STT5100 - Automne 2019 - Examen Final (GLM)

Arthur Charpentier

Examen **A**

Les calculatrices sont autorisées. Les documents sont en revanche interdits, sauf une page d'aide mémoire. L'examen dure 3 heures, mais toute sortie avant midi est autorisée, et sera définitive.

Dans les feuilles qui suivent, il y a 30 questions relatives au cours sur les modèles linéaires. Pour chaque question (sauf deux), cinq réponses sont proposées. Une seule est valide, et vous ne devez en retenir qu'une,

- vous gagnez 1 point par bonne réponse
- vous ne perdez pas de points pour une mauvaise réponse
- vous ne gagnez pas de point pour plusieurs réponses

Aucune justification n'est demandée. Une question repose sur un graphique qu'il faudra tracer sur la feuille de réponses (au dos). Votre note finale est le total des points (sur 30). Il y a une 31ème question, bonus. Une prédiction parfaite (sur 30) donnera un point bonus qui s'ajoutera à la note.

La page de réponses est au dos de celle que vous lisez présentement : merci de décrocher ladite feuille et de ne rendre que cette dernière, après avoir indiqué votre code permanent en haut à gauche.

Merci de cocher le carré en bleu ou en noir. En cas d'erreur, vous pouvez cocher une autre case en rouge. Seule cette dernière sera alors retenue.

Le surveillant ne répondra à aucune question durant l'épreuve : en cas de soucis sur une question (interprétation possiblement fausse, typo, etc), vous pouvez mettre un court commentaire sur la feuille de réponses.

Formulaire : Quantiles de lois usuelles. Exemple pour une loi normale - $Z \sim \mathcal{N}(0,1)$, $\mathbb{P}(Z \leqslant 2.326) = 99\%$.

	75%	90%	95%	97.5%	99%	99.5%	99.9%	99.95%
Loi normale	0.674	1.282	1.645	1.960	2.326	2.576	3.090	3.291
Chi^2, $\chi^2(5)$	6.626	9.236	11.070	12.833	15.086	16.750	20.515	22.105
Chi^2, $\chi^2(4)$	5.385	7.779	9.488	11.143	13.277	14.860	18.467	19.997
Chi^2, $\chi^2(3)$	4.108	6.251	7.815	9.348	11.345	12.838	16.266	17.730
Chi^2, $\chi^2(2)$	2.773	4.605	5.991	7.378	9.210	10.597	13.816	15.202
Chi^2, $\chi^2(1)$	1.323	2.706	3.841	5.024	6.635	7.879	10.828	12.116

La densité / mesure de probabilité d'une variable aléatoire dans la famille exponentielle s'écrit

$$f(y;\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right)$$

On appelle logit la fonction $(0,1) \mapsto \mathbb{R}$ définie par $\text{logit}(p) = \log(p/(1-p))$. Et la prime pure désigne l'espérance mathématique de la perte.

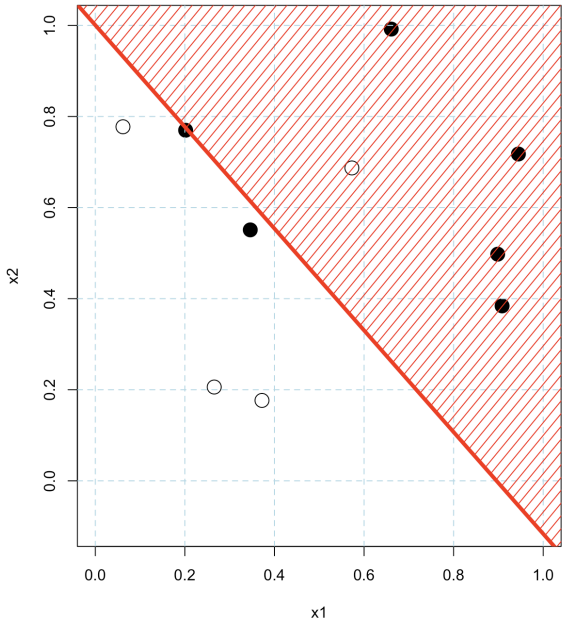
Je corrige ici le sujet A, le sujet B contenait les mêmes questions dans un ordre différents, et les applications numériques - ici 11 et 17 - sont différentes, mais proches en particulier pour l'exercice 11, les 0 et les 1 ont été intervertis pour y.

Code permanent :

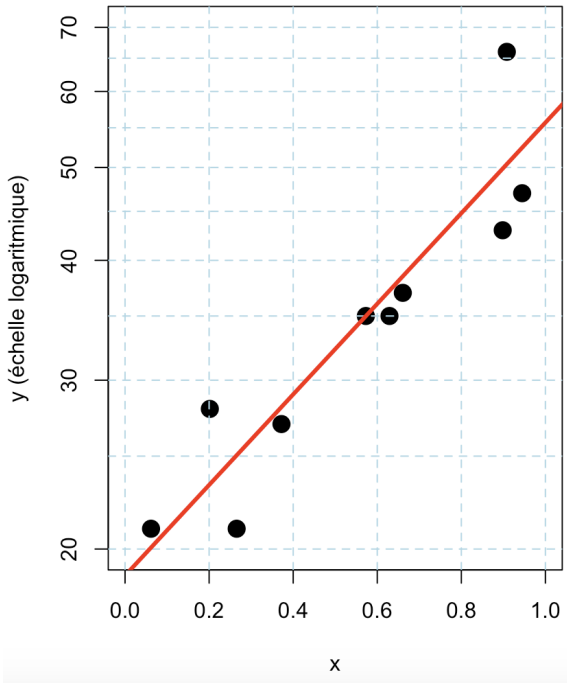
Sujet : A

question 1	☐ A	■ B	☐ C	☐ D	☐ E
question 2	☐ A	☐ B	■ C	☐ D	☐ E
question 3	☐ A	☐ B	■ C	☐ D	☐ E
question 4	■ A	☐ B	☐ C	☐ D	☐ E
question 5	☐ A	☐ B	☐ C	■ D	☐ E
question 6	☐ A	☐ B	☐ C	☐ D	■ E
question 7	■ A	☐ B	☐ C	☐ D	☐ E
question 8	☐ A	☐ B	☐ C	■ D	☐ E
question 9	☐ A	☐ B	☐ C	■ D	☐ E
question 10	☐ A	■ B	☐ C	☐ D	☐ E
question 11	Figure à droite (à compléter)				
question 12	☐ A	☐ B	☐ C	☐ D	■ E
question 13	■ A	☐ B	☐ C	☐ D	☐ E
question 14	☐ A	■ B	☐ C	☐ D	☐ E
question 15	☐ A	☐ B	☐ C	■ D	☐ E
question 16	☐ A	☐ B	☐ C	☐ D	■ E
question 17	Figure à droite (à compléter)				
question 18	☐ A	☐ B	■ C	☐ D	☐ E
question 19	☐ A	☐ B	■ C	☐ D	☐ E
question 20	☐ A	☐ B	■ C	☐ D	☐ E
question 21	■ A	☐ B	☐ C	☐ D	☐ E
question 22	■ A	☐ B	☐ C	☐ D	☐ E
question 23	☐ A	☐ B	■ C	☐ D	☐ E
question 24	☐ A	☐ B	■ C	☐ D	☐ E
question 25	☐ A	☐ B	■ C	☐ D	☐ E
question 26	☐ A	☐ B	☐ C	☐ D	■ E
question 27	☐ A	☐ B	☐ C	☐ D	■ E
question 28	☐ A	☐ B	☐ C	☐ D	■ E
question 29	☐ A	☐ B	☐ C	■ D	☐ E
question 30	☐ A	☐ B	☐ C	☐ D	■ E
question 31	Combien de bonnes réponses pensez vous avoir ?				

question 11 :



question 17 :



1 On estime trois modèles GLM, avec $g(\mathbb{E}[Y|X_1,X_2]) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$. On a (1) une régression Gamma (2) une régression Poisson et (3) une régression binomiale, toutes trois avec leur lien canonique. En estimant les modèles, on obtient les mêmes estimations, à savoir

$$\widehat{\beta}_0 = 2,\ \widehat{\beta}_1 = 1,\ \text{et}\ \widehat{\beta}_2 = -1.$$

Soit $\widehat{y}_m$ la prévision obtenue au point $x_1 = 2$ et $x_2 = 1$ par le modèle m (avec $m \in \{1,2,3\}$). Quelle affirmation est juste parmi les suivantes

A) $\widehat{y}_1 < \widehat{y}_2 < \widehat{y}_3$

B) $\widehat{y}_1 < \widehat{y}_3 < \widehat{y}_2$

C) $\widehat{y}_2 < \widehat{y}_1 < \widehat{y}_3$

D) $\widehat{y}_2 < \widehat{y}_3 < \widehat{y}_1$

E) la réponse n'est pas donnée par les affirmations ci-dessus

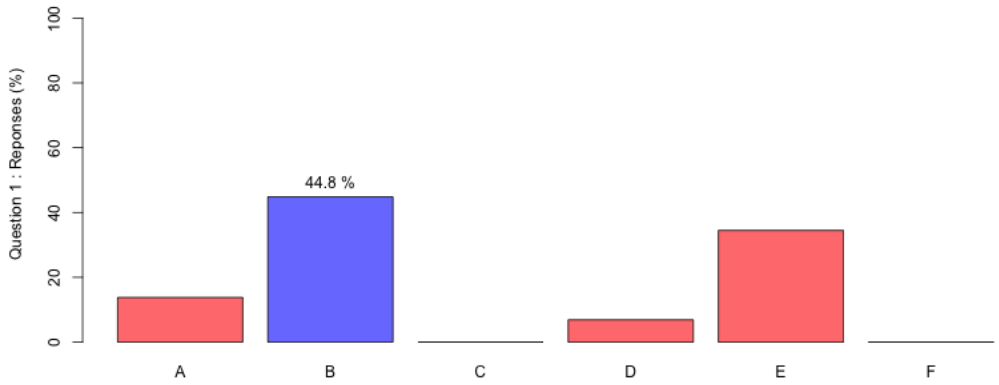
Il s'agissait d'un exercice de l'examen MAS-I de la CAS du printemps 2018. Pour les trois modèles, on a l'estimation de $g(\mathbb{E}[Y|X_1,X_2])$ qui vaut $2 + 1 \cdot 2 - 1 \cdot 1 = 3$. Donc

$$\widehat{y}_1 = g^{-1}(3) = \frac{1}{3} \text{ car } g(\mu) = \mu^{-1} \text{ et donc } g^{-1}(\mu) = \mu^{-1},$$

$$\widehat{y}_2 = g^{-1}(3) = \exp(3) \sim 20.08 \text{ car } g(\mu) = \log \mu \text{ et donc } g^{-1}(\mu) = \exp(\mu),$$

$$\widehat{y}_3 = g^{-1}(3) = \frac{e^3}{1+e^3} \sim 0.95 \text{ car } g(\mu) = \log \frac{\mu}{1-\mu} \text{ et donc } g^{-1}(\mu) = \frac{e^\mu}{1+e^\mu},$$

aussi $\widehat{y}_1 < \widehat{y}_3 < \widehat{y}_2$, qui correspond à la réponse B.



2 On nous donne les informations suivantes au sujet d'un GLM : les variables réponses y_i sont supposées suivre des lois normales indépendantes, de moyenne inconnue μ_i et de variance inconnue elle aussi σ^2. On note $\widehat{\mu}_i$ les moyennes prédites. Que vaut la scaled deviance de ce modèle ?

A) $\sum_{i=1}^{n} (y_i - \widehat{\mu}_i)^2$

B) $\sum_{i=1}^{n}(y_i-\mu_i)^2$

C) $\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\widehat{\mu}_i)^2$

D) $\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\mu_i)^2$

E) la réponse n'est pas donnée par les affirmations ci-dessus

Réponse C. Je vais le reprendre tranquillement, puisqu'il y a eu des questions, pour expliquer la distinction entre deviance et scaled deviance... Par définition, la scaled deviance est définie par

$$D^\star = 2\Big(\underbrace{\log \mathcal{L}(\boldsymbol{y})}_{=\log \mathcal{L}_{\text{saturée}}} - \underbrace{\log \mathcal{L}(\widehat{\boldsymbol{\mu}})}_{=\log \mathcal{L}_{\text{maximale}}} \Big)$$

et la deviance est définie par

$$D = \phi D^\star = 2\phi\left(\log \mathcal{L}(\boldsymbol{y}) - \log \mathcal{L}(\widehat{\boldsymbol{\mu}})\right)$$

Lançons nous : la log-vraisemblance dans un modèle Gaussien (avec les notations usuelles) s'écrit

$$\log \mathcal{L} = \log\left(\frac{1}{(2\pi\sigma^2)^{-n/2}}\exp\left[-\sum_{i=1}^{n}\frac{(y_i-\mu_i)^2}{2\sigma^2}\right]\right)$$

soit

$$\log \mathcal{L} = -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \sum_{i=1}^{n}\frac{(y_i-\mu_i)^2}{2\sigma^2}$$

On remplace alors μ_i par y_i pour la vraisemblance saturée, et par $\widehat{\mu}_i$ pour la valeur obtenue au maximum (de la vraisemblance), ce qui donne

$$\log \mathcal{L}_{\text{saturée}} = -\frac{n}{2}\log(2\pi\sigma^2)$$

et

$$\log \mathcal{L}_{\text{maximale}} = -\frac{n}{2}\log(2\pi\sigma^2) + \sum_{i=1}^{n}\frac{(y_i-\widehat{\mu}_i)^2}{2\sigma^2}$$

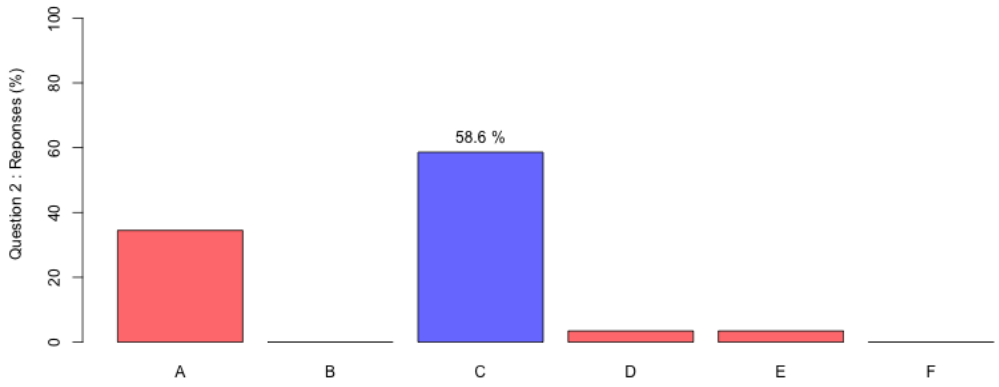
donc si maintenant on soustrait, on obtient

$$D^\star = \sum_{i=1}^{n}\frac{(y_i-\widehat{\mu}_i)^2}{2\sigma^2} = \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\widehat{\mu}_i)^2$$

ce qui correspond effectivement à la réponse C. En revanche, la déviance est obtenue en multipliant par ϕ, le paramètre de nuisance, qui correspond à σ^2 dans le modèle Gaussien,

$$D = \phi D^\star = \sum_{i=1}^{n}(y_i-\widehat{\mu}_i)^2$$

- qui correspondait à la réponse A (également fortement donnée).



