M1SID — Modélisation

TD, Durée : Libre

Exercice 1 : (La loi géométrique, membre de la famille exponentielle)

On étudie le temps d'attente (en milli-secondes) entre deux requêtes successives sur un serveur informatique.

1. Expliquer pourquoi on peut modéliser ce temps d'attente par une variable de loi géométrique ?
2. On rappelle qu'une variable aléatoire Y est distribuée selon une loi géométrique de paramètre p si, pour tout entier k strictement positif,

$$\mathbb{P}(Y = k) = p\,(1-p)^{k-1} \ ,$$

$$\mathbb{E}(Y) = \frac{1}{p} \ .$$

   On suppose qu'on observe un échantillon $Y_1, Y_2, \ldots, Y_n$ de variables géométriques de paramètre p inconnu.

   Calculer l'estimateur de maximum de vraisemblance de p.
3. Montrer que la loi géométrique appartient à la famille exponentielle en écrivant sa densité sous la forme :

$$\mathbb{P}(Y_i = y \ ; \ \theta, \phi) = \exp\left(\frac{\theta y - b(\theta)}{a(\phi)} + c(y, \phi)\right) \ .$$

   On identifiera les paramètres θ et ϕ, ainsi que les fonctions $a(\phi)$, $b(\theta)$ et $c(y, \phi)$.

   Donner les expressions de θ en fonction de p et réciproquement.
4. On dispose pour chaque observation i ($i = 1, 2, \ldots, n$) d'une variable explicative x_i mesurant le nombre d'utilisateurs connectés au serveur. Ecrire le modèle permettant d'exprimer le temps d'attente entre deux requêtes en fonction du nombre d'utilisateurs connectés.
5. Donner l'expression de la densité de la loi de Y_i conditionnellement à x_i, en y faisant apparaître les coefficients β_0 et β_1.

Exercice 2 : (Tabagisme)

Le tableau ci-dessous donne les résultats d'une étude réalisée auprès de 680 femmes adultes concernant la relation éventuelle entre la consommation de cigarettes et l'hypertension artérielle. On note Y_i la variable associée à la présence d'une hypertension artérielle (0 si absence, 1 si présence) et X_i la variable associée à la consommation de tabac (1 si fumeuse, 0 si non-fumeuse).

	Hypertension	
Tabagisme	Non ($Y_i = 0$)	Oui ($Y_i = 1$)
Non-fumeuses	368	13
Fumeuses	271	28

1. Calculer les probabilités de présence d'hypertension selon que les femmes sont fumeuses ou non-fumeuses, que l'on notera respectivement p_1 et p_0.
2. Calculer les odds chez les fumeuses et les non-fumeuses. Rappel : Lors d'une expérience de Bernoulli Y, les odds dans un groupe A est le quotient

$$\frac{\mathbb{P}(Y = 1|A)}{\mathbb{P}(Y = 0|A)} \ .$$

3. Calculer l'odds ratio des fumeuses par rapport au non-fumeuses. Commenter le fait que ce ratio soit > 1 ou < 1.

4. Pour étudier l'effet du tabagisme sur l'hypertension, écriver le modèle de régression logistique. Donnez son expression pour $x_i = 0$ et $x_i = 1$.
   En déduire β_0 et β_1 en fonction de p_0 et p_1. On fera en particulier apparaître l'odds ratio.
5. A partir de ces expressions, estimez les paramètres intervenant dans ce modèle, et écrivez le modèle correspondant.
6. Commentez les estimations obtenues (en terme d'odds-ratio).

Exercice 3 : (Sujet d'examen 2011-2012)

Une étude médicale a porté sur 76 patients chez qui on a étudié :
— la survenue d'un infarctus du myocarde (notée y et codée par 1 si présent, 0 si absent).
— le taux de cholestérol (notée x et exprimée en mg/dl.
— la présence d'une angine de poitrine au cours du dernier mois (notée z et codée par 1 si présent, 0 sinon).

Parmi les 76 patients, 40 ont eu un infarctus du myocarde.

L'objectif de cette étude est d'évaluer l'impact du taux de cholestérol et d'une angine de poitrine sur la présence d'un infarctus. Pour y répondre, un modèle de régression logistique a été mis en oeuvre sous la forme :

$$\text{logit}(p_i) = \beta_0 + \beta_1 x_i + \beta_2 z_i$$

où pour un individu i ($i = 1, \ldots, n$), p_i représente la probabilité d'avoir un infarctus du myocarde (c'est-à-dire que $y_i = 1$), x_i est la valeur du taux de cholestérol et z_i est la variable associée à la présence d'une angine de poitrine.

