ACT-2003 Modèles linéaires en actuariat
Exercices
Modèles linéaires généralisés

Marie-Pier Côté

Automne 2018

1. Est-ce que les distributions suivantes font partie de la famille exponentielle linéaire ? Si oui, écrire la densité sous la forme exponentielle linéaire, donner le paramètre canonique, le paramètre de dispersion, l'espérance et la variance de Y en termes de la fonction $b()$ et la relation $V()$ entre la moyenne et la variance.
   i. Normale(μ, σ^2)
   ii. Uniforme$(0, \beta)$
   iii. Poisson(λ)
   iv. Bernoulli(π)
   v. Binomiale(m, π), $m > 0$ est un entier et est connu (On considère $Y^* = Y/m$).
   vi. Pareto(α, λ)
   vii. Gamma(α, β)
   viii. Binomiale négative(r, π) avec r connu (On considère $Y^* = Y/r$).

2. Quelles fonctions de lien peut-on utiliser pour un GLM avec une loi de Poisson ?

3. Quel est le lien canonique pour la loi gamma ? Est-ce que ce lien est toujours approprié ?

4. On suppose que $Y_1, ..., Y_n$ sont des v.a.s indépendantes et $Y_i \sim Poisson(\mu_i)$. Pour chaque observation, on a une seule variable explicative x_i.
   i. Quel est le lien canonique ?
   ii. Trouver les fonctions de score (à résoudre pour l'estimation des paramètres par maximum de vraisemblance)

5. Montrer que la déviance pour le modèle binomial est

$$D(y, \hat{\mu}) = 2 \sum_{i=1}^{n} m_i \left[y_i \ln \left(\frac{y_i}{\hat{\mu}_i} \right) + (1 - y_i) \ln \left(\frac{1 - y_i}{1 - \hat{\mu}_i} \right) \right].$$

6. Trouver les expressions des résidus de Pearson, d'Anscombe et de déviance pour la loi Gamma.

7. Les données suivantes représentent des données de comptage, du nombre d'échec pour trois appareils médicaux (M1, M2 et M3) lors de tests de résistance sur 1000 appareils de chaque type et pour quatre niveaux de résistance mécanique différents (I, II, III, IV).

Device\ Stress Level	I	II	III	IV
M1	6	8	18	10
M2	13	18	29	20
M3	9	8	21	19

   À l'aide de la modélisation Poisson (lien canonique), évaluer s'il y a une différence significative entre les taux d'échec des appareils.

8. **Réclamations d'assurance automobile au Royaume-Uni.** [1] Les données pour cet exercice sot contenues dans le fichier `BritishCar.csv (sep=";")` disponible sur le site du cours. On y trouve les montants de réclamations moyens pour les dommages causés au véhicule du détenteur de la police pour les véhicules assurés au Royaume-Uni en 1975. Les moyennes sont en livres sterling ajustées pour l'inflation.

Variable	Description
`OwnerAge`	Âge du détenteur de la police (8 catégories)
`Model`	Type de voiture (4 groupes)
`CarAge`	Âge du véhicule, en années (4 catégories)
`NClaims`	Nombre de réclamations
`AvCost`	Coût moyen par réclamation, en livres sterling

   On s'intéresse à la modélisation du coût moyen par réclamation.

   a) Ajuster un modèle de régression Gamma avec lien inverse pour la variable endogène `AvCost`. Inclure les effets principaux `OwnerAge`, `Model` et `CarAge`.
   b) Quelle est l'espérance du coût moyen de la réclamation pour un détenteur de police âgé entre 17 et 20 ans, avec une auto de type A âgée de moins de 3 ans ?
   c) Interpréter brièvement les coefficients pour la variable exogène `OwnerAge`.
   d) Interpréter brièvement les coefficients pour la variable exogène `Model`.
   e) Interpréter brièvement les coefficients pour la variable exogène `CarAge`.
   f) Pour quelle combinaison de variables exogènes l'espérance du coût de réclamation est-elle la plus élevée ? Calculer sa valeur.
   g) Pour quelle combinaison de variables exogènes l'espérance du coût de réclamation est-elle la plus faible ? Calculer sa valeur.

1. Source des données : McCullagh, P., and Nelder, J. A. (1989). Generalized linear models. Chapman and Hall, London.

h) Quelle est la déviance pour ce modèle ? Est-ce que le modèle semble adéquat ?

i) Tracer le graphique des résidus de Pearson en fonction des valeurs prédites, des résidus d'Ascombe en fonction des valeurs prédites et des résidus de déviance en fonction des valeurs prédites.

j) Obtient-on les mêmes conclusions aux sous-questions a) à h) si on utilise un lien logarithmique plutôt que le lien inverse ?

9. On considère les données suivantes, qui contiennent le nombre Y_i de turbines sur m_i qui ont été fissurées après x_i heures d'opération.

x_i	m_i	Y_i
400	39	2
1000	53	4
1400	33	3
1800	73	7
2200	30	5
2600	39	9
3000	42	9
3400	13	6
3800	34	22
4200	40	21
4600	36	21

(a) En utilisant un GLM binomial avec lien canonique, dériver les estimateurs des paramètres lorsque x_i est traité comme une variable exogène dichotomique avec 11 niveaux, et lorsque le prédicteur linéaire pour la donnée i est

$$\eta_i = \beta_0 + \beta_i, \text{ pour } i = 1, ..., 11,$$

avec la contrainte d'identifiabilité que $\beta_1 = 0$.

(b) En utilisant R et un GLM binomial avec lien canonique, ajuster le modèle où le prédicteur linéaire est

$$\eta_i = \beta_0 + \beta_1 x_i, \text{ pour } i = 1, ..., 11.$$

Donner les estimations des paramètres et leur écart-type.