On va continuer un peu la discussion, parce que j'ai reçu plusieurs messages à ce sujet. La définition de la déviance que je donne est celle que j'avais donnée en cours. C'est celle de Mc Cullagh & Nelder (1989) Generalized linear models, publié chez Chapman & Hall (à gauche ci-dessous). Parmi les références du cours de Jong & Heller (2008) Generalized Linear Models for Insurance Data est malheureusement plus ambigu (à droite ci-dessous)

There are advantages in using as the goodness-of-fit criterion, not the log likelihood $l(\boldsymbol{\mu};\mathbf{y})$ but a particular linear function, namely

$$D^*(\mathbf{y};\boldsymbol{\mu}) = 2l(\mathbf{y};\mathbf{y}) - 2l(\boldsymbol{\mu};\mathbf{y}),$$

which we call the *scaled deviance*. Note that, for the exponential-family models considered here, $l(\mathbf{y};\mathbf{y})$ is the maximum likelihood achievable for an exact fit in which the fitted values are equal to the observed data. Because $l(\mathbf{y};\mathbf{y})$ does not depend on the parameters, maximizing $l(\boldsymbol{\mu};\mathbf{y})$ is equivalent to minimizing $D^*(\mathbf{y};\boldsymbol{\mu})$ with respect to $\boldsymbol{\mu}$, subject to the constraints imposed by the model.

For Normal-theory linear regression models with known variance σ^2, we have for a single observation

$$f(y;\mu) = \frac{1}{\sqrt{(2\pi\sigma^2)}}\exp\Big(-\frac{(y-\mu)^2}{2\sigma^2}\Big),$$

so that the log likelihood is

$$l(\mu;y) = -\tfrac{1}{2}\log(2\pi\sigma^2) - (y-\mu)^2/(2\sigma^2).$$

Setting $\mu = y$ gives the maximum achievable log likelihood, namely

$$l(y;y) = -\tfrac{1}{2}\log(2\pi\sigma^2),$$

so that the scaled deviance function is

$$D^*(y;\mu) = 2\{l(y;y) - l(\mu;y)\} = (y-\mu)^2/\sigma^2.$$

The value of the saturated log-likelihood is

$$\tilde{\ell} \equiv \sum_{i=1}^{n}\left\{\ln c(y_i,\phi) + \frac{y_i\tilde{\theta}_i - a(\tilde{\theta}_i)}{\phi}\right\},$$

which is the maximum possible log-likelihood for y given the response distribution specified by $a(\theta)$. This value is compared to $\hat{\ell}$, the value of the maximum of the log-likelihood based on y and the given explanatory variables. The "deviance," denoted as Δ, is defined as a measure of distance between the saturated and fitted models:

$$\Delta \equiv 2(\tilde{\ell} - \hat{\ell}).$$

Table 5.2. *Deviance for exponential family response distributions*

Distribution	Deviance Δ
Normal	$\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\hat{\mu}_i)^2$
Poisson	$2\sum_{i=1}^{n}\left\{y_i\ln\left(\frac{y_i}{\hat{\mu}_i}\right)-(y_i-\hat{\mu}_i)\right\}$
Binomial	$2\sum_{i=1}^{n}n_i\left\{y_i\ln\left(\frac{y_i}{\hat{\mu}_i}\right)+(n_i-y_i)\ln\left(\frac{n_i-y_i}{1-\hat{\mu}_i}\right)\right\}$
Gamma	$2\nu\sum_{i=1}^{n}\left\{-\ln\left(\frac{y_i}{\hat{\mu}_i}\right)+\frac{y_i-\hat{\mu}_i}{\hat{\mu}_i}\right\}$
Inverse Gaussian	$\frac{1}{\sigma^2}\sum_{i=1}^{n}\frac{(y_i-\hat{\mu}_i)^2}{\hat{\mu}_i^2 y_i}$
Negative binomial	$2\sum_{i=1}^{n}\left\{y_i\ln\left(\frac{y_i}{\hat{\mu}_i}\right)-(y_i+\frac{1}{\kappa})\ln\left(\frac{y_i+1/\kappa}{\hat{\mu}_i+1/\kappa}\right)\right\}$

SAS notes. In SAS the deviance is called the scaled deviance, and also the residual deviance by the SAS manual, while $\phi\,\Delta$ is called the deviance. This means that it is the *scaled deviance* that is relevant. (Both scaled and unscaled deviances are given in SAS output.)

Dans cet ouvrage,

$$2\Big(\underbrace{\log\mathcal{L}(\boldsymbol{y})}_{=\log\mathcal{L}_{\text{saturée}}} - \underbrace{\log\mathcal{L}(\widehat{\boldsymbol{\mu}})}_{=\log\mathcal{L}_{\text{maximale}}}\Big)$$

est appelé deviance - comme repris dans la table 5.2. Mais la note de bas de tableau est importante the deviance is called the scaled deviance. En réalité, c'est effectivement la scaled deviance, telle que définie par Mc Cullagh & Nelder (1989), parmi d'autres, comme Ohlsson & Johansson (2010) - que je cite aussi parmi les références

Let $\ell(\hat{\boldsymbol{\mu}})$ denote the log-likelihood for our data as a function of the estimated mean vector $\hat{\boldsymbol{\mu}}$. If the number of non-redundant parameters, r, equals the number of observations (tariff cells) n, we can get a perfect fit by setting all $\hat{\mu}_i = y_i$. These are in fact the ML-estimates in this case, since they trivially satisfy the ML equations in (2.28), including the restriction of (2.31). (The latter fact actually calls for some reflection.) This case with $r = n$ is called the *saturated model*. While this model is trivial and of no practical interest, it is often used as a benchmark in measuring the goodness-of-fit of other models, since it has perfect fit. Such a measure is the *scaled deviance* D^*, which is defined as the likelihood-ratio-test (LRT) statistic of the model under consideration, against the saturated model. The LRT statistic is two times the logarithm of the likelihood-ratio, i.e.

$$D^* = D^*(\mathbf{y}, \hat{\boldsymbol{\mu}}) = 2\big[\ell(\mathbf{y}) - \ell(\hat{\boldsymbol{\mu}})\big]. \tag{3.1}$$

Here, it is tacitly assumed that the same ϕ is used in $\ell(\mathbf{y})$ and $\ell(\hat{\boldsymbol{\mu}})$. Let h denote the inverse of b' in the relation $\mu_i = b'(\theta_i)$ of Lemma 2.1, so that $\theta_i = h(\mu_i)$. Then (2.1) gives the following expression for D^*,

$$D^* = \frac{2}{\phi}\sum_i w_i \left(y_i h(y_i) - b(h(y_i)) - y_i h(\hat{\mu}_i) + b(h(\hat{\mu}_i))\right). \tag{3.2}$$

qui donne justement de montrer comme exercice que la deviance est la somme des carrés des erreurs - ce qui correspond bien à ce que j'évoquais :

As a first example, suppose we have a GLM with normal distribution and all $w_i = 1$. We leave it as Exercise 3.1 to show that then

$$D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = \sum_i (y_i - \hat{\mu}_i)^2. \tag{3.3}$$

Thus the deviance equals the sum of squares used to measure goodness-of-fit in the analysis of variance (ANOVA).

Maintenant, un point plus problématique. Un élève m'a fait parvenir une copie d'écran du Exam MAS-I Study Manual :

Deviance

$$D = 2(l(\mathbf{b}_{max}) - l(\mathbf{b})) \sim \chi^2(n-p) \quad (46.1)$$

Binomial

$$D = 2\sum_{i=1}^{k}\left(y_i \ln\frac{y_i}{\hat{y}_i} + (n_i - y_i)\ln\frac{n_i - y_i}{n_i - \hat{y}_i}\right) \quad (46.2)$$

Normal (scaled deviance)

$$\sigma^2 D = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Poisson

$$D = 2\left(\sum_{i=1}^{n}\left(y_i \ln\frac{y_i}{\hat{y}_i} - (y_i - \hat{y}_i)\right)\right) \quad (46.5)$$

Gamma

$$D = 2a\left(\sum_{i=1}^{n}\left(-\ln\frac{y_i}{\hat{y}_i} + \frac{y_i - \hat{y}_i}{\hat{y}_i}\right)\right)$$

Standard error of regression (normal model)

$$s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-p}} \quad (46.4)$$

Likelihood ratio test

$$2(\hat{l} - \tilde{l}) = \tilde{D} - \hat{D} \sim \chi^2(q)$$

where hat indicates the unconstrained model, tilde indicates the constrained model, and there are q constraints.

Wald test

$$\frac{(b_j - r)^2}{\mathrm{Var}(b_j)} \sim \chi^2(1) \quad (46.7)$$

Score test

$$\mathbf{U}^T \mathfrak{I}^{-1} \mathbf{U} \sim \chi^2(p)$$

Type I/Type III tests Type I tests are sequential, adding one variable (or a group of variables) at a time to the model in a prescribed order. Type III tests check one variable (or a group of variables) assuming all other variables are in the model. Wald tests are Type III.

AIC

$$\mathrm{AIC} = -2l + 2p \quad (46.8)$$

BIC

$$\mathrm{BIC} = -2l + p \ln n \quad (46.9)$$

Je ne sais pas à quel point ce document est officiel, mais il est faux. Et j'en suis désolé. Donc pour résumer :

1) la bonne réponse est bien la réponse C, je persiste, et signe

2) lors de vos révisions, si vous voyez deux résultats contradictoires, merci de m'en faire part. Un élève (seulement) m'avait fait remarquer que ma définition n'était pas celle de de Jong & Heller (2008). Je lui avais dit que le livre n'était pas correct. Là encore, je persiste (même si la note en bas de tableau note que le terme plus approprié est la scaled deviance)

3) c'est le cours que je donne au tableau qui prévaut. C'est pour cela que j'insiste au premier cours sur le fait que la présence en cours est importante.

4) exceptionnelement, je veux bien accorder 1 point pour la réponse A car un des livres de références du cours (de Jong & Heller (2008)) contient une erreur, et que je ne l'avais pas mentionné en cours.

3 On estime trois modèles de Poisson, et on obtient les informations suivantes

modèle	variable(s)	degrés de liberté	log vraisemblance	AIC	BIC
(1)	classe de risque	5	-47,704	95,418	95,473.611
(2)	classe de risque + région	***	-47,495	***	***
(3)	classe de risque + région + indicatrice	10	-47,365	94,750	***

On sait de plus que la classe de risque prend les modalités $\{A, B, C, D, E\}$ et que la variable indicatrice prend deux modalités $\{0, 1\}$. Tous les modèles ont été estimés sur le même jeu de données.

Quelle est la différence (en valeur absolue) entre le AIC et le BIC dans le modèle (2) ?

A) moins de 85

B) entre 85 et 95

C) entre 95 et 105

D) entre 105 et 115

E) plus de 115

Il s'agissait d'un exerice de l'examen S de la CAS, du printemps 2016. Le modèle (2) a un degré de liberté de moins que le modèle 3 (correspondant à la variable indicatrice), aussi le nombre de degrés de liberté est 10-1=9. Aussi, le AIC pour le modèle (2) est ici

$$-2 \times (-47,495) + 2 \times 9 = 95,008,$$

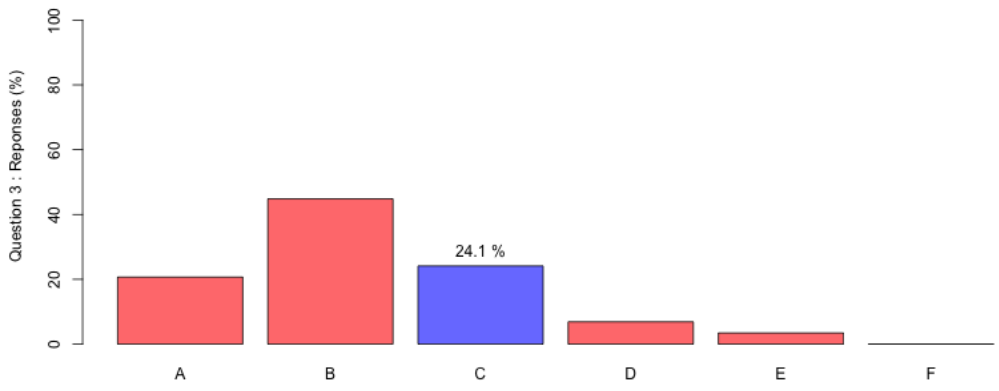
alors que le BIC est ici (toujours pour le modèle (2))

$$-2 \times (-47,495) + 9 \times \log n = 95,108.1$$

Aussi, la différence entre les deux est (en valeur absolue)

$$95,108.1 - 95,008 = 100.1$$

qui correspond à la réponse C.



4 On fait un sondage auprès de 100 personnes pour savoir pour qui ils voteront lors d'une prochaine élection parmi trois candidats $\{A, B, C\}$. On tient compte ici de l'âge des personnes interrogées à l'aide d'une classe d'âge, $[18, 30]$, $[31, 45]$, $[46, 61]$ et $[61, +]$. On utilise une régression logistique. On observe les résultats suivants : (i) le groupe $[18, 30]$, A et le candidat A sont considérées comme modalités de référence (ii) Pour le groupe $[18, 30]$, le logarithme de la côte de B et C sont respectivement -0.535 et -1.489. Pour une personne du groupe $[18, 30]$, quelle est la probabilité de préférer B ?

A) moins de 40%

B) entre 40% et 50%

C) entre 50% et 60%

D) entre 60% et 80%

E) plus de 80%

Il s'agissait d'un exercice de l'examen MAS-I de la CAS, de l'automne 2018. Ici, la modalité de référence est C (pour y). On nous dit (comme on se positionne par rapport à la référence) que

$$\log \frac{\mathbb{P}\big[Y = B | x = [18, 30]\big]}{\mathbb{P}\big[Y = A | x = [18, 30]\big]} = -0.535$$

et

$$\log \frac{\mathbb{P}\big[Y = C | x = [18, 30]\big]}{\mathbb{P}\big[Y = A | x = [18, 30]\big]} = -1.489$$

soit

$$\mathbb{P}\big[Y = B | x = [18, 30]\big] = e^{-0.535} \cdot \mathbb{P}\big[Y = A | x = [18, 30]\big]$$

et

$$\mathbb{P}\big[Y = C | x = [18, 30]\big] = e^{-1.489} \cdot \mathbb{P}\big[Y = A | x = [18, 30]\big]$$

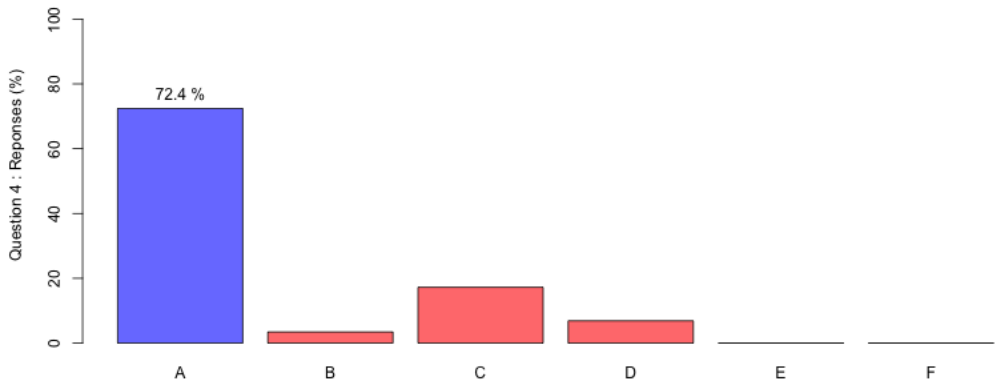
Or comme la somme des trois probabilités vaut 1 (par définition), on a

$$\big(1 + e^{-0.535} + e^{-1.489}\big) \cdot \mathbb{P}\big[Y = A | x = [18, 30]\big] = 1$$

aussi $\mathbb{P}\big[Y = A | x = [18, 30]\big] = 0.5521$ et donc, par substitution,

$$\mathbb{P}\big[Y = B | x = [18, 30]\big] = e^{-0.535} \cdot \mathbb{P}\big[Y = A | x = [18, 30]\big] = 0.3233$$

ce qui correspond à la réponse A.



