STT5100 - Hiver 2019 - Examen Final (GLM)

Arthur Charpentier

Examen **B**

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
probit 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

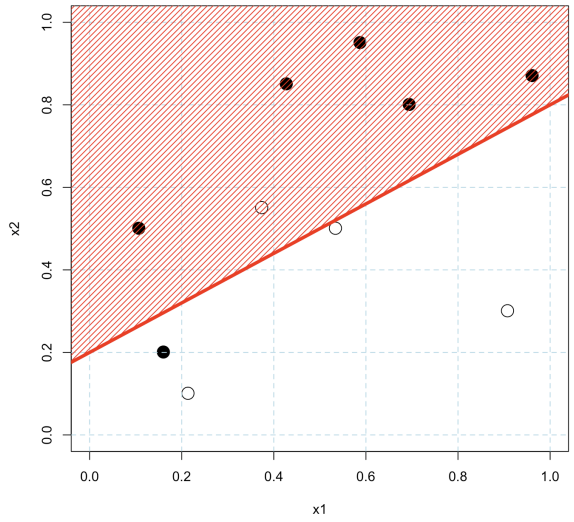
$$f(y;\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right)$$

Code permanent :

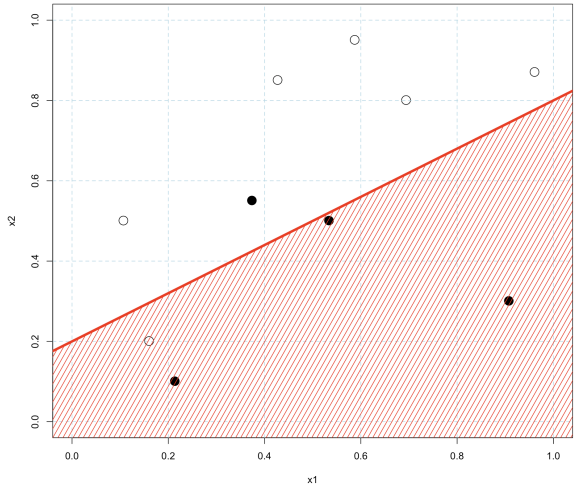
Sujet : B

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	Figure à droite (à compléter)				
question 8	☐ A	☐ B	☐ C	■ D	☐ E
question 9	☐ A	☐ B	☐ C	☐ D	■ E
question 10	☐ A	☐ B	☐ C	☐ D	■ E
question 11	☐ A	☐ B	☐ C	☐ D	■ E
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	☐ A	■ B	☐ C	☐ D	☐ E
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 7 (examen B):



question 17 (examen A):



1 Soit $f(y;\theta,\phi)$ une loi de la famille exponentielle. Quelles seraient les valeurs de θ, $b(\theta)$ et ϕ pour avoir une loi de Poisson de moyenne λ ?

A) $\theta=\lambda$, $b(\theta)=\theta$ et $\phi=1$

B) $\theta=\lambda$, $b(\theta)=e^{\theta}$ et $\phi=1$

C) $\theta=\log\lambda$, $b(\theta)=\theta$ et $\phi=1$

D) $\theta=\log\lambda$, $b(\theta)=\theta$ et $\phi=1$

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

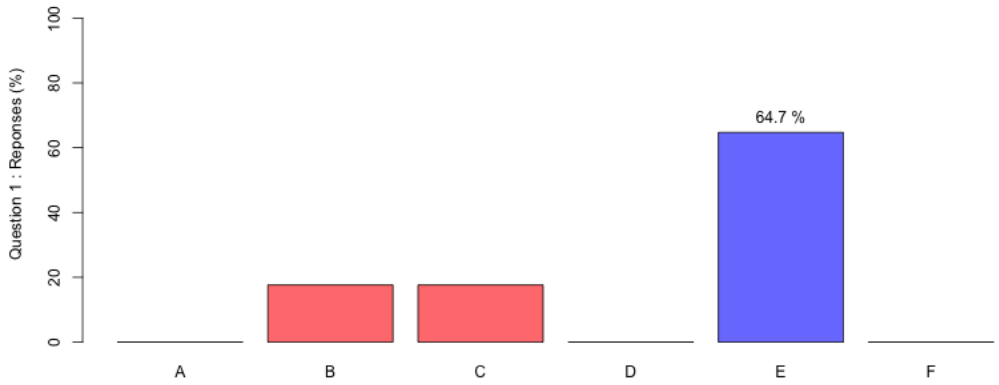
Il s'agit d'une question de l'examen S de la CAS d'automne 2017, qui avait été traitée plusieurs fois en cours. Rappelons que la loi de Poisson de moyenne λ admet pour fonction de masse