(c) Refaire (b) en utilisant un lien probit. Donner les estimations des paramètres et leur écart-type.

(d) Refaire (b) en utilisant un lien log-log complémentaire. Donner les estimations des paramètres et leur écart-type.

(e) Comparer les prévisions (et leurs mesures d'incertitude) sous les trois modèles ajustés en (b), (c) et (d) pour une turbine qui était en opération pour 2000 heures.

(f) Tracer un graphique pour montrer si les modèles en (b), (c) et (d) ajustent bien (ou non) les données. Commenter.

Solutions

1. i. Normale(μ, σ^2) : oui,

$$\begin{aligned} f_Y(y) &= \frac{1}{(2\pi\sigma^2)^1/2} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right), \ y \in \mathbb{R} \\ &= \exp\left(\frac{y\mu - \mu^2/2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{\ln(2\pi\sigma^2)}{2}\right), \ y \in \mathbb{R}. \end{aligned}$$

   - Paramètre canonique : $\theta = \mu$
   - Paramètre de dispersion : $\phi = \sigma^2$
   - $b(\theta) = \frac{\theta^2}{2}$
   - $E[Y] = \dot{b}(\theta) = \frac{\partial}{\partial\theta}\frac{\theta^2}{2} = \theta = \mu$
   - var(Y) $= \phi\ddot{\mathrm{b}}(\theta) = \sigma^2\frac{\partial}{\partial\theta}\theta = \sigma^2$
   - $V(\mu) = 1$.

   ii. Uniforme$(0, \beta)$: non. Le domaine dépend du paramètre β.

   iii. Poisson(λ) :

$$\begin{aligned} f_Y(y;\lambda) &= \frac{\lambda^y e^{-\lambda}}{y!}, \text{ pour } y \in \mathbb{N}^+ \\ &= \exp\{y\ln\lambda - \lambda - \ln y!\} \\ f_Y(y;\theta,\phi) &= \exp\left\{\frac{y\theta - e^\theta}{\phi} - \ln y!\right\}. \end{aligned}$$

   - Paramètre canonique : $\theta = \ln\lambda$
   - Paramètre de dispersion : $\phi = 1$
   - $b(\theta) = e^\theta$
   - $E[Y] = \dot{b}(\theta) = \frac{\partial}{\partial\theta}e^\theta = e^\theta = \lambda$
   - var(Y) $= \phi\ddot{\mathrm{b}}(\theta) = \frac{\partial}{\partial\theta}\mathrm{e}^\theta = \mathrm{e}^\theta = \lambda$
   - $V(\mu) = \mu$.

   iv. Bernoulli(π)

$$\begin{aligned} f_Y(y;\pi) &= \pi^y(1-\pi)^{1-y}1(y \in \{0,1\}) \\ &= \exp\left\{y\ln\left(\frac{\pi}{1-\pi}\right) + \ln(1-\pi)\right\}1(y \in \{0,1\}). \end{aligned}$$

   - Paramètre canonique : $\theta = \ln\left(\frac{\pi}{1-\pi}\right)$
   - Paramètre de dispersion : $\phi = 1$
   - $b(\theta) = \ln(1+e^\theta)$

- $E[Y] = \dot{b}(\theta) = \frac{\partial}{\partial\theta}\ln(1+e^\theta) = \frac{e^\theta}{1+e^\theta} = \pi$
- $\text{var(Y)} = \phi\ddot{\text{b}}(\theta) = \frac{\partial}{\partial\theta}\frac{\text{e}^\theta}{1+\text{e}^\theta} = \frac{\text{e}^\theta}{(1+\text{e}^\theta)^2} = \pi(1-\pi)$
- $V(\mu) = \mu(1-\mu)$.

v. Binomiale(m, π), $m > 0$ est un entier et est connu.

$$\begin{aligned} f_Y(y;\pi) &= \binom{m}{y}\pi^y(1-\pi)^{m-y}1(y \in \{0,1,...,m\}) \\ &= \exp\left\{y\ln\left(\frac{\pi}{1-\pi}\right) + m\ln(1-\pi) + \ln\binom{m}{y}\right\}1(y \in \{0,1,...,m\}). \end{aligned}$$

Dans cette représentation, on a

$$E[Y] = m\pi \text{ et } Var(Y) = m\pi(1-\pi).$$

Cette forme est moins utilisée car l'espérance de Y dépend de m, le paramètre de dispersion. Souvent, on transforme les données. On utilise plutôt $Y^* = Y/m$. Alors, pour ces données transformées,

$$\begin{aligned} f_{Y^*}(y;\pi) &= \exp\left\{my\ln\left(\frac{\pi}{1-\pi}\right) + m\ln(1-\pi) + \ln\binom{m}{my}\right\},\ y \in \{0,1/m,...,1\} \\ &= \exp\left\{\frac{y\ln\left(\frac{\pi}{1-\pi}\right) + \ln(1-\pi)}{1/m} + \ln\binom{m}{my}\right\},\ y \in \{0,1/m,...,1\}. \end{aligned}$$

- Paramètre canonique : $\theta = \ln\left(\frac{\pi}{1-\pi}\right)$
- Paramètre de dispersion : $\phi = 1/m$
- $b(\theta) = \ln(1+e^\theta)$
- $E[Y^*] = \dot{b}(\theta) = \frac{\partial}{\partial\theta}\ln(1+e^\theta) = \frac{e^\theta}{1+e^\theta} = \pi$
- $\text{var(Y}^*) = \phi\ddot{\text{b}}(\theta) = \frac{1}{\text{m}}\frac{\partial}{\partial\theta}\frac{\text{e}^\theta}{1+\text{e}^\theta} = \frac{\text{e}^\theta}{\text{m}(1+\text{e}^\theta)^2} = \frac{\pi(1-\pi)}{\text{m}}$
- $V(\mu) = \mu(1-\mu)$.