5 En corrigeant des copies d'étudiants, on peut lire les affirmations suivantes à propos de la déviance des GLM

(1) une petite déviance indique un mauvais ajustement du modèle

(2) la déviance peut être utilisée pour comparer la qualité de l'ajustement pour des modèles imbriqués

(3) un modèle saturé a une déviance nulle

A) aucune affirmation n'est juste

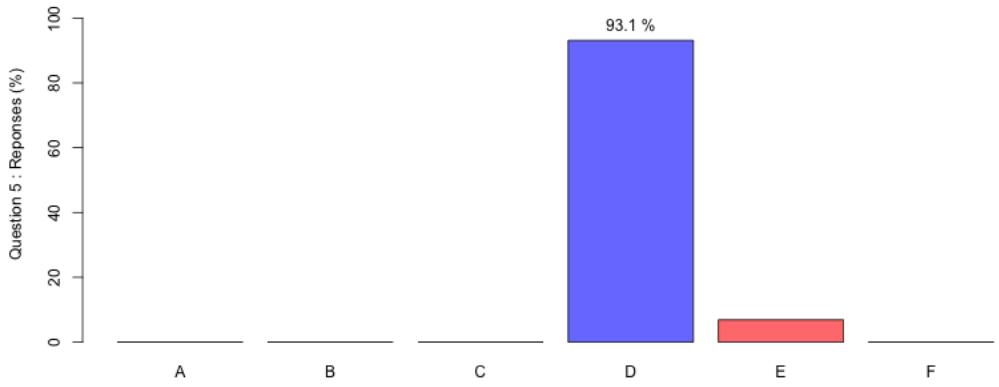
B) (1) et (2) sont justes

C) (1) et (3) sont justes

D) (2) et (3) sont justes

E) ni A, ni B, ni C et ni D

Il s'agissait d'une question de l'examen MAS-I de la CAS du printemps dernier. On va regarder les affirmations les unes après les autres. 1) La déviance est la distance du modèle au modèle saturé (ou parfait, quand $\widehat{\mu}_i = y_i$, $\forall i$). Donc au contraire, une petite distance signifie que le modèle est bon ! La première affirmation est donc fausse... 2) Oui, la déviance peut être utilisée pour juger la qualité de l'ajustement avec modèle GLM. Le soucis est qu'on ne veut pas seulement avoir un bon modèle (qualité de l'ajustemnet) mais aussi un modèle simple (ou disont parcimonieux) d'où l'utilisation des critères AIC ou BIC.... mais pour juger la qualité, on utilise la déviance ! 3) Oui, c'est la définition... Autrement dit, seules les affirmations 2 et 3 sont valides, ce qui correspond à la réponse D.



6 On modélise ici une variable y prenant trois modalités, $\{A, B, C\}$, et on suppose que C est la modalité de référence.

```
Coefficients (category A)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.591
genderM       -0.388
age2           1.128
age3           1.588

Coefficients (category B)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.039
genderM       -0.813
age2           1.478
age3           2.917
```

où la variable x_1 (`gender`) prend ici deux modalités $\{F, M\}$ et x_2 (`age`) prend ici trois modalités $\{1, 2, 3\}$. Quelle est la probabilité $\mathbb{P}[Y = A | x_1 = H, x_2 = 3]$?

A) moins de 5%

B) entre 5% et 10%

C) entre 10% et 15%

D) entre 15% et 20%

E) plus de 20%

Si on regarde les deux prévisions, on a

$$\boldsymbol{x}^\top \widehat{\boldsymbol{\beta}}_A = -0.591 - 0.388 + 1.588 = 0.609$$

et

$$\boldsymbol{x}^\top \widehat{\boldsymbol{\beta}}_B = -1.039 - 0.813 + 2.917 = 1.065$$

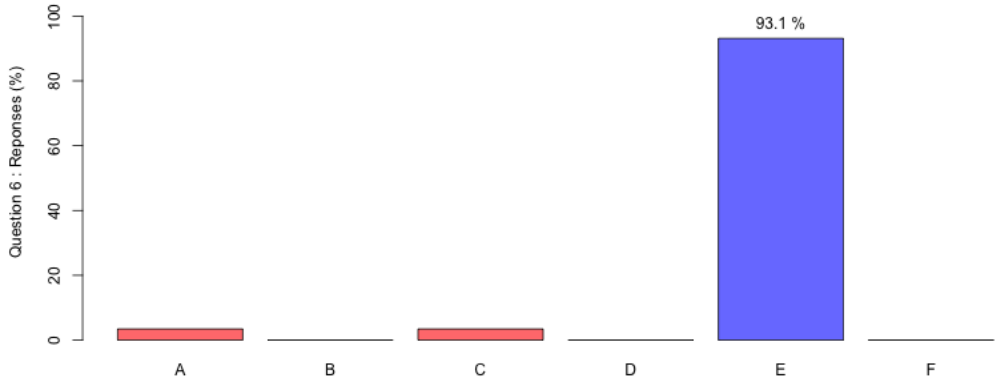
(la catégorie C étant ici la référence). Aussi, la probabilité d'être dans le groupe A est

$$\mathbb{P}[Y = A | x_1 = H, x_2 = 3] = \frac{e^{\boldsymbol{x}^\top \widehat{\boldsymbol{\beta}}_A}}{1 + e^{\boldsymbol{x}^\top \widehat{\boldsymbol{\beta}}_A} + e^{\boldsymbol{x}^\top \widehat{\boldsymbol{\beta}}_B}}$$

soit

$$\mathbb{P}[Y = A | x_1 = H, x_2 = 3] = \frac{e^{0.609}}{1 + e^{0.609} + e^{1.065}} = 0.3203$$

ce qui correspond à la réponse E.



7 On suppose que la loi Y est dans la famille exponentielle avec $b(\theta) = e^\theta$, $\phi = 1$ et que $\mu = \mathbb{E}[Y]$. Déterminez la variance de Y en fonction de μ ?

A) μ

B) μ^2

C) $1/\mu$

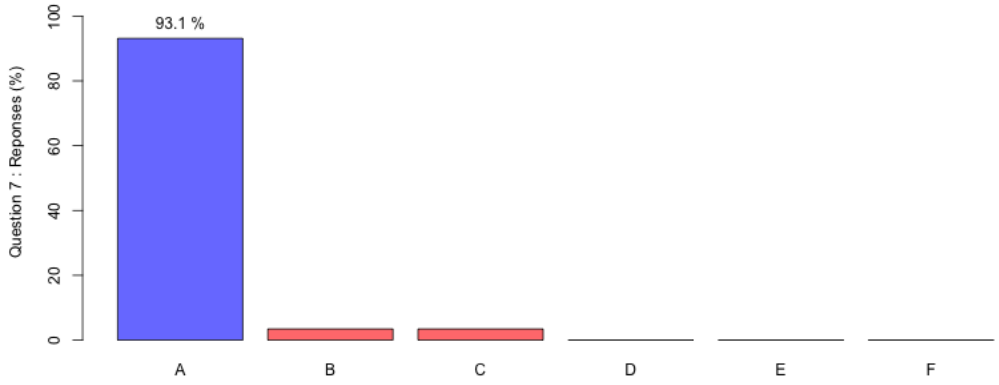
D) 1

E) e^{μ}

pour information, il s'agit d'une sample question de l'examen CAS-S (ou d'une question de cours, au choix). Comme $b(\theta) = e^{\theta}$, $b'(\theta) = e^{\theta}$ et $b''\theta) = e^{\theta}$. Or on sait que

$$\mu = \mathbb{E}[Y] = b'(\theta) = e^{\theta} \text{ et } \mathrm{Var}[Y] = \phi \cdot b''(\theta) = e^{\theta} = \mu$$

qui est la réponse A (on reconnaît au passage la loi de Poisson).



8 Sur un même jeu de données, avec y (strictement) positive et x, deux modèles sont estimés : le modèle (1)

```
Call:
lm(formula = log(y) ~ x)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.3055     0.2426   5.381 3.59e-07 ***
x             0.3043     0.4163   0.731    0.466
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.256 on 123 degrees of freedom
Multiple R-squared:  0.004325,        Adjusted R-squared:  -0.00377
F-statistic: 0.5343 on 1 and 123 DF,  p-value: 0.4662
```

et le modèle (2)

```
Call:
glm(formula = y ~ x, family = gaussian(link = "log"))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.2665     0.4266   5.313 4.89e-07 ***
x             0.1974     0.7035   0.266     0.79
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 509.611)
```

Pour une observation x, on veut comparer les prévisions avec les deux modèles. On note $\widehat{y}_{(j)}(x)$ la prévision - pour x - obtenue avec le modèle (j), correspond à l'estimateur (asymptotiquement) sans biais construit à l'aide de la méthode du maximum de vraisemblance. Que vaut $\widehat{y}_{(2)}(1) - \widehat{y}_{(1)}(1)$?

A) -6

B) -2.25

C) 0

D) 0.75

E) 6.75

Le premier modèle est un modèle log-normal, puisque $\log Y|X = x \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2)$. Or l'espérance d'une loi lognormale $LN(\mu, \sigma^2)$ est $\exp\left(\mu + \frac{\sigma^2}{2}\right)$, donc pour avoir un estimateur sans biais, on doit considérer

$$\exp\left(\widehat{\beta}_0 + \widehat{\beta}_1 x + \frac{\widehat{\sigma}^2}{2}\right) = \exp(1.3055 + 0.3043 + 1.256^2/2) = 11.007$$

Pour le second modèle, c'est simplement un GLM, donc on considère

$$\exp\left(\widehat{\beta}_0 + \widehat{\beta}_1 x\right) = \exp(2.2665 + .1974) = 11.7505$$

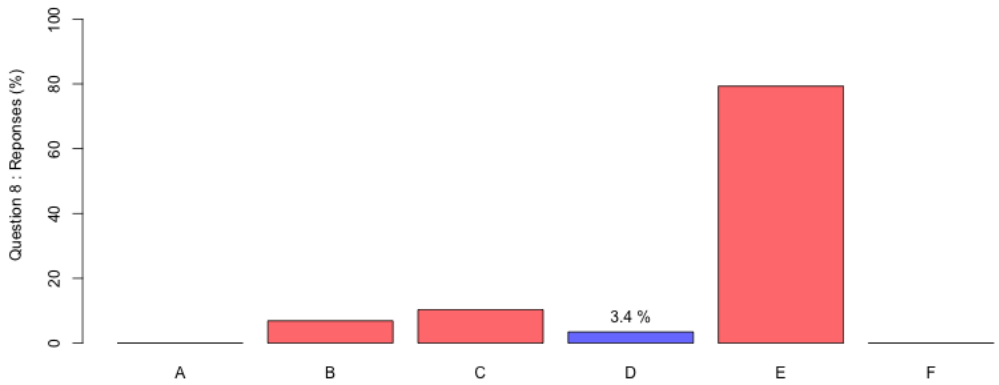
La différence est de 0.75, ce qui correspond à la réponse D, et non pas E, qui serait obtenu avec un estimateur biaisé pour le premier modèle, comme vu en cours, puisque

$$\exp[\mathbb{E}[\log(Y)]] \neq \mathbb{E}(Y)$$

En réalité, par l'inégalité de Jensen (le logarithme étant une fonction concave)

$$\mathbb{E}[\log(Y) < \log(\mathbb{E}[(Y)] \text{ donc } \exp[\mathbb{E}[\log(Y)]] < \mathbb{E}(Y)$$

autrement dit, en prenant l'exponentiel de la prévision, on sous-estime toujours : on a en effet un biais multiplicatif de $\exp(\sigma^2/2)$.



9 On dispose de la sortie de régression suivante, d'une régression sur deux variables catégorielles

```
Call:
glm(formula = loss ~ region + group, family = Gamma(link = "log"))

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)      5.26
regionB          0.18
regionC          0.37
group2           0.12
group3           0.25
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 1.44)
```

Calculer la prime pure pour un assuré du groupe 2 dans la région B,

A) moins de 200

B) entre 200 et 225

C) entre 225 et 250

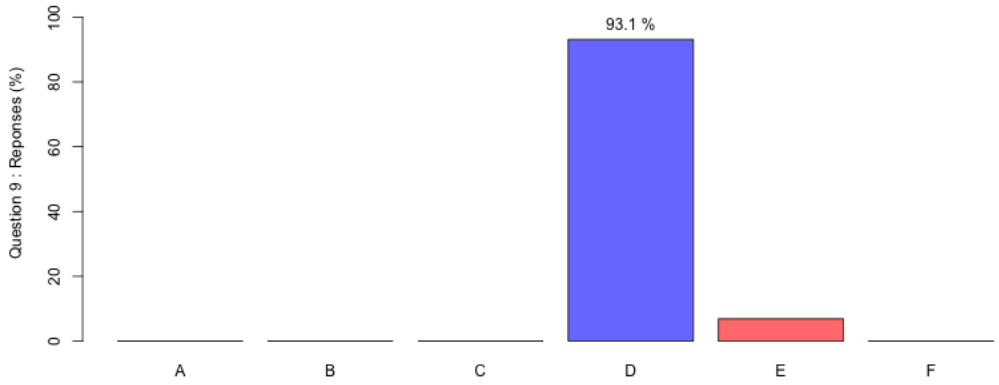
D) entre 250 et 275

E) plus de 275

Il s'agissait d'une question de l'examen S de la CAS de l'automne 2017. **il y avait toutefois une typo, corrigée au cours de l'examen : il s'agit du 'risque' (groupe) B et du 'territoire' (région) '2'** On nous dit qu'on a un lien logarithmique, aussi

$$\log \widehat{\mu} = \boldsymbol{x}^{\mathsf{T}}\boldsymbol{\beta} = 5.26 + \underbrace{0.18}_{\text{risk B}} + \underbrace{0.12}_{\text{territory 2}} = 5.56$$

donc $\widehat{\mu} = \exp[5.56] \sim 259.822$ ce qui correspond à la réponse D. Et comme vu en cours, la loi ne sert à rien pour la prévision.