$$f(y;\lambda)=e^{-\lambda}\frac{\lambda^y}{y!}=\exp\left(y\log(\lambda)-\lambda-\log(y!)\right)=\exp\left(\frac{y\log\lambda-\lambda}{1}-\log(y!)\right)$$

que l'on va identifier à

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

(les couleurs vont simplifier et éviter de longs discours). 1) on commence par le terme en rouge, qui nous permet d'écrire $\theta=\log\lambda$. 2) on passe au terme en vert : $b(\theta)=\lambda$ avec $\lambda=e^{\theta}$, donc $b(\theta)=e^{\theta}$. 3) pour le paramètre de nuisance, on a 1 (sinon, techniquement, on aurait une loi dite *quasi*-Poisson). On retrouve les trois éléments de la réponse... ben aucune parce que j'ai fait une faute de frappe.... (dans l'examen de la CAS, la réponse D était l'analogue de B, autrement dit $b(\theta)=e^{\theta}$ - qui est la bonne).



2 Soit $f(y;\theta,\phi)$ une loi de la famille exponentielle, et Y une variable aléatoire suivant cette loi. On nous donne

$$b(\theta)=-\sqrt{-2\theta},\ \theta=-0.3 \text{ et } \phi=1.6$$

Que vaut $\mathbb{E}[Y]$?

A) moins de -1

B) entre -1 et 0

A) entre 0 et 1

D) entre 1 et 2

E) plus que 2

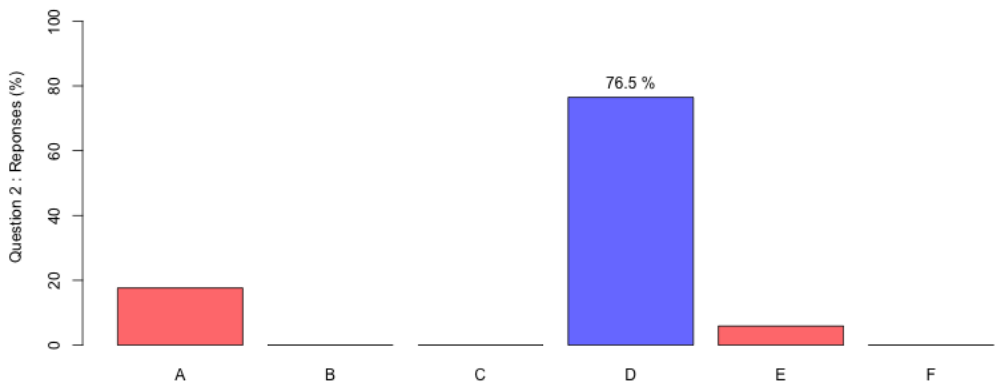
Il s'agit d'une question de l'examen S de la CAS d'automne 2015. On avait mentionné plusieurs fois cette propriété en cours : pour un variable dont la loi est dans la famille exponentielle, $\mathbb{E}[Y] = b'(\theta)$, soit ici

$$b'(\theta) = -\frac{1}{2}(-2\theta)^{-1/2}(-2) = (-2\theta)^{-1/2}.$$

Or ici on nous dit que θ vaut -0.3, donc

$$b'(\theta) = (-2 \cdot (-0.3))^{-1/2} = \frac{1}{\sqrt{0.6}} = 1.2910$$

ce qui correspond à la réponse D. Pour information, on peut reconnaitre une distribution inverse Gaussienne... Et oui, ϕ ne sert à rien, ϕ intervient juste dans les calculs de variance.



3 Une famille de distributions de la famille exponentielle est définie par une relation de la forme $\text{Var}[Y] = a\mathbb{E}[Y]^p$. Considérons les trois affirmations suivantes

1. si $p = 0$ on a une loi normale
2. si $p \in (1, 2)$ on a une loi composée Poisson-Gamma
3. si $p = -1$ on a une loi inverse Gaussienne

Quelle(s) affirmation(s) est(sont) vraie(s) ?