vi. Pareto(α, λ) : non.

vii. Gamma(α, β) Soit $Y \sim Gamma(\alpha, \beta)$. Alors, avec un peu de travail, la densité peut être écrite sous la forme exponentielle linéaire.

$$f_Y(y;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}y^{\alpha-1}e^{-\beta y},$$

pour $y > 0$. On reparamétrise : $\mu = \alpha/\beta = E[Y]$ et α, on a donc $\beta = \alpha/\mu$ et

$$f_Y(y;\alpha,\mu) = \frac{1}{y\Gamma(\alpha)}\left(\frac{\alpha y}{\mu}\right)^{\alpha}\exp\left\{-\frac{\alpha y}{\mu}\right\}.$$

Posons $\theta = -1/\mu$, et $a(\phi) = 1/\alpha$, alors on trouve

$$f_Y(y;\theta,\phi) = \exp\left\{\frac{y\theta + \ln(-\theta)}{\phi} + \alpha\ln\alpha + (\alpha-1)\ln y - \ln\Gamma(\alpha)\right\}.$$

Donc, $b(\theta) = -\ln(-\theta)$ et $a(\phi) = 1/\alpha \Rightarrow \dot{b}(\theta) = \frac{-1}{\theta} = \mu$ et $\ddot{b}(\theta) = \frac{1}{\theta^2} = \mu^2$. Finalement,

$$E[Y] = \frac{-1}{\theta} = \mu \text{ et } Var(Y) = \frac{1}{\alpha}\mu^2.$$

viii. Binomiale négative(r,π) avec r connu. On considère $Y^* = Y/r$:

$$\begin{aligned} f_Y^*(y) &= \binom{r+ry-1}{ry}\pi^r(1-\pi)^{ry}, \text{ pour } y \in \{0, \frac{1}{r}, \frac{2}{r}, ...\} \\ &= \exp\left(ry\ln(1-\pi) + r\ln\pi + \ln\binom{r+ry-1}{ry}\right). \end{aligned}$$

- Paramètre canonique : $\theta = \ln(1-\pi)$
- Paramètre de dispersion : $\phi = 1/r$
- $b(\theta) = -\ln(1-e^{\theta})$
- $E[Y^*] = \dot{b}(\theta) = \frac{\partial}{\partial\theta} - \ln(1-e^{\theta}) = \frac{e^{\theta}}{1-e^{\theta}} = \frac{1-\pi}{\pi}$
- $\mathrm{var(Y^*)} = \phi\ddot{\mathrm{b}}(\theta) = \frac{1}{\mathrm{r}}\frac{\partial}{\partial\theta}\frac{\mathrm{e}^{\theta}}{1-\mathrm{e}^{\theta}} = \frac{\mathrm{e}^{\theta}}{\mathrm{r}(1-\mathrm{e}^{\theta})^2} = \frac{(1-\pi)}{\mathrm{r}\pi^2}$
- $V(\mu) = \mu(\mu+1)$.

—

2. Le lien canonique est le lien log : $\eta = \ln(\mu)$. On pourrait aussi utiliser d'autres fonctions de lien, telle que le lien identité $\eta = \mu$, le lien inverse $\eta = \frac{1}{\mu}$, mais le lien log est le plus approprié parce que son utilisation garantit une moyenne μ positive, ce qui est nécessaire pour la loi de Poisson.
3. Le lien canonique pour la loi Gamma est le lien inverse $\eta = 1/\mu$. Comme la moyenne d'une loi Gamma est toujours positive, ce lien n'est pas toujours approprié parce qu'il ne restreint pas le domaine de μ aux réels positifs. Le lien log serait plus approprié dans certains cas.
4. i. $\eta = g(\mu) = \ln(\mu)$

   ii. On a

$$\ln(\mu_i) = \eta_i = \beta_0 + \beta_1 x_i.$$

La densité de la loi Poisson est

$$\begin{aligned} f_{Y_i}(y_i;\mu_i) &= \exp\left(y_i\ln\mu_i - \mu_i - \ln y_i!\right) \\ f_{Y_i}(y_i;\beta_0,\beta_1) &= \exp\left(y_i(\beta_0+\beta_1 x_i) - e^{\beta_0+\beta_1 x_i} - \ln y_i!\right). \end{aligned}$$

La fonction de vraisemblance et la log-vraisemblance sont donc :

$$\mathcal{L}(\beta_0, \beta_1) = \prod_{i=1}^{n} f_{Y_i}(y_i; \beta_0, \beta_1) = \prod_{i=1}^{n} \exp\left(y_i(\beta_0 + \beta_1 x_i) - e^{\beta_0+\beta_1 x_i} - \ln y_i!\right)$$
$$\ell(\beta_0, \beta_1) = \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 x_i) - e^{\beta_0+\beta_1 x_i} + \text{constante.}$$

On maximise la log-vraisemblance :

$$\frac{\partial}{\partial \beta_0}\ell(\beta_0, \beta_1) = \sum_{i=1}^{n} y_i - e^{\beta_0+\beta_1 x_i}$$
$$\frac{\partial}{\partial \beta_1}\ell(\beta_0, \beta_1) = \sum_{i=1}^{n} y_i x_i - x_i e^{\beta_0+\beta_1 x_i}$$

Donc, les équations à résoudre sont

$$\sum_{i=1}^{n} y_i - e^{\beta_0+\beta_1 x_i} = 0$$
$$\sum_{i=1}^{n} x_i(y_i - e^{\beta_0+\beta_1 x_i}) = 0.$$

5. La déviance est

$$D(y; \hat{\mu}) = 2(\ell_n(\tilde{\theta}) - \ell_n(\hat{\theta})).$$

Pour le modèle Binomial, on a que

$$\ell_n(\theta) = \sum_{i=1}^{n} \frac{y_i \ln\left(\frac{\mu_i}{1-\mu_i}\right) + \ln(1-\mu_i)}{1/m_i}.$$