10 On cherche à modéliser le nombre d'accident, par année, par police d'assurance. On suppose que y suit une loi de Poisson et on utilise un lien logarithmique. Deux variables explicatives sont prises en compte : le nombre de jeunes conducteurs, de moins de 25 ans ($x_1 \in \{0, 1 \text{ ou plus}\}$) et le nombre de conducteurs de plus de 25 ans ($x_2 \in \{1, 2 \text{ ou plus}\}$). On obtient la sortie de régression suivante

```
Call:
glm(formula = y ~ youth + old, family = poisson(link = "log"))

Coefficients
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -2.663
youth1+        0.132
old2+         -0.031
```

On considère une police avec un jeune conducteur de moins de 25 ans, et un autre de plus de 25 ans. Quelle est la probabilité d'avoir deux sinistres ou plus ?

A) moins de 0.1%

B) entre 0.1% et 0.5%

C) entre 0.5% et 1%

D) entre 1% et 5%

E) plus de 5%

Il s'agissait d'un exercice de l'examen S de la CAS de l'automne 2015. Ici x_1 vaut ' ou plus' et x_2 vaut '1' - qui est la modalitée de référence. Donc le paramètre de la loi de Poisson (conditionnelle aux variables explicatives) est

$$\widehat{\lambda} = \exp[\widehat{\beta}_0 + \widehat{\beta}_1 \cdot 1 + \widehat{\beta}_2 \cdot 0] = \exp[-2.663 + 0.132] = 0.07958$$

On nous demande ici

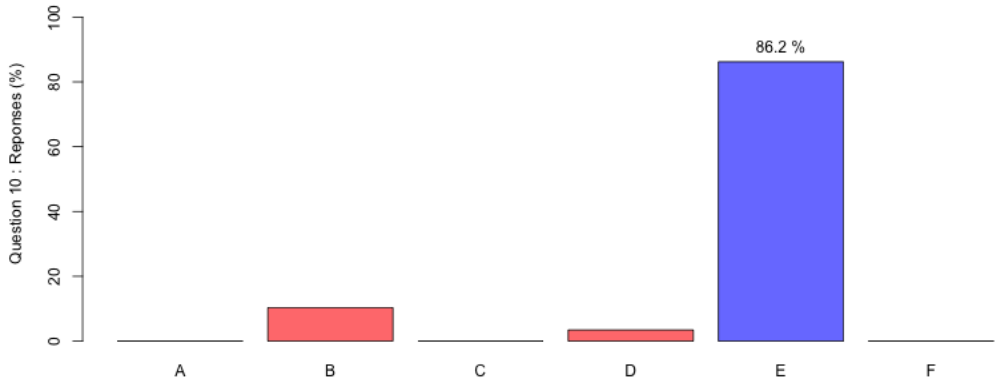
$$\mathbb{P}[Y \geqslant 2|\boldsymbol{x}] = 1 - \big(\underbrace{\mathbb{P}[Y = 0|\boldsymbol{x}]}_{=e^{-\lambda}} + \underbrace{\mathbb{P}[Y = 1|\boldsymbol{x}]}_{=e^{-\lambda}\lambda^1}\big)$$

soit, en remplaçant

$$\mathbb{P}[Y \geqslant 2|\boldsymbol{x}] = 1 - \big(e^{-0.07958} + 0.07958 \times e^{-0.07958}\big) = 1 - 0.9969966 = 0.0030034 = 0.3\%.$$

ce qui correspond à la réponse B. Je ne comprends pas pourquoi autant d'élèves ont répondu E.

Malheureusement, il semble que dans l'énoncé distributé, $\widehat{\beta}_0$ valait -0.663 ce qui change le résulat, et donne la réponse E. Pour la correction (et l'attribution des points) c'est la réponse E qui donnera 1 point.



11 On nous donne la sortie de régression suivante, en fonction de deux variables continues x_1 et x_2

```
Call:
glm(formula = y ~ x1 + x2, family = binomial(link = "logit"))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    4.542      3.254   1.396    0.163
x1            -5.070      3.246  -1.562    0.118
x2            -4.539      4.224  -1.074    0.283

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 13.4602  on 9  degrees of freedom
Residual deviance:  8.4035  on 7  degrees of freedom
AIC: 14.403

Number of Fisher Scoring iterations: 5
```

Le nuage des points $(x_{1,i}, x_{2,i})$ est représenté sur la Figure page 2 (en haut), les points étant noirs ($\bullet$) si $y_i = 1$ et blancs ($\circ$) si $y_i = 0$. Représentez (en la hachurant) la région pour laquelle

$$\mathbb{P}[Y = 1|\boldsymbol{x} = (x_1, x_2)] > \mathbb{P}[Y = 0|\boldsymbol{x} = (x_1, x_2)]$$

.

On va commencer par tracer la frontière, autrement dit, trouver l'ensemble des points (x_1, x_2) tels que

$$\mathbb{P}[Y = 1|\boldsymbol{x} = (x_1, x_2)] > \mathbb{P}[Y = 0|\boldsymbol{x} = (x_1, x_2)]$$

On sait (c'est le principe des modèles linéires qu'il s'agit d'une droite (dans le plan (x_1, x_2)). Si les deux probabilités sont égales, c'est que $\mathbb{P}[Y = 1|\boldsymbol{x} = (x_1, x_2)] = 1/2$, soit, puisqu'on a un lien logistique

$$\frac{\exp[\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2]}{1 + \exp[\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2]} = \frac{1}{2} = \frac{1}{1+1}$$

soit $\exp[\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2] = 1$ ou encore $\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 = 0$. Autrement dit (c'était la première étape), on doit voir sur le dessin la droite d'équation

$$\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 = 4.542 - 5.070 x_1 - 4.539 x_2 = 0.$$

Comme on est dans le plan (x_1, x_2), le plus simple est de tracer la droite soit la forme

$$x_2 = \frac{4.542}{4.539} - \frac{5.0539}{4.539} x_1 = 1 - 1.12 x_1.$$

La droite va donc passer par les points $P_1 = (0, 1)$ et $P2 = (1, -0.12)$, comme sur la figure ci-dessous.

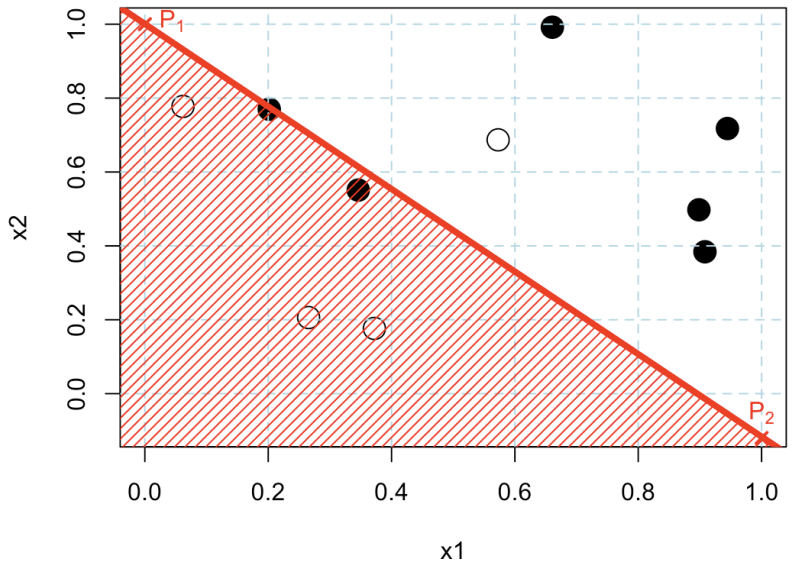
Pour savoir quelle région correspond à celle demandée, on peut prendre un point au hasard, d'un côté, ou de l'autre. Pour simplifier les calculs, prenons $(x_1, x_2) = (0, 0)$. Dans ce cas, on a

$$\mathbb{P}[Y = 1|\boldsymbol{x} = (0, 0)] = \frac{\exp[\widehat{\beta}_0]}{1 + \exp[\widehat{\beta}_0]} = \frac{\exp[4.542]}{1 + \exp[4.542]} \sim 98.9\%$$

(ou plus simplement, comme $\widehat{\beta}_0 > 0$, on sait que la valeur prédite en $\mathbf{0}$ excède 50%). Donc le point $\mathbf{0}$ est bien dans la région

$$\mathbb{P}[Y = 1|\boldsymbol{x} = (x_1, x_2)] > \mathbb{P}[Y = 0|\boldsymbol{x} = (x_1, x_2)]$$

et on va alors hachurer cette partie là...



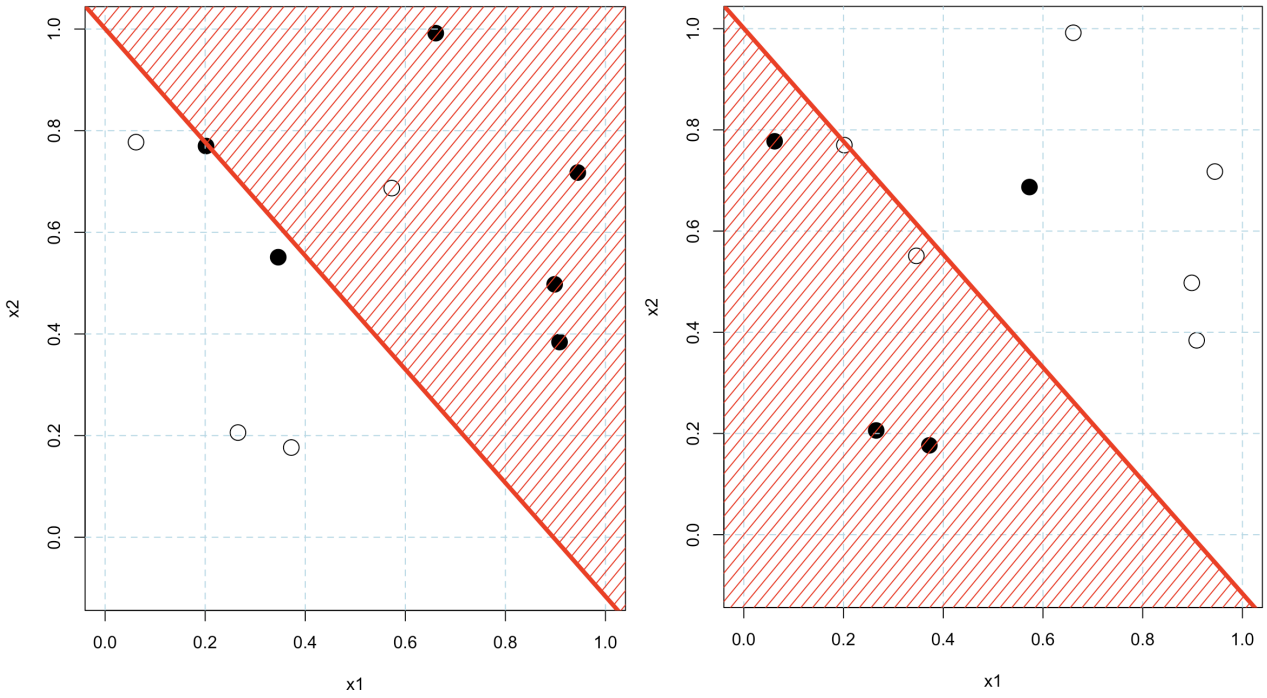
En fait, on peut faire un peu plus simple, puisque pour le point (0,0)

$$\log \frac{\mathbb{P}[Y=1|x_1=0,x_2=0]}{\mathbb{P}[Y=0|x_1=0,x_2=0]} = \beta_0 = 4.542$$

donc $\mathbb{P}[Y=1|x_1=0,x_2=0] > \mathbb{P}[Y=0|x_1=0,x_2=0]$ ce qui est bien ce qu'on cherche ! Donc $(0,0)$ est dans la région que l'on cherche !Autrement dit, on va hachurer la région en bas à gauche. On peut vérifier avec un autre point si on n'est pas convaincu !

$$\log \frac{\mathbb{P}[Y=1|x_1=1,x_2=1]}{\mathbb{P}[Y=0|x_1=1,x_2=1]} = \beta_0+\beta_1+\beta_2 = 4.542-5.070-4.539 = -5.067$$

donc $\mathbb{P}[Y=1|x_1=0,x_2=0] < \mathbb{P}[Y=0|x_1=0,x_2=0]$ ce qui n'est pas ce qu'on nous demande, donc $(1,1)$ n'est pas dans la région que l'on cherche ! Pour le sujet B, la sortie de la régression était très légèrement différente. En fait, si on comparer les figures, on avait les mêmes observations $\boldsymbol{x}_i$ mais les y_i ont été permutés. Aussi, pour le sujet A, on avait la sortie de gauche, et pour le sujet B, la sortie de droite



L'énoncé suivant est relatif aux questions 12 et 13

On cherche à modéliser la survenance, ou non, d'un accident pour une police d'assurance. On utilise deux variables explicatives : le genre (male / female) et la classe d'âge (1 / 2 / 3). La classe d'âge est prise comme une variable numérique, et on considère une régression quadratique sur cette variable. Pour une police d'assurance i, la survenance, ou non, d'un sinistre est une variable Y_i telle que

$$g(\mathbb{E}[Y_i]) = \beta_0 + \beta_1 \mathbf{1}(x_{1,i} = \text{male}) + \beta_2 \text{age} + \beta_3 \text{age}^2,$$

où g est la fonction de lien logistique. On a les statistiques suivantes, qui donne le nombre de polices dans chacun des cas

	genre					
	male			female		
	age			*age*		
	1	2	3	1	2	3
pas de sinistre	20	28	30	24	28	22
un sinistre	8	7	3	16	13	1

On dispose de la sortie suivante

```
Coefficients
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.1155
genreM        -0.4192
age .          1.2167
age2          -0.5412
```

12 Calculer la variance du nombre de sinistres qui seront déclarés pour des assurés hommes du groupe 2 (on arrondira à la valeur la plus proche)

A) 0.2

B) 1.2

C) 2.4

D) 4.8

E) 6.0

Il s'agissait d'un exercice ACTEX de 2018. On a ici un loi binomiale, ou plus précisément, $Y|x_1 = \text{male}, \text{age} = 2 \sim \mathcal{B}(n,p)$ avec $n = 28 + 7 = 35$ d'après la sortie (ci-dessus) et p est la probabilité,

$$p = \mathbb{P}[Y = 1|x_1 = \text{male}, \text{age} = 2] = \frac{e^{\boldsymbol{x}^\top \hat{\boldsymbol{\beta}}}}{1 + e^{\boldsymbol{x}^\top \hat{\boldsymbol{\beta}}}} = \frac{1}{1 + e^{-\boldsymbol{x}^\top \hat{\boldsymbol{\beta}}}}$$

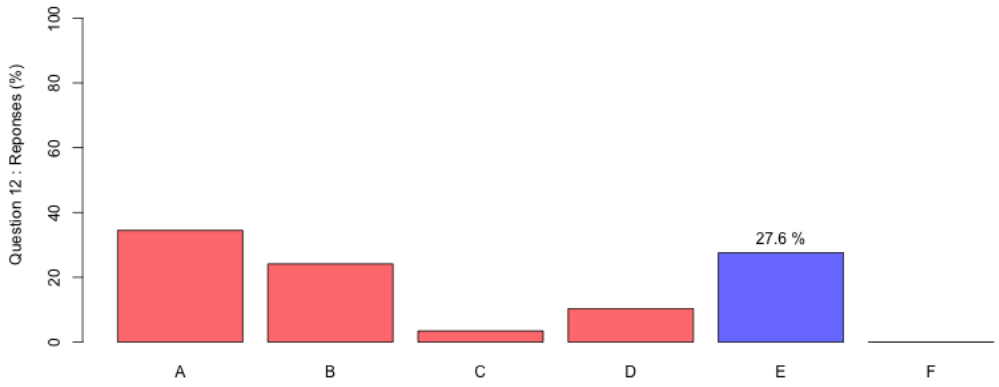
(si on veut simplifier un peu les calculs), aussi

$$p = (1 + \exp(-[-1.1155 - 0.4192 + 1.2167 \times 2 - 0.5412 \times 2^2]))^{-1} = 0.2199$$

Nous avons vu que le nombre d'accident suivait ici une loi binomiale, dont la variance est (cf cours de probabilités) $np(1-p)$ soit ici

$$\text{Var}[Y|x_1 = \text{male}, \text{age} = 2] = np(1-p) = 34 \times 0.2199 \times (1 - 0.2199) = 6.004$$

ce qui correspond à la réponse E.



