

3.3 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 $\{0, 1/m, 2/m, \ldots, 1\}$. 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 Y 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 5

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 :

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

où

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

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.

3.4 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 (3.1) devient

$$E(Y \mid \boldsymbol{X} = \boldsymbol{x}) = \exp(\boldsymbol{x}\boldsymbol{\beta}).$$

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

Exercice 6

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 :

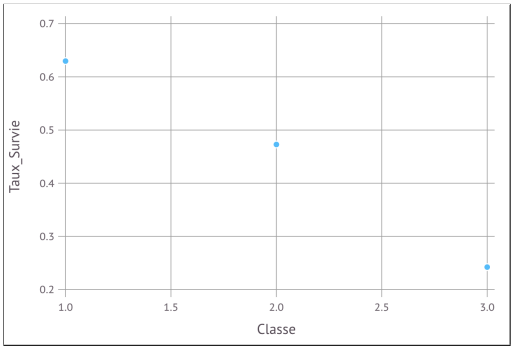
$$f_{(\boldsymbol{Y}|\boldsymbol{\beta})}(\boldsymbol{y}) = \prod_{i=1}^{n} \frac{\exp\left(y_i \boldsymbol{x}_i \boldsymbol{\beta} - e^{\boldsymbol{x}_i \boldsymbol{\beta}}\right)}{y_i!}$$

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.

3.5 Exercices

1. Dans le Travail Dirigé de ce chapitre, j'ai construit un modèle de régression logistique pour prédire la survie des 418 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) ?