Alors, dans le modèle complet, $\mu_i = y_i$ et on trouve

$$\ell_n(\tilde{\theta}) = \sum_{i=1}^{n} \frac{y_i \ln\left(\frac{y_i}{1-y_i}\right) + \ln(1-y_i)}{1/m_i}.$$

Dans le modèle développé avec le lien log, $\mu_i = \hat{\mu}_i$ et on trouve

$$\ell_n(\hat{\theta}) = \sum_{i=1}^{n} \frac{y_i \ln\left(\frac{\hat{\mu}_i}{1-\hat{\mu}_i}\right) + \ln(1-\hat{\mu}_i)}{1/m_i}.$$

Finalement, la déviance est

$$\begin{aligned} D(y;\hat{\mu}) &= \sum_{i=1}^{n} \frac{y_i \ln\left(\frac{y_i}{1-y_i}\right) + \ln(1-y_i)}{1/m_i} - \frac{y_i \ln\left(\frac{\hat{\mu}_i}{1-\hat{\mu}_i}\right) + \ln(1-\hat{\mu}_i)}{1/m_i} \\ &= \sum_{i=1}^{n} m_i \left[y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) + (1-y_i)\ln\left(\frac{1-y_i}{1-\hat{\mu}_i}\right)\right]. \end{aligned}$$

6. Pour la distribution Gamma, on a $V(t) = t^2$ et $b(t) = -\ln(-t)$.
   Résidus de Pearson :
   $$r_{P_i} = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i^2}} = \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}.$$
   Résidus d'Anscombe :
   $$\begin{aligned} A(t) &= \int_0^t \frac{\mathrm{d}s}{s^{2/3}} = 3t^{1/3} \\ \dot{A}(t) &= \frac{1}{s^{2/3}} \\ r_{A_i} &= \frac{A(y_i) - A(\hat{\mu}_i)}{\dot{A}(\hat{\mu}_i)\sqrt{V(\hat{\mu}_i)}} = \frac{3(y_i^{1/3} - \hat{\mu}_i^{1/3})}{\hat{\mu}_i^{1/3}}. \end{aligned}$$
   Résidus de déviance :
   $$\begin{aligned} D_i &= 2\left(-\frac{y_i}{y_i} - \ln(y_i) + \frac{y_i}{\hat{\mu}_i} + \ln(\hat{\mu}_i)\right) \\ &= 2\left(\ln\left(\frac{\hat{\mu}_i}{y_i}\right) + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}\right) \\ r_{D_i} &= sign(y_i - \hat{\mu}_i)\sqrt{2\left(\ln\left(\frac{\hat{\mu}_i}{y_i}\right) + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}\right)}. \end{aligned}$$

7. Cette solution est en anglais, vous pouvez poser vos questions sur le forum, s'il y a lieu.
   This is a two-factor model, «Device» takes three levels (M1, M2 and M3) and «Stress» takes 4 levels. The baseline group is M1 device at stress level I. An analysis of deviance is carried out to assess if the parameters for the devices are significant.

```
> glm4 <- glm(Failures~Level*Machine,family=poisson,data=stresstest)
> anova(glm4)
Analysis of Deviance Table

Model: poisson, link: log

Response: Failures
```

```
Terms added sequentially (first to last)

             Df Deviance Resid. Df Resid. Dev
NULL                           11     35.844
Level         3  20.8567         8     14.987
Machine       2  12.2154         6      2.772
Level:Machine 6   2.7719         0      0.000
> qchisq(0.95,6)
[1] 12.59159
> qchisq(0.95,2)
[1] 5.991465
```

The model

$$\text{Stress+Device+Stress.Device}$$

is fitted first. The change in deviance from the simpler model Stress+Device is 2.7719 on 6 degrees of freedom, which is not significant when compared to $\chi^2_{(6,0.95)} = 12.59$. Hence, the model Stress+Device is an adequate simplification of the more complex model. If we then test for the significance of the Device parameters, we find that the change in deviance from the simpler model Stress is 12.2154 on 2 degrees of freedom, which is significant because $\chi^2_{(2,0.95)} = 5.99$. From this analysis, we can conclude that there is a significant difference between the failure rates of the different devices.

8. **Réclamations d'assurance automobile au Royaume-Uni.**

   a) En R, on obtient

```
> Bcar <- read.table("BritishCar.csv",header=TRUE,sep=";")
> modinv <- glm(AvCost~OwnerAge+Model+CarAge,family=Gamma,data=Bcar)
> summary(modinv)

Call:
glm(formula = AvCost ~ OwnerAge + Model + CarAge, family = Gamma,
    data = Bcar)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.85536  -0.13930  -0.00821   0.07444   1.49969

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.0033233  0.0004038   8.230 4.42e-13 ***
OwnerAge21-24  0.0006043  0.0004159   1.453  0.14908
```

```
OwnerAge25-29  0.0003529  0.0003933   0.897  0.37163
OwnerAge30-34  0.0011783  0.0004572   2.577  0.01130 *
OwnerAge35-39  0.0016372  0.0004990   3.281  0.00139 **
OwnerAge40-49  0.0012039  0.0004592   2.622  0.01000 **
OwnerAge50-59  0.0010998  0.0004511   2.438  0.01638 *
OwnerAge60+    0.0012390  0.0004619   2.682  0.00845 **
ModelB        -0.0002817  0.0004049  -0.696  0.48806
ModelC        -0.0006502  0.0003906  -1.664  0.09893 .
ModelD        -0.0018235  0.0003481  -5.239 7.96e-07 ***
CarAge10+      0.0033776  0.0004747   7.115 1.24e-10 ***
CarAge4-7      0.0003393  0.0002723   1.246  0.21539
CarAge8-9      0.0017423  0.0003575   4.873 3.75e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.1074529)

    Null deviance: 27.841  on 122  degrees of freedom
Residual deviance: 11.511  on 109  degrees of freedom
  (5 observations deleted due to missingness)
AIC: 1400.7

Number of Fisher Scoring iterations: 5
```