13 Parmi les polices suivantes, laquelle (ou lesquelles) ont une probabilité d'avoir un accident qui excède 25% ?

(i) un assuré homme de la classe d'âge 1

(ii) un assuré homme de la classe d'âge 2

(iii) une assurée femme de la classe d'âge 3

A) (i) seulement

B) (ii) seulement

C) (i) et (ii)

D) (i) et (iii)

E) (ii) et (iii)

Il faut se lancer dans les calculs ! Rappelons que la probabilité pour un homme de la classe d'âge 2 est (cf question précédante)

$$p_2 = (1+\exp(-[-1.1155 - 0.4192 + 1.2167 \times 2 - 0.5412 \times 2^2]))^{-1} = 0.2199$$

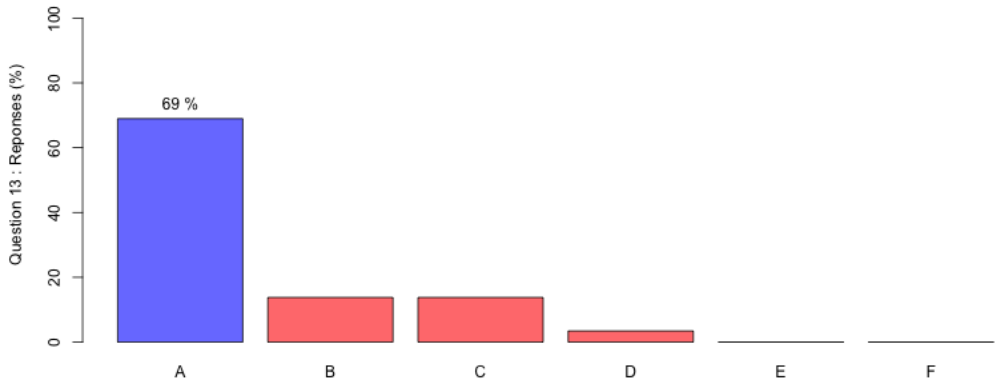
Pour avoir les deux autres probabilités, il faut remplacer 2 par 1 ou 3, ce qui donne

$$p_1 = (1+\exp(-[-1.1155 - 0.4192 + 1.2167 \times 1 - 0.5412 \times 1^2]))^{-1} = 0.2975$$

$$p_3 = (1+\exp(-[-1.1155 - 0.4192 + 1.2167 \times 3 - 0.5412 \times 3^2]))^{-1} = 0.0882$$

Aussi, seuls les hommes de la classe 1 ont plus de 25% (presque 30% en réalité) d'avoir un sinistre. Ce qui est la réponse A. Pour ceux qui voulait répondre en moins de 30 secondes avec une chance assez grande d'avoir bon, il était possible de prendre les probabilités empiriques (qui sont différentes des prévisions par GLM, mais si le modèle est bon, on ne doit pas être trop éloigné !)

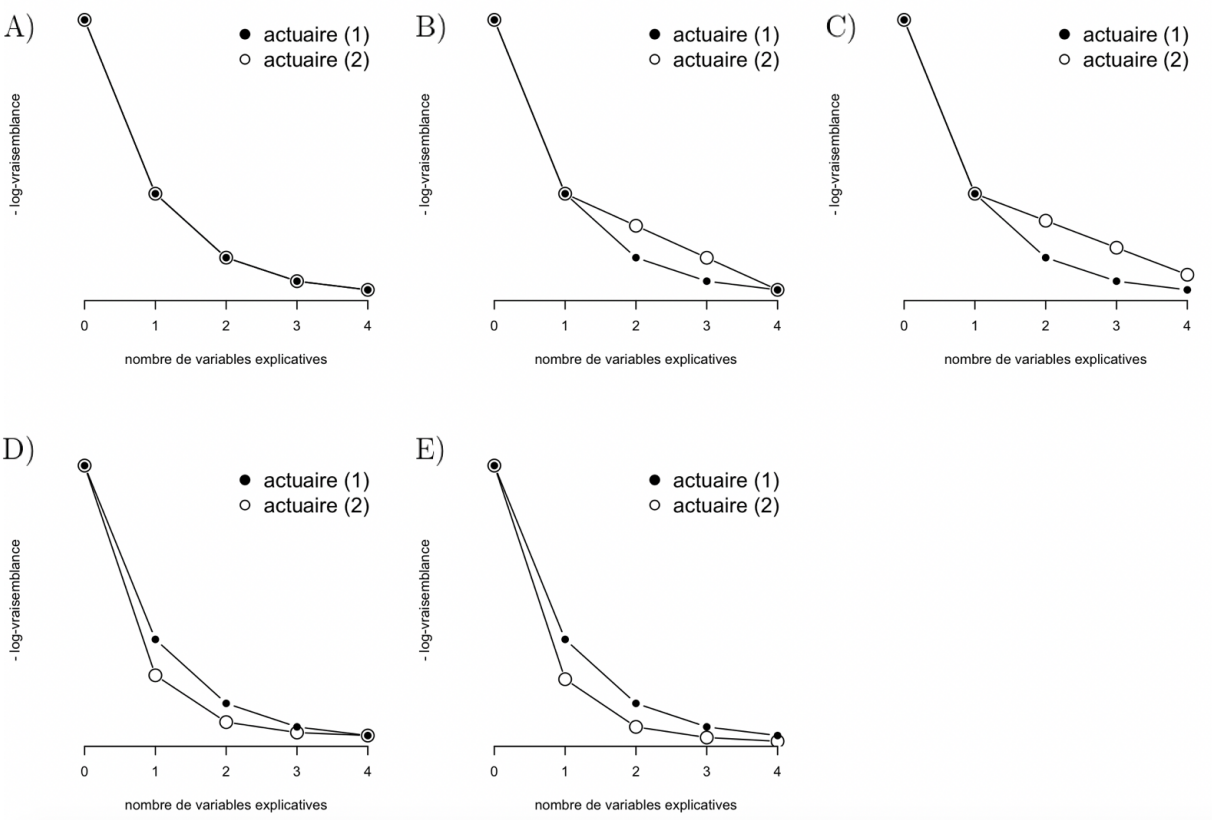
$$p_1 = \frac{8}{20+8} = 28.57\% > 25\% \text{ alors que } p_2 = \frac{7}{28+7} = 20.00\% \text{ et } p_3 = \frac{3}{30+3} = 9.09\% < 25\%.$$



14 On suppose que deux actuaries utilisent une régression GLM sur la même base, comptenant ~~5~~ 4 variables explicatives, avec le même type de modèle (même loi et même fonction de lien). On suppose

- que l'actuaire (1) choisi le modèle avec k variables explicatives en comparant tous les modèles avec k variables explicatives (et compare alors leur log-vraisemblance, à k fixé)
- que l'actuaire (2) utilise une approche itérative - stepwise forward - pour sélectionner un modèle avec k variables explicatives (et compare alors leur log-vraisemblance, à k fixé)
- en utilisant le même critère le AIC, les deux actuaires n'obtiennent par le même modèle.

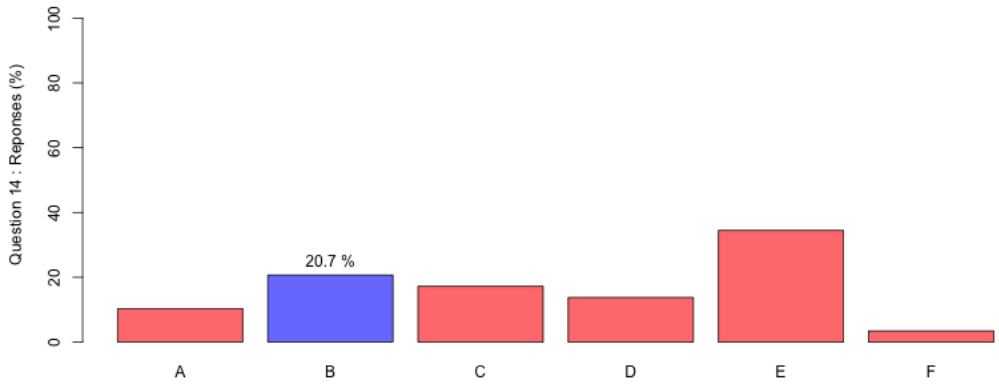
Les graphiques suivant montrent l'évolution de l'opposé de la log-vraisemblance du modèle retenu en fonction de k pour les deux actuaires. Lequelle vous semble le plus réaliste ?



Il s'agissait d'un exercice ACTEX de 2018. Avant de se lancer dans les calculs, on sait que l'approche best subset (où on regarde tous les modèles) et l'approche stepwise donnent exactement la même chose quand on a 0, 1 ou 4 variables explicatives. Ce qui élimine automatiquement C, D et E.

On notera d'ailleurs que les deux graphiques (D et E) sont aussi éliminés car l'approche best subset donne forcéement une log-vraisemblance plus grande, donc les points noirs (•) doivent être en dessous des points blancs (○), a priori, pour 2 ou 3 variables explicatives.

On veut maintenant départager entre A et B. Or on sait, selon la dernière hypothèse que le modèle retenu avec le AIC n'est pas le même, donc forcément, la bonne réponse est B. Le soucis c'est que j'avais fait une typo en indiquant qu'il y avait 5 variables explicates... alors qu'il y en avait 5. Je supprime la question...



15 On nous donne la densité suivante pour Y,

$$f(y) = \sqrt{\frac{\lambda}{2\pi y^3}} \exp\left[-\frac{\lambda(y-\mu)^2}{2\mu^2 y}\right] \text{ pour } y > 0$$

(et 0 sinon), pour des paramètres μ et λ positifs. On sait que la distribution de Y est dans la famille exponentielle. Donnez l'expression du paramètre canonique θ et la forme de $b(\theta)$ pour cette distribution.

A) $\theta = -1/\mu$ et $b(\theta) = -\log(-\theta)$

B) $\theta = -1/\mu$ et $b(\theta) = -\sqrt{-2\theta}$

C) $\theta = -1/2\mu^2$ et $b(\theta) = -\log(-\theta)$

D) $\theta = -1/2\mu^2$ et $b(\theta) = -\sqrt{-2\theta}$

E) $\theta = -1/2\mu^2$ et $b(\theta) = e^\theta$

Pour information, il s'agit de la loi inverse Gaussienne. Mais on connaît rarement les lois par coeur, donc reprenons tranquillement. Pour ça, on prend le logarithme de la densité

$$\log f(y) = \log\left(\sqrt{\frac{\lambda}{2\pi y^3}}\right) + \left[-\frac{\lambda(y-\mu)^2}{2\mu^2 y}\right]$$

Comme pour la loi normale, on va développer ici le carré $(y-\mu)^2$, soit

$$\log f(y) = \frac{1}{2}\log\left(\frac{\lambda}{2\pi y^3}\right) + \left[-\frac{\lambda(y^2+\mu^2-2y\mu)}{2\mu^2 y}\right]$$

On va simplifier un peu dans le terme de droite

$$\log f(y) = \frac{1}{2}\log\left(\frac{\lambda}{2\pi y^3}\right) - \frac{\lambda y}{2\mu^2} + \frac{\lambda}{\mu} - \frac{\lambda}{2y}.$$

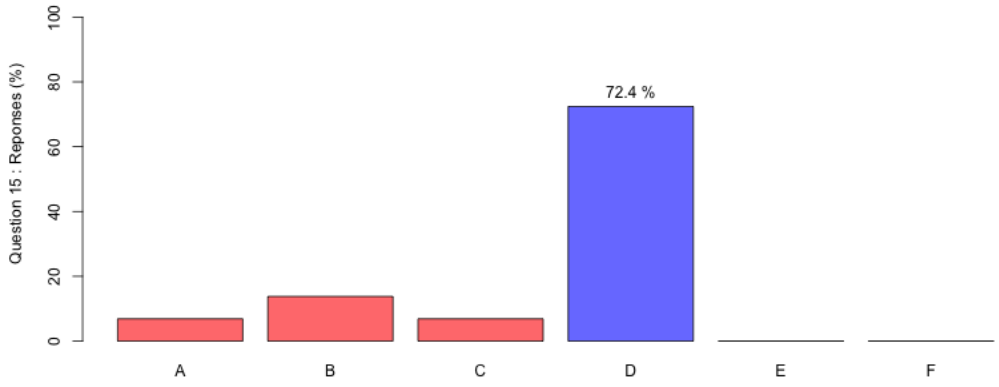
On y est presque : on peut alors écrire

$$\log f(y) = \frac{y\cdot(-1/(2\mu^2)) - (-1/\mu)}{1/\lambda} + \frac{1}{2}\log\left(\frac{\lambda}{2\pi y^3}\right) - \frac{\lambda}{2y}$$

Cette fois c'est bon ! Si on revient à l'expression de la première page, on doit avoir

$$f(y) = \exp\left(\frac{y \cdot \theta - b(\theta)}{\phi} + c(y, \phi)\right),$$

et il reste juste à identifier. Pour le paramètre canonique $\theta = (-1/(2\mu^2))$ et pour le second terme, $b(\theta) = (-1/\mu)$, qu'on réécrit tout simplement $b(\theta) = -\sqrt{-2\theta}$, ce qui correspond à la réponse D. Accessoirement, le paramètre de nuisance est $\phi = 1/\lambda$.



16 Quelles sont les affirmations justes parmi les suivantes, au sujet de la sélection de variable par une méthode itérative, forward stepwise, qui serait déroulée sans critère d'arrêt ?