1. Que vaut la probabilité d'avoir un infarctus du myocarde en l'absence d'effet des variables explicatives.
2. On désire tester l'ajustement global du modèle aux données par la statistique du rapport de vraisemblance, en comparant le modèle estimé et le modèle blanc.
   Pour cela, on vous donne la valeur de la log-vraisemblance du modèle estimé :

   $$-2 \log L\left(y; \widehat{\beta}\right) = 27.47 \ .$$

   Pour calculer celle du modèle blanc, on rappelle que la log-vraisemblance d'un modèle logistique (sans facteurs) est de la forme :

   $$\log L\left(y; p\right) = \sum_{i=1}^{n} y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \ .$$

   On rappelle que la statistique du rapport de vraisemblance sous l'hypothèse θ_0 s'écrit :

   $$\Lambda_n = -2 \log \frac{L(y; H_1)}{L(y; H_0)} \ ,$$

   et est asymptotiquement donné par une loi du χ^2.
3. La mise en oeuvre du modèle a permis d'obtenir les résultats suivants :

   $$\widehat{\beta}_0(y) = -41.1752, \quad \widehat{\beta}_1(y) = 0.2559, \quad \widehat{\beta}_2(y) = 3.4228 \ .$$

   $$\widehat{\text{Var}}(\widehat{\beta}) = \begin{pmatrix} 138.9552 & -0.8703 & -5.1169 \\ -0.870 & 0.0055 & 0.0288 \\ -5.117 & 0.0288 & 1.1721 \end{pmatrix} \ .$$

   Calculer l'intervalle de confiance du paramètre β_1 de niveau 95%. Concluez.
4. Testez la nullité du paramètre β_2 associé à la présence d'une angine de poitrine, par la statistique de Wald. Concluez.
5. Ecrivez la probabilité prédite de présence d'un problème cardiaque en fonction des 3 paramètres β_0, β_1 et β_2.
   A partir de cette formule, calculez la probabilité prédite de présence d'un problème cardiaque pour un individu dont le taux de cholestérol est égal à 150 mg/dl, dans le cas où il a une angine de poitrine, puis dans le cas où il n'en a pas.

Exercice 4 : (Le modèle log-linéaire)

On s'intéresse au nombre de décès du COVID-19 dans différent pays. Cette variable d'intérêt est notée Y et nous désirons l'expliquer la différents facteurs : l'âge moyen de la population, le PIB, la température moyenne de l'année, la taille moyenne des habitants etc... Une dimension de plus est obtenue en ajoutant l'intercept. Et nous notons les variables explicatives par $X \in \mathbb{R}^k$.

Par la loi des évennements rares, la loi naturelle pour Y est la loi de Poisson :

$$\forall k \in \mathbb{N},\ \mathbb{P}(Y = k) = e^{-\lambda}\frac{\lambda^k}{k!}\ ,$$

où $\lambda > 0$ est le paramètre. Mais conformément à la logique du modèle linéaire généralisé, la régression de Poisson suppose que la paramètre λ est lié à la variable explicative X. On pourra se référer à la page Wikipedia suivante :

https://en.wikipedia.org/wiki/Poisson_regression

1. La fonction de lien est le logarithme

   $$\log \mathbb{E}Y = \beta X$$

   où βX est le produit scalaire entre les paramètres $\beta \in \mathbb{R}^k$ et les facteurs $X \in \mathbb{R}^k$. D'où les noms de "Modèle log-linéaire" ou "Modèle de regression Poisson". Donner l'expression de la loi conditionnelle $(Y|X)$.

2. Démontrer que la log-vraisemblance d'écrit :

   $$\sum_{i=1}^{n}\left[y_i(\beta x_i) - e^{\beta x_i} - \log y_i!\right]\ .$$

3. Donner le programme de maximisation pour calculer l'estimateur de maximum de vraisemblance. Dans la maximisation de la vraisemblance, expliquer pourquoi nous avons le droit de poser :

   $$\ell(y, x, \beta) = \frac{1}{n}\sum_{i=1}^{n}\left[y_i(\beta x_i) - e^{\beta x_i}\right]$$

   et que les expressions des dérivées nécessaires à l'optimisation restent inchangées.

4. En déduire :

   $$\nabla_\beta \ell(y, x, \beta) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - e^{\beta x_i}\right)x_i = \frac{1}{n}VX\ ,$$

   $$\nabla^2_\beta \ell(y, x, \beta) = -\frac{1}{n}\sum_{i=1}^{n} e^{\beta x_i}x_i^T x_i = \ \frac{1}{n}X^T W X,$$

   où V et W sont les matrices explicites :

   $$V = \left(y_i - e^{\beta x_i}\right)_i \in M_{1,n}(\mathbb{R}),$$

   $$W = \operatorname{diag}\left(-e^{\beta x_i}\right)_{1\leq i\leq n}\ .$$

   La matrice $X \in M_{n,k}(\mathbb{R})$ est la matrice dont les lignes sont les facteurs explicatifs.

5. Pensez-vous qu'il soit possible d'utiliser l'algorithme de Newton-Raphson comme vu en cours et en TP, ou bien un algorithme de descente de gradient ? (Hint : Attention, il y a une exponentielle !)

   Remarque de fin : Dans le cas d'une implémentation pratique, vous pouvez utiliser statsmodels pour faire la régression Poisson en regardant statmodels sur

   https://www.statsmodels.org/stable/generated/statsmodels.discrete.discrete_model.Poisson.html