b) On a utilisé un lien inverse, alors $E[Y_i] = \frac{1}{\eta_i}$. Puisque les variables explicatives prennent toutes leur niveau de base, on a que $\hat{\eta}_i = \hat{\beta}_0 = 0.0033233$ et

$$\widehat{E[Y_i]} = 0.0033233^{-1} = 300.91.$$

c) Puisqu'on a utilisé un lien inverse, un coefficient plus élevé implique une diminution de l'espérance du coût de la réclamation, alors qu'un coefficient négatif signifie une augmentation de cette espérance. Ici on observe que les sept coefficients sont positifs, alors la catégorie d'âge ayant une espérance de coût la plus élevée est la catégorie de base, 17-20 ans. Le coût moyen semble ensuite relativement élevé pour les jeunes entre 21 et 29 ans. La catégorie d'âge avec coût de réclamation minimal est 35-39 ans, puis la moyenne semble relativement stable pour les détenteurs de police plus âgés.

d) Les trois coefficients pour la variable modèle sont négatifs, ce qui signifie que les réclamations pour les véhicules de type A (niveau de base) sont moins élevées en moyenne que celles pour les autres types de véhicule. Les réclamatios pour les véhicules du modèle D semblent particulièrement coûteuse car le coefficient est beaucoup plus grand en valeur absolue que les autres.

e) De la même façon, on observe que d'augmenter l'âge du véhicule diminue le coût moyen des réclamations.

f) Pour un détenteur de police entre 17 et 20 ans, avec un véhicule de type D âgé de un à 3 ans, on trouve que

$$\widehat{E[Y_i]} = \frac{1}{\hat{\beta}_0 + \hat{\beta}_D^{MODEL}} = \frac{1}{0.0033233 - 0.0018235} = 666.76.$$

g) Pour un détenteur de police entre 35 et 39 ans, avec un véhicule de type A âgé de plus de 10 ans, on trouve que

$$\widehat{E[Y_i]} = \frac{1}{\hat{\beta}_0 + \hat{\beta}_{35-39}^{OWNERAGE} + \hat{\beta}_{10+}^{CARAGE}} = \frac{1}{0.0033233 + 0.0016372 + 0.0033776} = 119.93.$$

h) La déviance $D(y, \hat{\mu}) = 11.511$ est donnée dans la sortie `R` pour la sous-question a). On a que

$$\frac{D(y, \hat{\mu})}{\hat{\phi}} = \frac{11.511}{0.1074529} = 107.126,$$

ce qui est très près de $n - p' = 109$. Le modèle semble donc adéquat.

i) Les résidus sont calculés avec les formules trouvées à la question 6. Il faut d'abord enlever les données manquantes du vecteur contenant les coûts moyens. On obtient les graphiques de la Figure 1.

j) a. Le modèle avec le lien logarithmique est

```
> modlog <- glm(AvCost~OwnerAge+Model+CarAge,family=Gamma(link=log),data=Bcar)
> summary(modlog)

Call:
glm(formula = AvCost ~ OwnerAge + Model + CarAge, family = Gamma(link = log),
    data = Bcar)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.84819  -0.12796  -0.00834   0.08552   1.20066

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.711739   0.103835  55.008  < 2e-16 ***
OwnerAge21-24 -0.108159   0.114547  -0.944   0.3471
OwnerAge25-29  0.005223   0.113170   0.046   0.9633
OwnerAge30-34 -0.288090   0.113170  -2.546   0.0123 *
OwnerAge35-39 -0.331420   0.114547  -2.893   0.0046 **
```