(i) si p est le nombre de variables explicatives, il y a 2^{p-1} modèles à estimer

(ii) la log-vraisemblance du modèle optimal à k variables (obtenu par l'approche forward) est plus petite que la log-vraisemblance du modèle optimal à $k+1$ variables (obtenu par l'approche forward)

(iii) les variables retenues dans le modèle optimal à k variables (obtenu par l'approche forward) est un sous-ensemble des variables retenues dans le modèle optimal à $k+1$ variables (obtenu par l'approche forward)

A) (i) seulement

B) (ii) seulement

C) (iii) seulement

D) (i), (ii) et (iii)

E) aucune des propositions précédantes (ni A, ni B, ni C, ni D)

Il s'agissait d'un exercice ACTEX de 2018. L'affirmation (i) est fausse ! 2^{p-1} c'est le nombre de modèles à estimer en best subset, par en stepwise ! En forward stepwise on a

- 1 modèle à estimer, sans variable explicatives
- p modèles à estimer avec 1 seule variable explicative
- $p-1$ modèles à estimer avec 2 variables explicatives (on a celle de l'étape précédante, et on rajoute individuellement parmi les $p-1$ qui restent)
- $p-2$ modèles à estimer avec 3 variables explicatives (on a celles de l'étape précédante, et on rajoute individuellement parmi les $p-2$ qui restent)

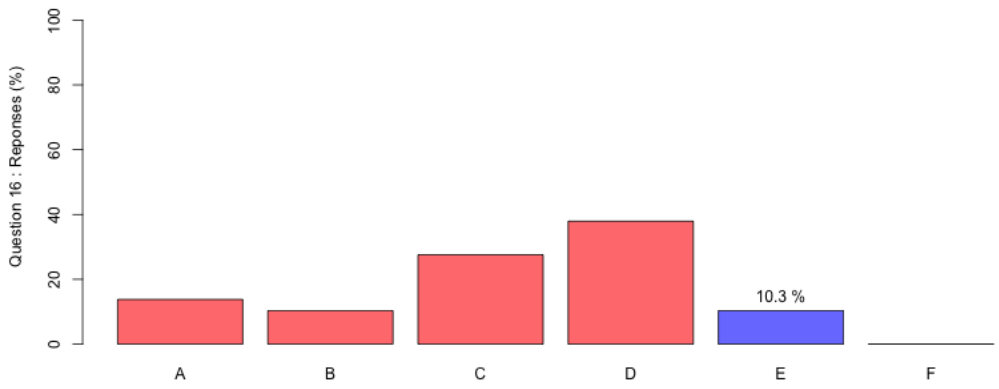
- ...
- 2 modèles à estimer avec $p-1$ variables explicatives (on a celles de l'étape précédante, et on rajoute individuellement parmi les 2 qui restent)
- 1 modèle à estimer avec toutes les variables explicatives

soit

$$1+\underbrace{p+(p-1)+(p-2)+\cdots+2+1}_{=\frac{p(p+1)}{2}}=1+\frac{p(p+1)}{2}$$

(sans critère d'arrêt).

L'affirmation (iii) est vraie ! Les modèles construits par l'approche forward sont imbriqués, par construction... Et l'affirmation (ii) est une conséquence de (iii) donc elle est également valide. Il fallait donc retenir E.



17 On dispose de $n=10$ observations (y_i,x_i), avec $y\in\mathbb{N}$. Une régression de Poisson donne la sortie suivante

```
Call:
glm(formula = y ~ x, family = poisson)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.9320     0.1313   22.33  < 2e-16 ***
x             1.0877     0.1866    5.83 5.56e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 43.5531  on 9  degrees of freedom
Residual deviance:  7.8088  on 8  degrees of freedom
AIC: 65.479
```

Le nuage des points $(x_{1,i}, x_{2,i})$ est représenté sur la figure page 2 (en bas). Représentez prévision $x \mapsto \mathbb{E}[Y|x]$ sur la Figure.

Comme on a une régression de Poisson, avec un lien logarithmique (je me répète probablement, mais c'est le lien canonique), donc l'estimateur de $\mathbb{E}[Y|x]$ est

$$\exp\left(\widehat{\beta}_0 + \widehat{\beta}_1 \cdot x\right)$$

Mais comme on est en échelle logarithmique sur l'axe des ordonnéees, on note que

$$\log\left(\mathbb{E}[Y|x]\right) = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot x$$

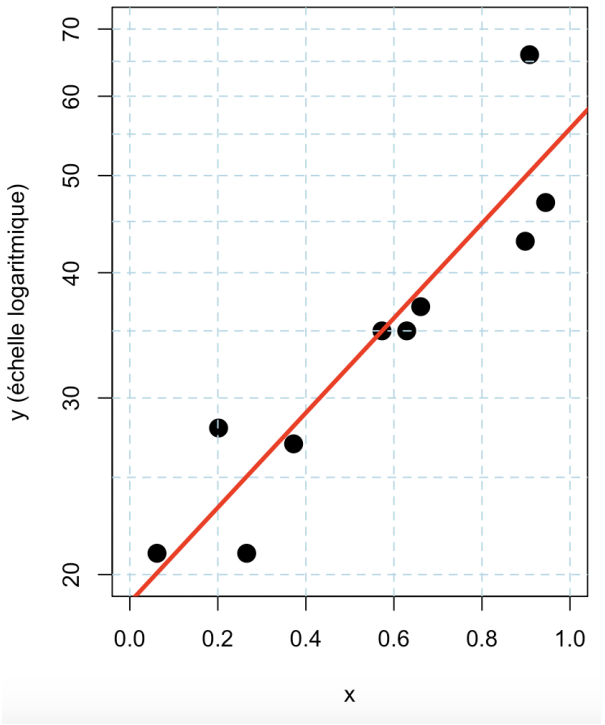
donc en fait, $x \mapsto \mathbb{E}[Y|x]$ est une droite (ce qui simplifie grandement le tracé ! Il suffit d'avoir deux points,

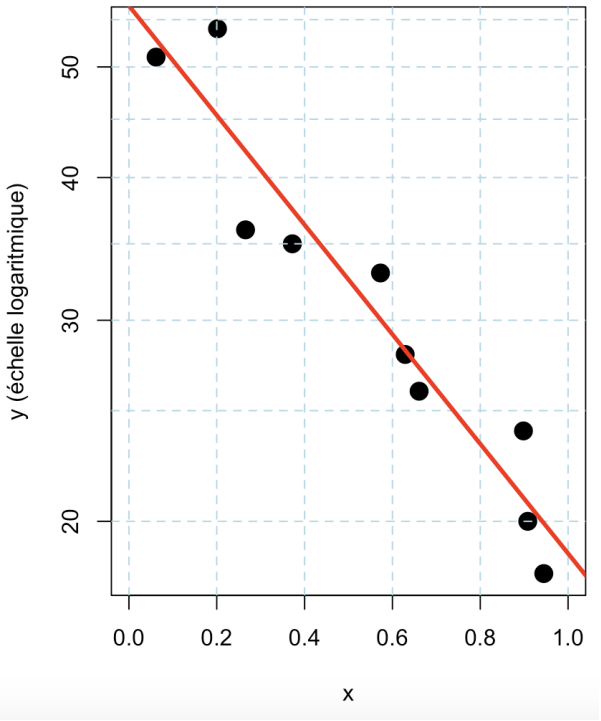
$$(0, \exp(\widehat{\beta}_0) = (0, e^{2.9320}) \sim (0, 18.76)$$

et

$$(1, \exp(\widehat{\beta}_0 + \widehat{\beta}_1) = (0, e^{2.9320+1.0877}) \sim (1, 55.68)$$

Pour l'autre sujet (le sujet B), on avait une autre sortie, mais la prévision était encore une droite quand on est en échelle logarithmique sur l'axe des ordonnées.





18 On considère un modèle GLM, avec une loi Gaussienne et une fonction de lien identité. Une des observations est $y_i = 805$ et le modèle prédit $\widehat{y}_i = 740$. On sait de plus que $\widehat{\sigma} = 44$. Quelle est la valeur du résidu de déviance $\widehat{\varepsilon}_i^D$?

A) 1.1

B) 1.3

C) 1.5

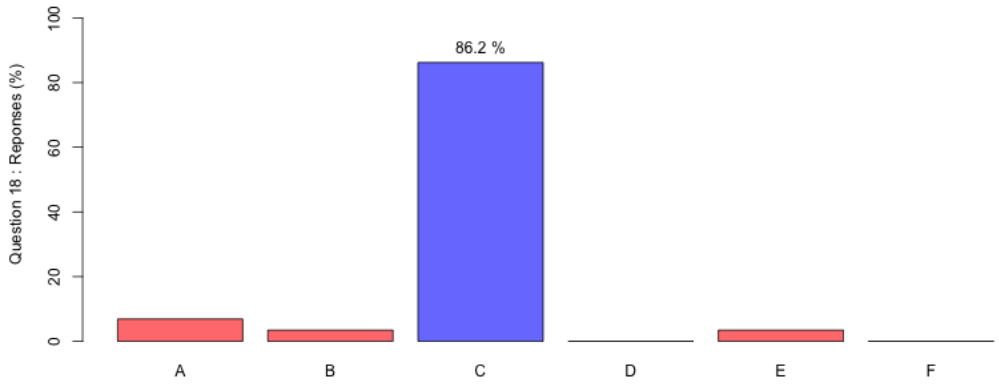
D) 1.7

E) 1.9

Il s'agit d'un exercice de l'examen MAS-I de la CAS de 2018. Pour rappel, les résidus de déviance sont égaux aux résidus de Pearson aux dans le cas du modèle Gaussien, qui correspondent aux résidus dit Studentisés, dans le modèle linéaire ! Aussi, ici,

$$\widehat{\varepsilon}_i^D = \frac{y_i - \widehat{y}_i}{\widehat{\sigma}} = \frac{805 - 740}{44} = 1.477$$

ce qui correspond à la réponse C (classiquement, on retient la valeur la plus proche).



19 On essaye de modéliser la probabilité qu'une personne ait la grippe un hiver (y est ici grippe) à l'aide de trois variables explicatives, l'âge `Age` qui est une une variable catégorielle prenant 2 modalités (`Less25` (entre 0 et 25 ans) et `More25` (plus de 25 ans)), une zone géographique (variable `Zone`) prenant trois modalités (A, B et C), et enfin `Income` qui est une une variable numérique (correspondant à un montant en '000 de dollars).

```
Call:
glm(grippe ~ Age + PriorRate + Zone, Income = binomial(link = "probit"))

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)         0.4270
AgeMore25           0.1320
ZoneB               0.0160
ZoneC              -0.0920
Income              0.0150
```

Estimer la probabilité d'avoir la grippe pour un assuré de moins de 25 ans, habitant dans la zone C et ayant un revenu de $ 22,500 ?

A) moins de 60%

B) entre 60% et 70%

C) entre 70% et 80%

D) entre 80% et 90%

E) au moins 90%

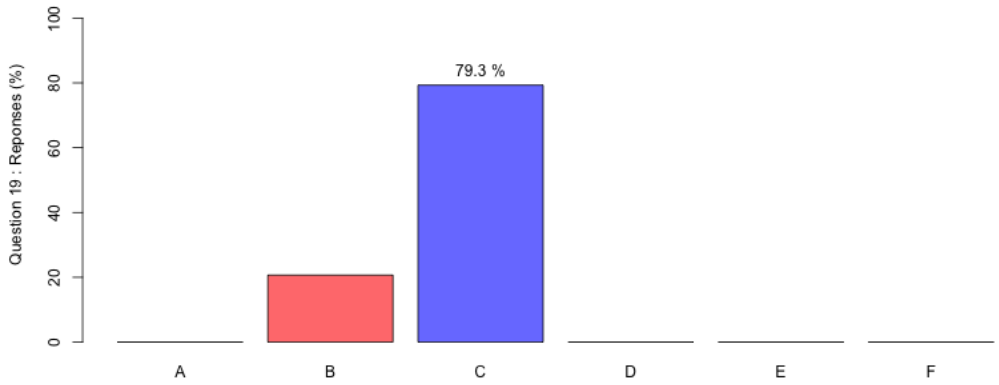
En utilisant la notation classique

$$g(\widehat{p}) = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot 0 + \widehat{\beta}_2 \cdot 0 + \widehat{\beta}_3 \cdot 1 + \widehat{\beta}_4 \cdot \frac{22,5000}{1000} = 0.4270 - 0.0920 + 0.0150 \cdot 225 = 0.6725$$

or on a une régression probit donc $g = \Phi^{-1}$, donc $\widehat{p} = \Phi(0.6725)$. Or page 1, si on reprend le tableau

	75%	90%	95%	97.5%	99%	99.5%	99.9%	99.95%
Loi normale	**0.674**	1.282	1.645	1.960	2.326	2.576	3.090	3.291

on a l'impression que $\Phi(0.6725) \sim 75\%$ - en réalité, la valeur exacte est 74.86% - ce qui correspond à la réponse C.



20 On suppose que Y suit une loi de la famille exponentielle, $\log f(y) = y \cdot \beta(\theta) + \gamma(\theta) + \delta(y)$, avec $\beta(\theta) = \log[\theta/(1-\theta)]$ et $\gamma(\theta) = 20\log(1-\theta)$. Que vaut la variance de Y ?