A) (1) seulement

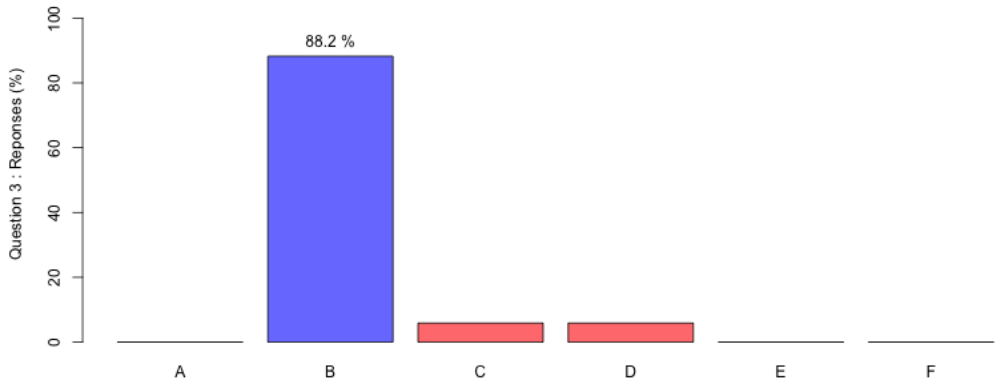
B) (1) et (2) seulement

C) (1) et (3) seulement

D) (2) et (3) seulement

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

Il s'agissait d'une question de l'examen S de la CAS de l'automne 2017. Et ça correspond à des choses vue en cours... Oui, le modèle Gaussien est homoscédastique, de variance (conditionnelle) constante, soit $p = 0$, donc 1 est juste. Oui, la loi Tweedie - correspondant à un modèle Poisson composé avec des sauts Gamma - est obtenu avec $p \in (1, 2)$. Techniquement, je pense qu'on pourrait rajouter les bords, $p = 1$ correspond à une loi de Poisson (les sauts sont déterministes) et $p = 2$ à la loi Gamma (le nombre de sauts est alors déterministe). Donc 2 est juste aussi. Par contre, la loi inverse Gaussienne correspond à $p = 3$. 3 est fausse, donc il fallait choisir la réponse B.



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

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

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)     5.26
riskB           0.18
riskC           0.37
territory2      0.12
territory3      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 250

B) entre 250 et 275

C) entre 275 et 300

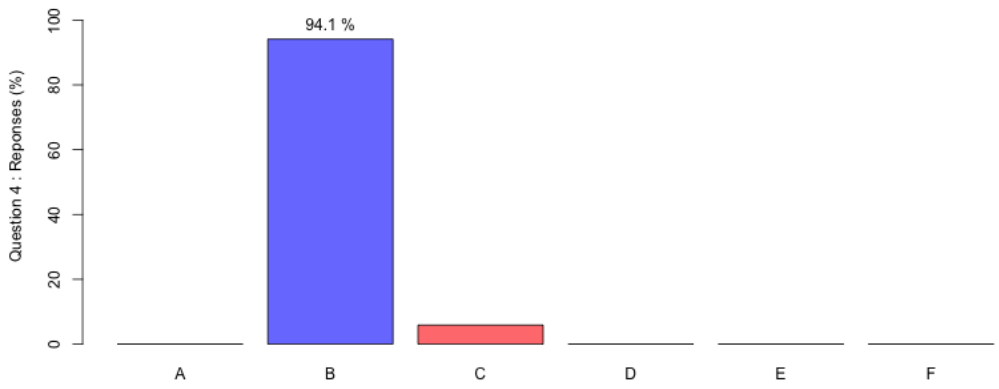
D) entre 300 et 325

E) plus de 325

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 B. Et comme vu en cours, la loi ne sert à rien pour la prévision.