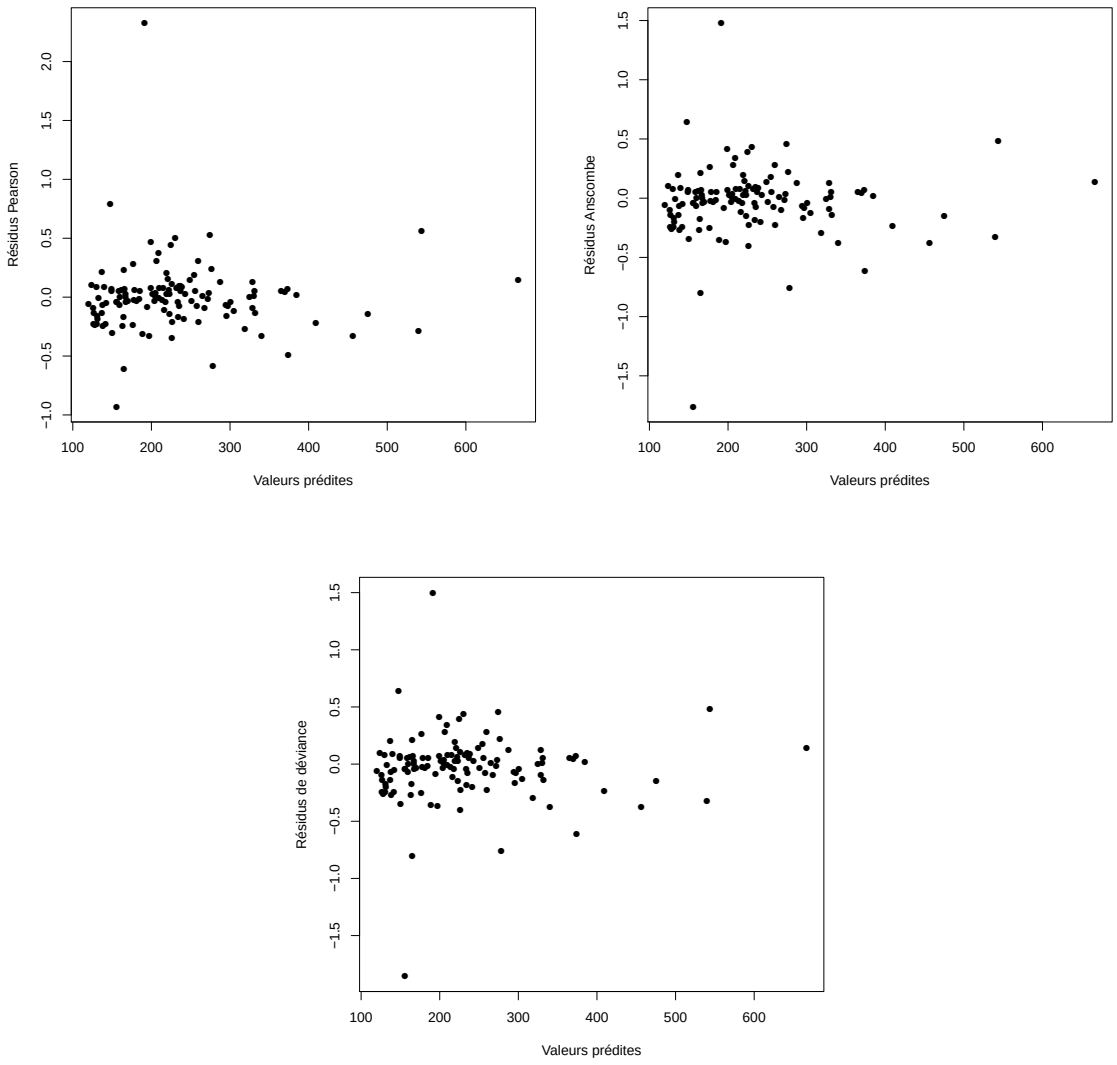

FIGURE 1: Résidus pour GLM Gamma

```
OwnerAge40-49 -0.280775   0.113170  -2.481   0.0146 *
OwnerAge50-59 -0.238136   0.113170  -2.104   0.0377 *
OwnerAge60+   -0.283521   0.113170  -2.505   0.0137 *
ModelB         0.057951   0.075447   0.768   0.4441
ModelC         0.154588   0.076115   2.031   0.0447 *
ModelD         0.472290   0.078497   6.017 2.43e-08 ***
CarAge10+     -0.735513   0.078497  -9.370 1.17e-15 ***
CarAge4-7     -0.111412   0.075447  -1.477   0.1426
CarAge8-9     -0.422538   0.076115  -5.551 2.02e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.0910768)

    Null deviance: 27.841  on 122  degrees of freedom
Residual deviance: 11.263  on 109  degrees of freedom
  (5 observations deleted due to missingness)
AIC: 1398

Number of Fisher Scoring iterations: 7
```

b. Avec ce modèle $E[Y_i] = e^{\eta_i}$. Puisque les variables explicatives prennent toutes leur niveau de base, on a que $\hat{\eta}_i = \hat{\beta}_0 = 5.711739$ et

$$\widehat{E[Y_i]} = e^5.711739 = 302.39.$$

Cela ne diffère pas beaucoup du résultat trouvé en b).

c-d-e. Puisqu'on a utilisé un lien logarithmique, on a un modèle multiplicatif. Si $e^\beta > 1$, alors l'espérance du coût augmente, alors que si $e^\beta < 1$ alors l'espérance du coût diminue. On peut donc tirer des conclusions similaires à celles en c), d) et e).

f. Pour un détenteur de police entre 17 et 20 ans, avec un véhicule de type D âgé de un à 3 ans, on trouve que

$$\widehat{E[Y_i]} = \exp \hat{\beta}_0 + \hat{\beta}_D^{MODEL} = \exp 5.711739 + 0.472290 = 484.94.$$

On note que cette valeur est beaucoup moins élevée que celle obtenue en f).

g. Pour un détenteur de police entre 35 et 39 ans, avec un véhicule de type A âgé de plus de 10 ans, on trouve que

$$\widehat{E[Y_i]} = \exp \hat{\beta}_0 + \hat{\beta}_{35-39}^{OWNERAGE} + \hat{\beta}_{10+}^{CARAGE} = \exp 5.711739 - 0.331420 - 0.735513 = 104.04.$$

h. La déviance $D(y, \hat{\mu}) = 11.263$ est donnée dans la sortie `R` pour la sous-question a). On a que

$$\frac{D(y, \hat{\mu})}{\hat{\phi}} = \frac{11.263}{0.0910768} = 123.66,$$

ce qui est moins près de $n - p' = 109$ que pour le modèle avec le lien inverse. Le modèle semble donc moins adéquat.

9. (a) If x_i is treated as a factor predictor with 11 levels, the linear predictor is written as

$$\eta_i = \beta_0 + \beta_i,\ i = 1, ..., 11$$

and $\beta_1 = 0$. The binomial density is the following :

$$f_Y(y_i) = \binom{m_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{m_i - y_i},$$

which can be rewritten in a exponential family representation as :

$$f_Y(y_i) = \exp\left[y_i \ln\left(\frac{\pi_i}{1 - \pi_i}\right) + m \ln(1 - \pi_i) + \ln\binom{m_i}{y_i} \right].$$

Hence, the canonical parameter is $\theta_i = \ln\left(\frac{\pi_i}{1-\pi_i}\right)$ and the canonical link is the logit link. Thus,

$$\eta_i = \ln\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_i$$
$$\pi_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}} = \frac{e^{\beta_0 + \beta_i}}{1 + e^{\beta_0 + \beta_i}}$$