A) 20θ

B) $\theta(1-\theta)$

C) $20\theta(1-\theta)$

D) $20\theta(1+\theta)$

E) $20\theta^2$

Il s'agit d'un exercice de l'examen de pratique MAS-I de la CAS. Attention, ce ne sont pas les formes usuelles de la loi exponentielle. Ici $\beta(\theta)$ est le paramètre canonique et γ correspond à la fonction usuelle b. En fait, on reconnait la loi binomiale, et θ est le paramètre usuel de la loi de Bernoulli, i.e. la probabilité. Mais refaisons les calculs. Revenons à la base: on nous dit que $\log f(y) = y \cdot \beta(\theta) + \gamma(\theta) + \delta(y)$, donc

$$\frac{d}{d\theta} f(y) = \big[y\beta'(\theta) + \gamma'(\theta)\big] f(y)$$

donc

$$\int \frac{d}{d\theta} f(y) dy = \int \big[y\beta'(\theta) + \gamma'(\theta)\big] f(y) dy$$

soit

$$0 = \beta'(\theta)\int yf(y)dy + \gamma'(\theta)\int f(y)dy = \beta'(\theta)\cdot\mathbb{E}[Y] + \gamma'(\theta)$$

soit

$$\mathbb{E}[Y] = -\frac{\gamma'(\theta)}{\beta'(\theta)}$$

Allons un cran plus loin,

$$\frac{d^2}{d\theta^2}f(y) = \big[y\beta''(\theta) + \gamma''(\theta)\big]f(y) + \big[y\beta'(\theta) + \gamma'(\theta)\big]^2 f(y)$$

i.e.

$$\int \frac{d^2}{d\theta^2}f(y)dy = \int \big[y\beta''(\theta) + \gamma''(\theta)\big]f(y)dy + \int \big[y\beta'(\theta) + \gamma'(\theta)\big]^2 f(y)dy$$

soit

$$0 = \beta''(\theta)\cdot\mathbb{E}[Y] + \gamma''(\theta) + \big(\beta'(\theta)\big)^2\int\left(y + \frac{\gamma'(\theta)}{\beta'(\theta)}\right)f(y)dy$$

aussi, on peut écrire

$$0 = -\beta''(\theta)\frac{\gamma'(\theta)}{\beta'(\theta)} + \gamma''(\theta) + \big(\beta'(\theta)\big)^2\text{Var}[Y]$$

soit

$$\text{Var}[Y] = \frac{\beta''(\theta)\gamma'(\theta) - \beta'(\theta)\gamma''(\theta)}{[\beta'(\theta)]^3}$$

Ici, $\beta(\theta) = \log[\theta/(1-\theta)]$ i.e. $\beta(\theta) = \log(\theta) - \log(1-\theta)$ de telle sorte que

$$\beta'(\theta) = \frac{1}{\theta} + \frac{1}{1-\theta} = \frac{1}{\theta(1-\theta)}$$

et

$$\beta''(\theta) = \frac{-1}{\theta^2} + \frac{1}{(1-\theta)^2}$$

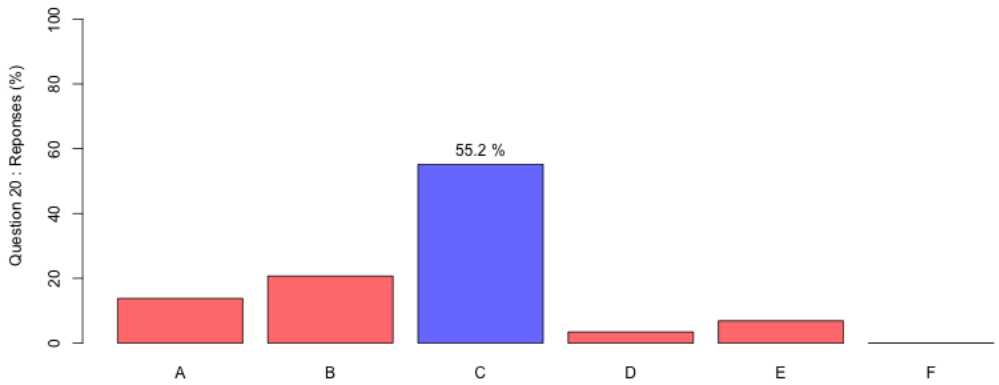
alors que $\gamma(\theta) = 20\log(1-\theta)$, soit

$$\gamma'(\theta) = \frac{-20}{1-\theta} \text{ et } \gamma''(\theta) = \frac{-20}{(1-\theta)^2}$$

Aussi, en substituant

$$\text{Var}[Y] = \frac{\beta''(\theta)\gamma'(\theta) - \beta'(\theta)\gamma''(\theta)}{[\beta'(\theta)]^3} = \cdots = 20\theta(1-\theta)$$

Bref, tout ça pour retrouver la formule de la variance d'une loi binomiale ! La bonne réponse était C.



21 On nous donne les informations suivantes sur une régression logistique, visant à prédire la probabilité d'avoir un accident en fonction de trois variables explicatives x_1, x_2 et x_3.

- la transformation logit de la probabilité diminue de 0.12 quand x_1 augmente de 1
- la dérivée du logarithme de la cote par rapport à x_2 vaut 0.23
- le logarithme de la cote d'un sinistre décroit de 0.34 quand x_3 augmente de 1
- quand $x_1 = x_2 = x_3 = 1$, on prédit une probabilité de 50%.

Quelle est la probabilité d'avoir un accident quand $x_1 = 1$, $x_2 = 2$ et $x_3 = 3$?

A) 40%

B) 50%

C) 60%

D) 70%

E) 80%

Il s'agit d'un exercice ACTEX de 2018. Si $\widehat{p}$ est la probabilité prédite d'avoir un accident avec un modèle logistique

$$\text{logit}(\widehat{p}) \log \frac{\widehat{p}}{1-\widehat{p}} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \widehat{\beta}_3 x_3$$

La première information qu'on nous donne est que $\widehat{\beta}_1 = -0.12$. Pour la second information, notons que

$$\frac{\partial}{\partial x_2} \log \underbrace{\frac{\widehat{p}}{1-\widehat{p}}}_{=\text{cote}} = \frac{\partial}{\partial x_2} \text{logit}(\widehat{p}) = \widehat{\beta}_2$$

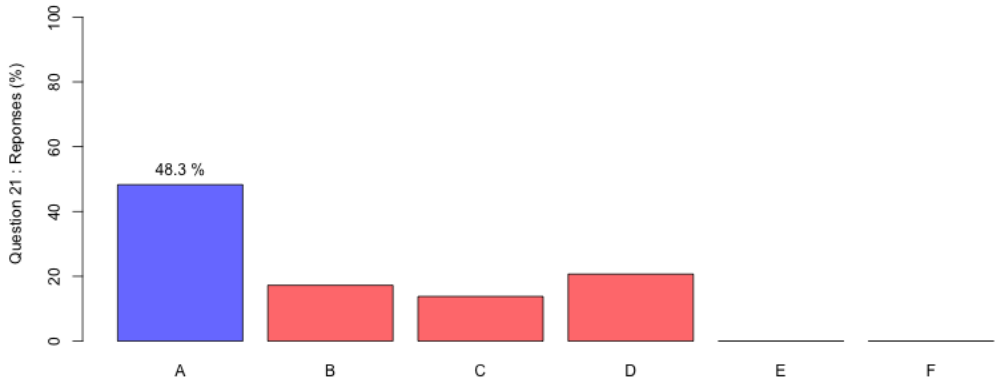
et donc $\widehat{\beta}_2 = 0.23$. La troisième information qu'on nous donne est du même genre que la première, et donc $\widehat{\beta}_3 = -0.34$. Et finalement, on nous dit que

$$\widehat{\beta}_0 + \widehat{\beta}_1 + \widehat{\beta}_2 + \widehat{\beta}_3 = \log 1 = 0$$

de telle sorte que $\widehat{\beta}_0 = 0.23$. Maintenant que l'on a toutes les informations, on peut faire la prévision demandée,

$$\widehat{p} = \frac{e^{\widehat{\beta}_0 + \widehat{\beta}_1 \cdot 1 + \widehat{\beta}_2 \cdot 2 + \widehat{\beta}_3 \cdot 3}}{1 + e^{\widehat{\beta}_0 + \widehat{\beta}_1 \cdot 1 + \widehat{\beta}_2 \cdot 2 + \widehat{\beta}_3 \cdot 3}} = \frac{e^{-0.45}}{1 + e^{-0.45}} = 38.9\%$$

ce qui correspond à la réponse A.



22 On a compté le nombre de décès dans une population donnée, suivant deux critères : le genre - $x_1 \in \{H/F\}$ - et le lieu d'habitation - $x_2 \in \{A/B/C\}$,

N	A	B	C	total
F	21	19	56	96
H	29	33	84	146
total	50	52	140	242

Y	A	B	C	total
F	8	1	21	30
H	19	7	67	93
total	27	8	88	123

On suppose ici que $Y|N, x_1, x_2 \sim \mathcal{B}(N, p_{x_1,x_2})$.

On dispose de la sortie de régression suivante

```
Call:
glm(formula = y ~ x1 + x2, family = quasipoisson(link = "log"))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.1047     0.1729  -6.388 8.77e-10 ***
x1H           0.7353     0.1466   5.014 1.04e-06 ***
x2B          -1.2914     0.2811  -4.594 7.07e-06 ***
x2C           0.1386     0.1536   0.902    0.368
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be ? )

    Null deviance: 166.48  on 241  degrees of freedom
Residual deviance: 131.26  on 238  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

Donnez un ordre de grandeur pour la valeur du paramètre de dispersion ϕ

A) moins de 0.9

B) entre 0.9 et 1

C) exactement 1

D) entre 1 et 2

E) plus de 2

Il s'agissait d'un sample exercice pour MAS-I de la CAS. On n'a pas beaucoup d'information ici.. On nous donne

```
Residual deviance:  131.26 on 238 degrees of freedom
```

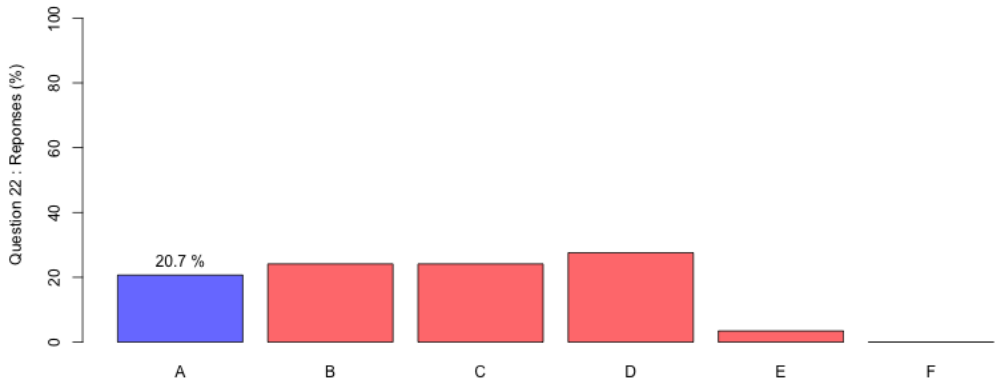
Classiquement, on utiliserait

$$\widehat{\phi} = \frac{1}{\text{degrés de libertés}} \sum \text{carrés des résidus de Pearson}$$

mais comme on a seulement la somme des carrés des résidus de déviance, on pourrait dire

$$\widehat{\phi} \approx \frac{1}{\text{degrés de libertés}} \sum \text{carrés des résidus de Pearson} = \frac{238}{131.26} = 0.5515$$

donc a priori la réponse A est la plus crédible. En réalité, si on refait les calculs sur ordinateur, on obtient 0.4875992.



23 On dispose de la sortie de régression suivante

```
Call:
glm(formula = y ~ color + zone + Age, family = Gamma(link = "log"))

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)          2.100
colorBlack           1.200
colorBlue            1.300
colorGrey            2.000
colorRed             1.406
colorWhite           1.875
zone2                7.678
zone3                4.227
zone4                1.336
zone5                1.734
AgeOld               1.800
AgeYoung             2.000
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 1.0000)
```

Calculer l'écart-type de la charge sinistre pour une observation dans la zone 1, pour un véhicule gris (`Grey`) et un conducteur d'âge intermédiaire (ni `Old` ni `Young`).

A) moins de 55

B) entre 55 et 60

C) entre 60 et 65

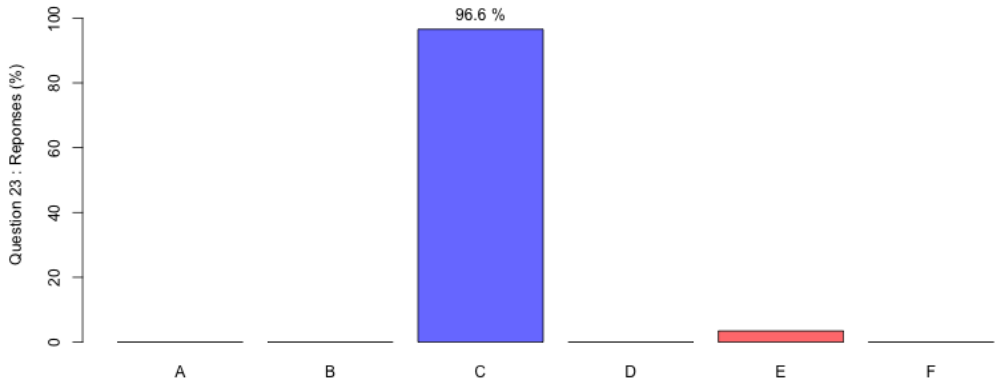
D) entre 65 et 70

E) plus de 70

C'est une question de l'examen CAS-S de printemps 2016 (adapté). On a un lien log, ce qui signifie que

$$\log \widehat{\mu} = 2.100 + 2.000 = 4.100$$

de telle sorte que $\widehat{\mu} = \exp[4.1] \sim 60.340$. Or on sait que la fonction variance est μ^2 pour une loi Gamma, donc la variance serait 60.34^2 et l'écart-type $\sqrt{60.34^2} = 60.34$, qui correspond à la réponse C.



24 On obtient la sortie suivante dans une régression logistique, estimant la probabilité d'un sinistre en fonction du revenu hebomadaire (`weekly_income`, en '000$) et des dépenses hebdomadaires (`weekly_expenditure`, en '000$).

```
Coefficients
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)           0.3321
weekly_income        -2.3700
weekly_expenditure    4.0400
```

Quelle serait la probabilité d'avoir un sinistre pour un assuré ayant un revenu hebdomadaire de 518$ et qui dépense, toutes les semaines, 50.2% de son revenu ?

A) 46%

B) 50%

C) 54%

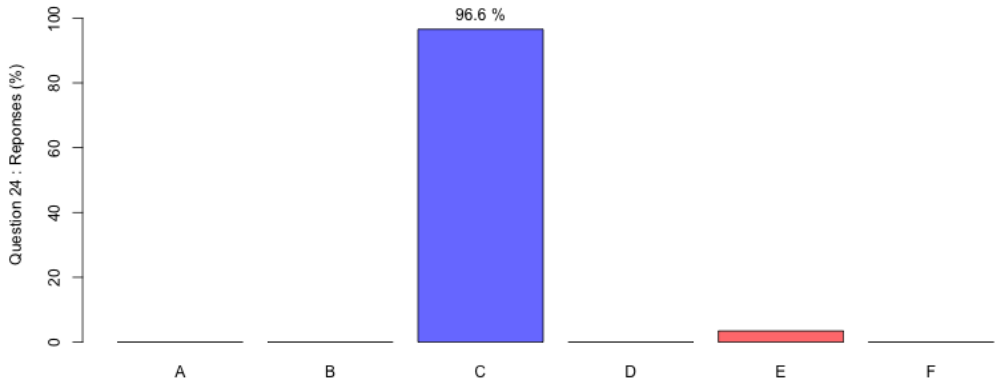
D) 58%

E) 62%

Il s'agissait d'un exercice ACTEX de 2018 (désolé pour la formulation un peu tortueuse). Sur le principe, il s'agit simplement de calculer une prévision pour x_1 (`weekly_income`, en '000$) valant 0.518, et x_2 `weekly_expenditure`, en '000$) valant 50.2% de 0.518 soit 0.260. Comme on a un modèle logistique, on utilise

$$\mathbb{P}[Y=1|x_1,x_2] = \frac{1}{1+e^{-0.3321-2.3700\times 0.518+4.0400\times 0.260}} = \frac{1}{1+e^{-0.154985}} = 0.5387$$

qui correspond à la réponse C.



25 Un actuaire souhaite prédire une probabilité, à partir d'observations d'occurence ($y = 1$) ou pas ($y = 0$) d'un évènement. Il tente un modèle linéaire, et obtient des prévisions inférieures à 0 (pour certaines) ou supérieures à 1 (pour d'autres). Quelle solution serait la plus appropriée

A) se limiter aux obervations telle que la prévision soit comprise entre 0 et 1

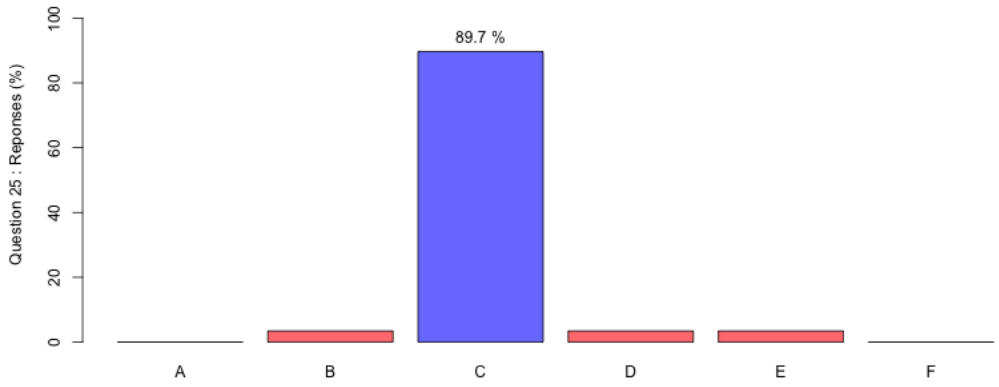
B) transformer la prévision : toute prévision inférieure à 0 devient 0, et toute prévision supérieure à 1 devient 1

C) utiliser une régression probit afin de transformer le modèle linéaire en un modèle dont la prévision sera comprise entre 0 et 1

D) transformer la variable d'intérêt à l'aide la fonction de lien canonique du modèle binomial et estimer ensuite un modèle linéaire

E) aucune des propositions

A) on évite de retirer des données, surtout certaines données car cela induit un biais dans les estimation. B) non, pour la même raison qu'auparavant. C) oui, c'est l'idée de la régression logistique. D) non ! la fonction de lien canonique c'est la fonction quantile de la loi logistique, i.e. $p \mapsto \log(p/(1-p))$. Autrement dit, on transforme 0 en $-\infty$ et 1 en $+\infty$. On retient donc C.



26 Lors d'une enquête, une question propose les réponses suivantes, 'negatif', 'neutre' et 'positif'. On veut expliquer les choix à l'aide de trois variables,

- $x_1 = 1$ si la personne a étudié à l'université
- $x_2 = 1$ si la personne a de fortes convictions religieuses
- $x_3 = x_1 \cdot x_2$

Un modèle logistique-multinomial donne la sortie suivante ('positif' est ici la référence)