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

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

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)       2.32
locationRural    -0.64
genderMale        0.76
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 2.00)
```

Calculer la variance de la prévision de la perte pour un homme habitant à la campagne

A) moins de 25

B) entre 25 et 100

C) entre 100 et 175

D) entre 175 et 250

E) plus de 250

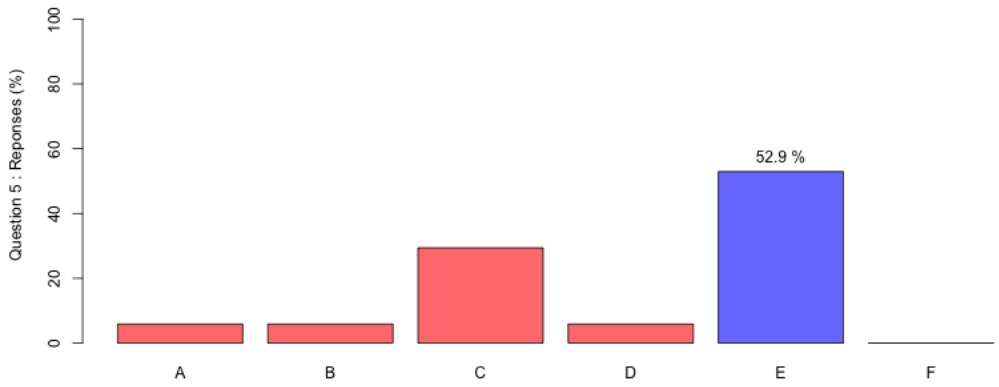
Il s'agit là encore d'une question de l'examen S de la CAS (d'automne 2015). On va procéder en deux temps : on commence par calculer la prévision (fonction lien) puis on calcule la variance (fonction variance - loi Gamma). On a un lien logarithmique, donc

$$\log \widehat{\mu} = \boldsymbol{x}^\top \boldsymbol{\beta} = 2.32 + \underbrace{(-0.64)}_{\text{campagne}} + \underbrace{0.76}_{\text{homme}} 2.44 \text{ donc } \widehat{\mu} = e^{2.44} \ (\sim 11.473)$$

mais la valeur numérique importe peu... Maintenant, on sait que la fonction variance est en μ^2, donc ici,

$$\widehat{\text{Var}}[Y|\boldsymbol{X} = \boldsymbol{x}] = \phi \cdot \widehat{\mu}^2 = 2 \cdot (e^{2.44})^2 \sim 263.2613$$

ce qui correspondrait à la réponse E.



6 On peut lire les affirmations suivantes à propos de la déviance des GLM

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

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

(3) un modèle saturé a une déviance nulle

A) aucune affirmation n'est juste

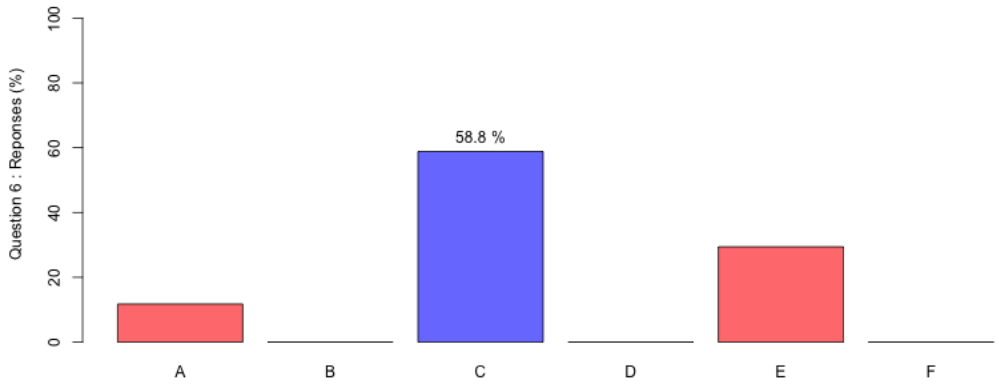
B) (1) et (2) sont justes

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) 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 ! 2) 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 seconde affirmation est donc fausse... 3) Oui, c'est la définition... Autrement dit, seules les affirmations 1 et 3 sont valides, ce qui correspond à la réponse C.