The expression of the density in the reparametrization is then

$$\begin{aligned} f_Y(y_i) &= \binom{m_i}{y_i} \left(\frac{e^{\beta_0+\beta_i}}{1 + e^{\beta_0+\beta_i}}\right)^{y_i} \left(\frac{1}{1 + e^{\beta_0+\beta_i}}\right)^{m_i - y_i} \\ &= \binom{m_i}{y_i} \frac{e^{y_i(\beta_0+\beta_i)}}{(1 + e^{\beta_0+\beta_i})^{m_i}} \end{aligned}$$

The likelihood L and the log-likelihood l are shown below :

$$L(\beta_0, ..., \beta_{11}; y_1, ..., y_{11}) = \prod_{i=1}^{11} \binom{m_i}{y_i} \frac{e^{y_i(\beta_0+\beta_i)}}{(1 + e^{\beta_0+\beta_i})^{m_i}}$$
$$\ell(\beta_0, ..., \beta_{11}; y_1, ..., y_{11}) = \sum_{i=1}^{11} \left[\ln\binom{m_i}{y_i} + y_i(\beta_0 + \beta_i) - m_i \ln(1 + e^{\beta_0+\beta_i}) \right]$$

We have

$$\frac{\partial \ell}{\partial \beta_0} = \sum_{i=1}^{11} \left[y_i - m_i \frac{e^{\beta_0+\beta_i}}{1+e^{\beta_0+\beta_i}} \right]$$
$$\frac{\partial \ell}{\partial \beta_i} = y_i - m_i \frac{e^{\beta_0+\beta_i}}{1+e^{\beta_0+\beta_i}},\ i = 2, ..., 11,$$

and $\beta_1 = 0$ by constraint of the model. The maximum likelihood estimators for the parameters are derived by solving the system of equations $\frac{\partial \ell}{\partial \beta_i} = 0$, $i = 0, ..., 11$:

$$\sum_{i=1}^{11} \left[y_i - m_i \frac{e^{\hat{\beta}_0+\hat{\beta}_i}}{1+e^{\hat{\beta}_0+\hat{\beta}_i}} \right] = 0$$
$$y_i - m_i \frac{e^{\hat{\beta}_0+\hat{\beta}_i}}{1+e^{\hat{\beta}_0+\hat{\beta}_i}} = 0,\ i = 2, ..., 11,$$
$$\Rightarrow \hat{\beta}_0 + \hat{\beta}_i = \ln\left(\frac{y_i}{m_i - y_i}\right),\ i = 2, ..., 11,$$

Using the first equation and replacing $\hat{\beta}_0 + \hat{\beta}_i$ by $\ln\left(\frac{y_i}{m_i-y_i}\right)$,

$$y_1 - m_1 \frac{e^{\hat{\beta}_0}}{1+e^{\hat{\beta}_0}} + \sum_{i=2}^{11} \left[y_i - m_i \frac{\left(\frac{y_i}{m_i-y_i}\right)}{1+\left(\frac{y_i}{m_i-y_i}\right)} \right] = 0$$
$$y_1 - m_1 \frac{e^{\hat{\beta}_0}}{1+e^{\hat{\beta}_0}} + \sum_{i=2}^{11} \left[y_i - m_i \frac{y_i}{m_i} \right] = 0$$
$$y_1 - m_1 \frac{e^{\hat{\beta}_0}}{1+e^{\hat{\beta}_0}} = 0$$
$$\hat{\beta}_0 = \ln\left(\frac{y_1}{m_1 - y_1}\right)$$
$$\hat{\beta}_i = \ln\left(\frac{y_i}{m_i - y_i}\right) - \hat{\beta}_0 = \ln\left(\frac{y_i/(m_i - y_i)}{y_1/(m_1 - y_1)}\right),\ i = 2, ..., 11.$$

The estimates of the model parameters are easily found in R as follows :

```
> (beta0 <- log(y[1]/(m[1]-y[1])))
[1] -2.917771
```

```
> (beta <- c(0,log(y[-1]/(m[-1]-y[-1]))-beta0))
[1] 0.0000000 0.4122448 0.6151856 0.6740261 1.3083328 1.7137979 1.6184877
[8] 2.7636201 3.5239065 3.0178542 3.2542430
```

Hence, here

$$\hat{\beta} = (-2.9178, 0, 0.4122, 0.6152, 0.6740, 1.3083, 1.7138, 1.6185, 2.7636, 3.5239, 3.0179, 3.2542)^T.$$

As a consistency check following from the invariance property of maximum likelihood estimation, we can verify that the estimates of π_i using the expit function are equal to the MLE estimates $\hat{\pi}_i = \frac{y_i}{m_i}$:

```
> (pi <- exp(beta0+beta)/(1+exp(beta0+beta)))
 [1] 0.05128205 0.07547170 0.09090909 0.09589041 0.16666667 0.23076923
 [7] 0.21428571 0.46153846 0.64705882 0.52500000 0.58333333
> y/m
 [1] 0.05128205 0.07547170 0.09090909 0.09589041 0.16666667 0.23076923
 [7] 0.21428571 0.46153846 0.64705882 0.52500000 0.58333333
```

(b) The Binomial GLM model with logit link and the linear predictor $\eta_i = \beta_0 + \beta_1 x_i$, $i = 1, ..., 11$ is fitted to the data using `R` and the command :

```
glm(cbind(y,m-y)~x,family=binomial).
```

The estimates of the parameters are $\hat{\beta}_0 = -3.6070615$ and $\hat{\beta}_1 = 0.0009121$, with standard error $SE(\hat{\beta}_0) = 0.3533875$ and $SE(\hat{\beta}_1) = 0.0001084$.