```
Coefficients (category 'negative')
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    -0.70
x1             -0.56
x2              1.45
x3             -0.28

Coefficients (category 'neutre')
            Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.60
x1             -1.25
x2              1.84
x3              0.23
```

Pour une personne ayant étudié à l'université et ayant de fortes convictions religieuse, quelle est la probabilité d'avoir un avis 'neutre' à la question ?

A) moins de 35%

B) entre 35% et 45%

C) entre 45% et 55%

D) entre 55% et 65%

E) plus de 65%

On reconnait la régression multinomiale logistique, et comme on l'avait vu en cours (et auparavant à l'exercice 6), le principe est que pour la modalité de référence, $\mathbb{P}[Y = \text{positive}|\boldsymbol{x}] \propto 1$ alors que $\mathbb{P}[Y = \text{negative}|\boldsymbol{x}] \propto \exp[\boldsymbol{x}^\top \boldsymbol{\alpha}]$ et $\mathbb{P}[Y = \text{neutre}|\boldsymbol{x}] \propto \exp[\boldsymbol{x}^\top \boldsymbol{\beta}]$. Aussi, si on normalise, on a (par exemple)

$$\mathbb{P}[Y = \text{neutre}|\boldsymbol{x}] = \frac{\exp[\boldsymbol{x}^\top \boldsymbol{\beta}]}{1 + \exp[\boldsymbol{x}^\top \boldsymbol{\alpha}] + \exp[\boldsymbol{x}^\top \boldsymbol{\beta}]}$$

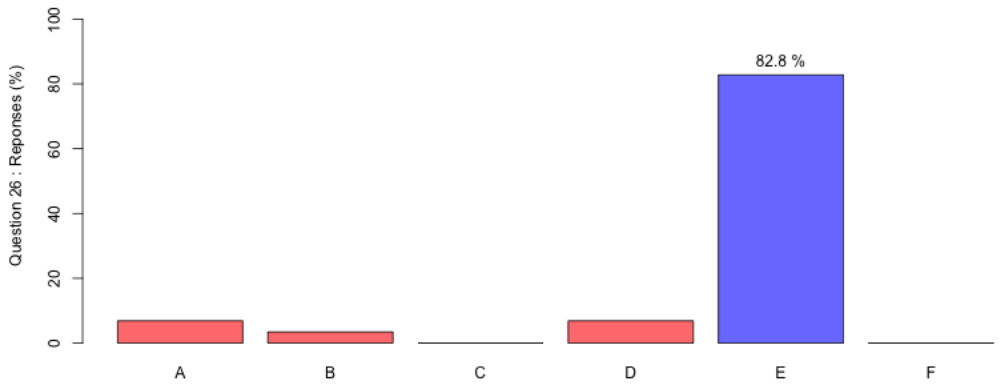
Pour avoir la valeur de la prévision, on se place en $(x_1, x_2, x_3) = (1, 1, 1)$, autrement dit, on doit calculer l'exponentielle de la somme des quatre coefficients,

$$\exp(-0.7 - 0.56 + 1.45 - 0.28) = 0.9139312 \text{ et } \exp(0.60 - 1.25 + 1.84 + 0.23) = 4.13712$$

donc ici l'estimateur de $\mathbb{P}[Y = \text{neutre}|\boldsymbol{x}]$ est

$$\frac{4.13712}{1 + 4.13712 + 0.9139312} \sim 68.37\%$$

On retiendra donc la réponse E.



27 On a construit un modèle logistique pour modéliser la probabilité d'avoir un incendie dans un logement, à l'aide d'une variable explicative x correspondant au nombre de pièces du logement. Pour 3 pièces, la probabilité est de 6%, contre 7% pour 4 pièces. Que vaut le paramètre $\widehat{\beta}_0$ associé à la constante dans la régression logistique ?

A) moins de -5

B) entre -5 et -4.5

C) entre -4.5 et -4

D) entre -4 et -3.5

E) plus de -3.5

Il s'agissait d'un exercice (adapté) de l'examen MAS-I de la CAS du printemps 2019. On sait que

$$\text{logit}(\widehat{\beta}_0 + \widehat{\beta}_1 \cdot 3) = 0.06 \text{ et } \text{logit}(\widehat{\beta}_0 + \widehat{\beta}_1 \cdot 5) = 0.07$$

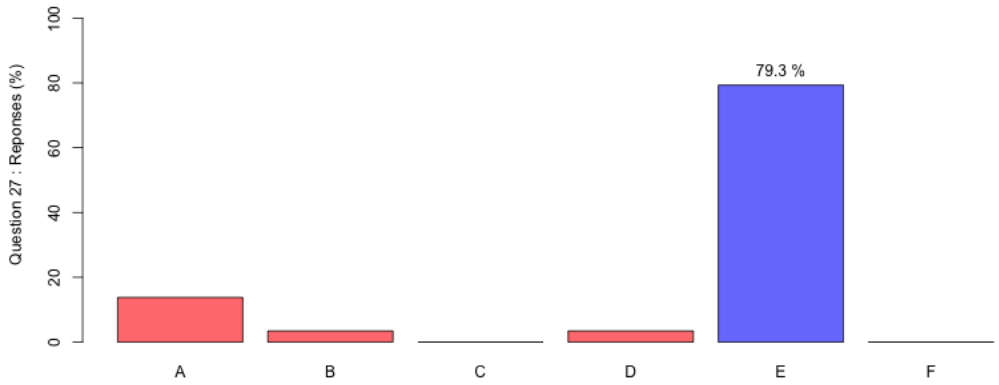
donc

$$\widehat{\beta}_0 + \widehat{\beta}_1 \cdot 3 = \log \frac{0.06}{1-0.06} \text{ et } \widehat{\beta}_0 + \widehat{\beta}_1 \cdot 4 = \log \frac{0.07}{1-0.07}$$

Par soustraction, on en déduit que $\widehat{\beta}_1 = 0.164846$ et donc

$$\widehat{\beta}_0 = \log \frac{0.06}{1-0.06} - \widehat{\beta}_1 \cdot 3 = -2.751535 - 3 \times 0.164846 = -3.24607$$

ce qui correspond à la réponse E.



Les question 28 à 30 sont basées sur les deux sorties de régression suivantes

Deux modèles de Poisson sont envisager pour modéliser une variable de comptage y en fonction de trois variables continues x_1, x_2 et x_3. Le modèle (123) est

```
Call:
glm(formula = y ~ x1+x2+x3, family = poisson)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.7719     0.8910   1.989   0.0467 *
X1            1.6350    15.3774   0.106   0.9153
X2            1.0897     0.1035  10.528   <2e-16 ***
X3           -4.1656    30.7287  -0.136   0.8922

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 269.730  on 19  degrees of freedom
Residual deviance:  16.068  on 16  degrees of freedom
AIC: 73.272
```

et le modèle (13) est

```
Call:
glm(formula = y ~ x1+x3, family = poisson)
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.8354     0.5813  -1.437    0.151
X1            68.8018    14.2170   4.839 1.30e-06 ***
X3          -136.0553    28.5669  -4.763 1.91e-06 ***

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 269.73  on 19  degrees of freedom
Residual deviance: 190.89  on 17  degrees of freedom
AIC: 246.10
```

28 On considère une nouvelle observation, $\boldsymbol{x} = (x_1, x_2, x_3) = (2, 3, 1)$, et on note $\widehat{y}_{(123)}$ et $\widehat{y}_{(13)}$ les prévisions avec les deux modèles. Que vaut $\widehat{y}_{(123)} - \widehat{y}_{(13)}$?

A) moins de 0

B) entre 0 et 10

C) entre 10 et 20

D) entre 20 et 50

E) plus de 50

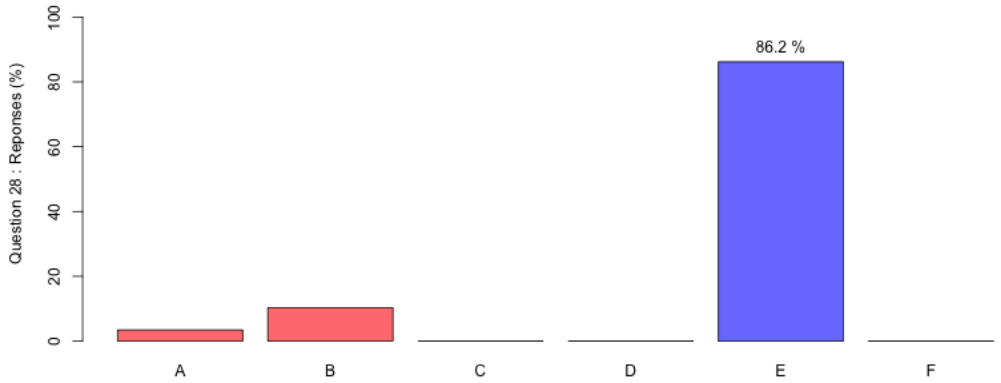
On a ici deux régressions de Poisson, on devrait pouvoir y arriver.

$$\widehat{y}_{(123)} = \exp\left[\boldsymbol{x}^\top \widehat{\boldsymbol{\beta}}\right] = \exp\left[1.7719 + 1.6350 \times 2 + 1.0897 \times 3 - 4.1656 \times 1\right] = \exp[4.14556] = 63.15$$

alors que

$$\widehat{y}_{(13)} = \exp\left[\boldsymbol{x}^\top \widehat{\boldsymbol{\beta}}\right] = \exp\left[-0.8354 + 68.8018 \times 2 - 136.0553 \times 1\right] = \exp[0.7129] = 2.04$$

ce qui fait une différence de l'ordre de 60, ce qui est la réponse E.



29 Dans le premier modèle, donner un intervalle de confiance (approximatif) à 95% où devrait se trouver y, pour une observation $\boldsymbol{x} = (x_1, x_2, x_3) = (2, 3, 1)$ (on retiendra les valeurs les plus proches)

A) $[62.3, 63.9]$

B) $[60, 66]$

C) $[55, 70]$

D) $[48, 80]$

E) $[30, 100]$

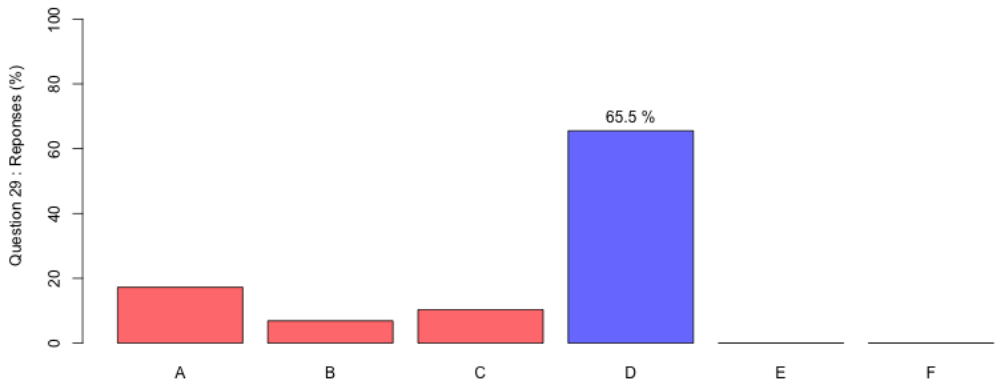
Pour rappel, Y suit une loi de Poisson de paramètre 63.15, d'après la question précédante. On peut alors utiliser une approximation Gaussienne de la loi de Poisson,

$$\mathcal{P}(\lambda) \approx \mathcal{N}(\lambda, \lambda)$$

donc un intervalle de confiance (approximatif) à 95% sera de la forme

$$\left[\widehat{\lambda} \pm 1.96\sqrt{\widehat{\lambda}}\right] = \left[63.15 \pm 1.96\sqrt{63.15}\right] = [47.57; 78.73]$$

Pour information, si on regarde les quantiles à 2.5% et 97.5% de la loi de Poisson de paramètre 63.15, on obtient $[48, 79]$. La bonne réponse sera ici la réponse D.



30 Avec le second modèle, on sait que la somme des carrés des résidus de Pearson vaut 211.9841. Si on faisait une régression quasi-Poisson, donnez la valeur de $\widehat{\phi}$

```
Call:
glm(formula = y ~ x1+x3, family = quasipoisson)

(Dispersion parameter for poisson family taken to be  ? )
```

A) moins de 0.9

B) entre 0.9 et 1.1

C) entre 1.1 et 2.0

D) entre 2 et 10

E) plus de 10

Dans la sortie initiale, on nous disait

```
Residual deviance:  190.89 on 17 degrees of freedom
```

On nous dit maintenant que la somme des carrés des résidus de Pearson vaut 211.9841. On utilise alors

$$\widehat{\phi} = \frac{1}{\text{degrés de libertés}} \sum \text{carrés des résidus de Pearson} = \frac{211.9841}{17} = 12.46965$$

ce qui correspond à la réponse E. Pour information, ces trois questions venaient d'une partie d'un exercice du cours Applied Statistical Regression à l'ETH Zürich (de 2011).