7 On dispose de $n = 10$ observations $(y_i, x_{1,i}, x_{2,i})$, avec $y \in \{0, 1\}$. Une régression logistique donne la sortie suivante

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

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.714      2.001  -0.856    0.392
x1            -5.141      5.021  -1.024    0.306
x2             8.568      5.515   1.554    0.120
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

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

je corrige ici le sujet B ! pour le sujet A... je vous laisse convertir la correction Cherchons pour commencer la frontière de l'ensemble, c'est à dire l'ensemble des points (x_1, x_2) tels que

$$\mathbb{P}[Y = 1|\boldsymbol{x} = (x_1, x_2)] = \frac{1}{2}$$

c'est à dire tels que

$$\frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2}} = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2)}} = \frac{1}{2}$$

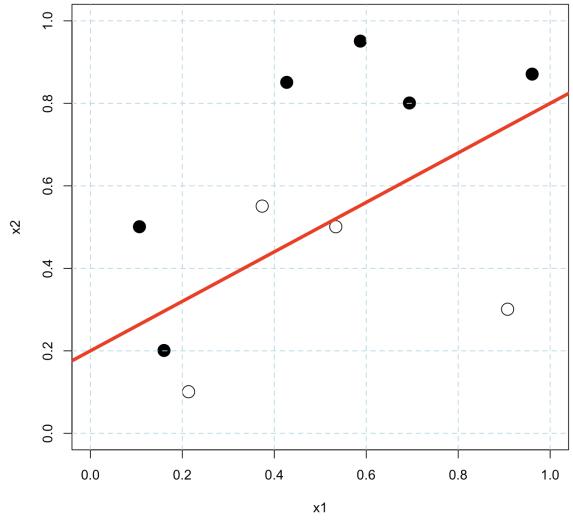
soit

$$e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2)} = 1, \text{ i.e. } \beta_0 + \beta_1 x_1 + \beta_2 x_2 = 0$$

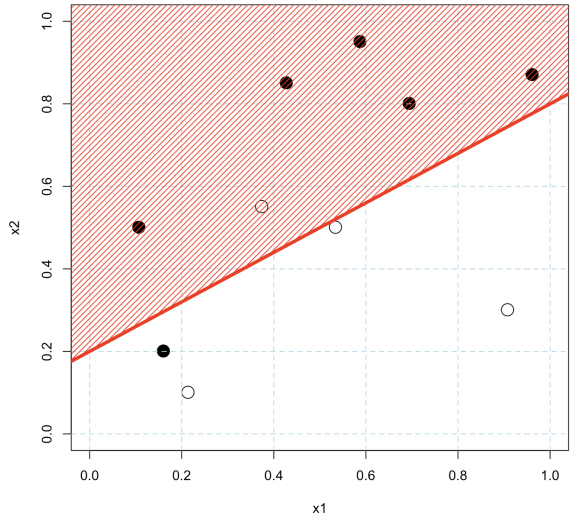
Comme vu en cours, on retrouve une *droite*. Plus précisément, la droite

$$x_2 = \frac{-1}{\beta_2}(\beta_0 + \beta_1 x_1) = \frac{-1}{8.568}(-1.714 - 5.141 x_1) = 0.2 + 0.6 x_2$$

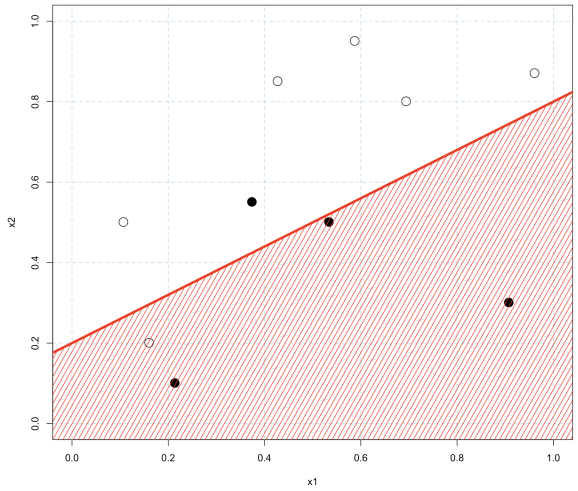
ce qui correspond à une droite qui passe par $(0, 0.2)$ et $(0, 0.6)$. C'est la droite rouge ci-dessous



On faire des calculs, mais un peu de bon sens laisse à penser que les points noirs $(y = 1)$ sont en haut à gauche, alors que les points blancs $(y = 0)$ sont en bas à droite. Donc la partie qu'on cherche est au dessus de la droite.



Pour le sujet A, les calculs sont proches, aux signes près, aussi, c'est la partie inférieure du plan qui correspond à la région demandée