(c) The Binomial GLM model with probit link and the linear predictor $\eta_i = \beta_0 + \beta_1 x_i$, $i = 1, ..., 11$ is fitted to the data using `R` and the command :

```
glm(cbind(y,m-y)~x,family=binomial(link=probit)).
```

The estimates of the parameters are $\hat{\beta}_0 = -2.080$ and $\hat{\beta}_1 = 5.230 \times 10^{-4}$, with standard error $SE(\hat{\beta}_0) = 0.1852$ and $SE(\hat{\beta}_1) = 5.973 \times 10^{-5}$.

(d) The Binomial GLM model with complementary log-log link and the linear predictor $\eta_i = \beta_0 + \beta_1 x_i$, $i = 1, ..., 11$ is fitted to the data using `R` and the command :

```
glm(cbind(y,m-y)~x,family=binomial(link=cloglog)).
```

The estimates of the parameters are $\hat{\beta}_0 = -3.360$ and $\hat{\beta}_1 = 7.480 \times 10^{-4}$, with standard error $SE(\hat{\beta}_0) = 0.3061$ and $SE(\hat{\beta}_1) = 8.622 \times 10^{-5}$.

(e) Predictions can be found using the inverse of the link function. For the model with canonical link (model from b), we find that

$$\hat{y}_{2000} = \frac{e^{\hat{\beta}_0 + 2000\hat{\beta}_1}}{1 + e^{\hat{\beta}_0 + 2000\hat{\beta}_1}} = 0.1439472.$$

Alternatively, the command `predict(modelb,data.frame(x=2000),type="response",se.fit=TRUE)` can be used to calculate the predictions and associated standard errors. The resulting predictions and standard errors are presented in Table 1. The probability of developing fissures after 2000 hours of operations is 14.39% according to model b, and the standard deviation is 2.08%. This prediciton is quite comparable with the complementary log-log model (d), for which the estimated probability of developing fissures is 14.36%, with a standard error of 2.03%. The precision is slightly better in this model than the two others due to a smaller variance. The estimated probability with Model c, using the probit link, is higher at 15.06%, with standard error of 2.06%.

	Model b (logit link)	Model c (probit link)	Model d (compl. log-log link)
$\hat{y}_{2000}$	0.1439472	0.150615	0.1436447
$SE(\hat{y}_{2000})$	0.02080256	0.02062154	0.02030803

TABLE 1: Predictions and Standard Errors for Predictions for the 3 Models

(f) Figure 2 shows a plot of the data points along with the three fitted lines. It was obtained using the following code in `R`, where `ilogit, iprobit` and `icloglog` are the inverse of the corresponding link functions :

```
plot(x,y/m,pch=19,xlab="Number of Hours of Operation (x)",
  ylab="Prob of Developing Fissures")
j <- seq(0,4800,1)
lines(j,ilogit(coef(modelb)[1]+coef(modelb)[2]*j),lwd=2)
lines(j,iprobit(coef(modelc)[1]+coef(modelc)[2]*j),lty=2,col=2,lwd=2)
lines(j,icloglog(coef(modeld)[1]+coef(modeld)[2]*j),lty=3,col=4,lwd=2)
legend("topleft",legend=c("Logit","Probit","Complementary log-log"),
lty=c(1,2,3),col=c(1,2,4),lwd=rep(2,3))
```

It is easy to observe that the fit is better when the number of hours of operations is lower, it seems that the variance of the observations is increasing with the predictor. This is expected in a generalized linear model framework. The three fitted lines are slightly different. The probit link produces lower estimates in the tails and higher estimates in the middle of the range of the predictors. It seems like this model is less representative of the data than the others. The complementary log-log model (d) predicts higher probabilities of failures in the extremes of the range of the predictors. This seems to fit the data well, and recall that the variance of the predictions where also smaller than other models in this case, which is a desirable property. The line for the model with canonical link is between the two others. It could also be a reasonable model for the data.

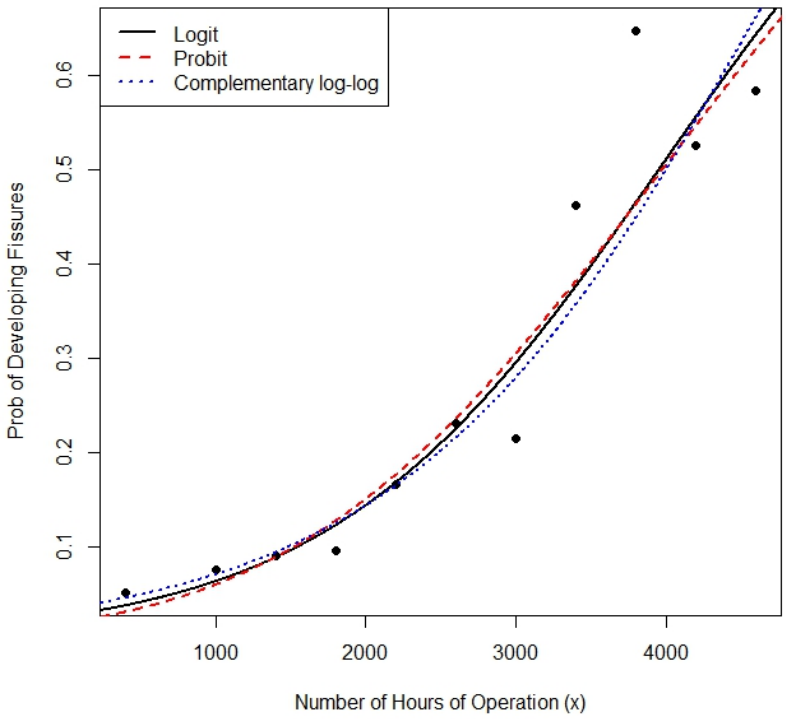

FIGURE 2: Logistic, Probit and Complementary Log-Log Model Fit