8 On suppose que les variables Y_i sont des variables indépendantes, de distribution $\mathcal{P}(\mu_i)$, et on note $\widehat{\mu}_i$ les moyennes prédites. Donnez l'expression de la déviance

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

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

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

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

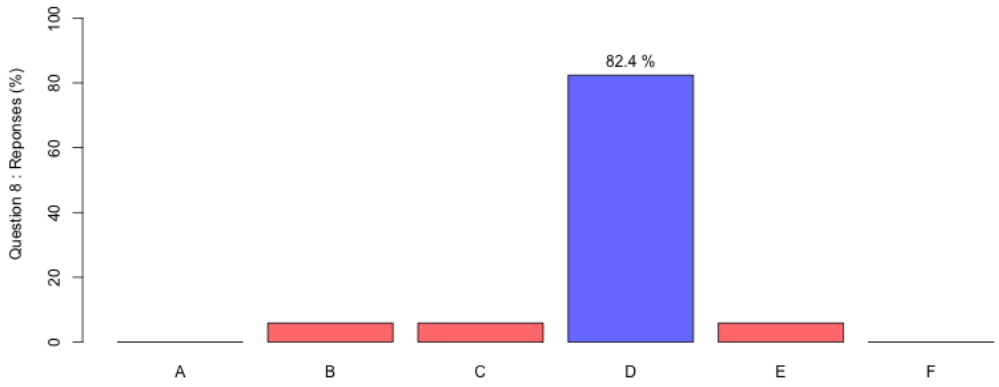
E) $2\sum_{i=1}^{n} y_i \log \frac{y_i}{\mu_i} - (y_i - \mu_i)$

On va faire les calculs : la log-vraisemblance est ici

$$\log L(\boldsymbol{\mu}) = \sum_{i=1}^{n} (-\mu_i + y_i \log \mu_i - \log(y_i!))$$

Comme la déviance est la distance entre le modèle saturé ($\mu_i = y_i$) et le modèle estimé ($\mu_i = \widehat{\mu}_i$), (avec un facteur 2)

$$D = 2(\log \mathcal{L}(\boldsymbol{y}) - \log \mathcal{L}(\widehat{\boldsymbol{\mu}})) = 2\sum_{i=1}^{n} y_i \log \frac{y_i}{\widehat{\mu}_i} - (y_i - \widehat{\mu}_i)$$



9 On nous donne les affirmations suivantes :

1 la déviance est utile pour tester la significativité de variables explicatives dans des modèles imbriqués

2 la déviance dans le cas Gaussien est proportionnelle à la somme des carrés des résidus

3 la déviance est définie comme une distance entre le modèle saturé et le modèle estimé

Lesquelles sont justes ?

A) 1 seulement

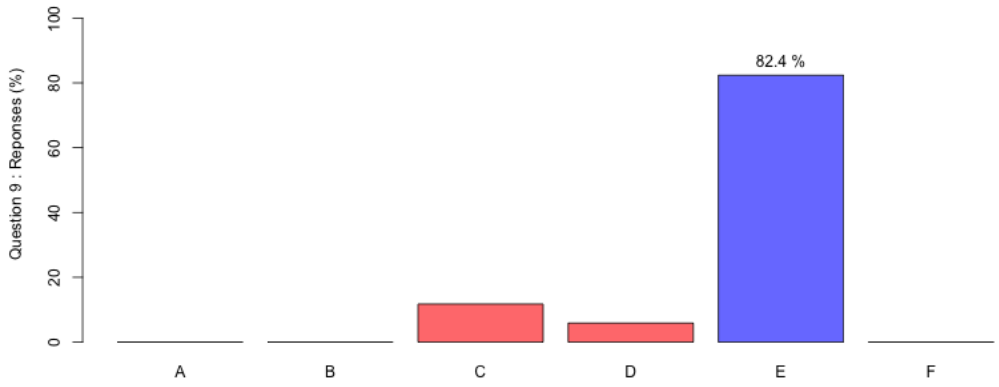
B) 2 seulement

C) 3 seulement

D) 1 et 2

E) 1, 2 et 3

Il s'agissait d'une question de l'examen S de la CAS du printemps 2016. Les trois affirmations sont justes. 1) xxxxxxx 2) On l'avait fait en cours, la *scaled deviance* est égale à la somme des carrés des résidus, le coefficient de proportionnalité est ici juste l'inverse de la variance des résidus. 3) on l'a mentionné à plusieurs reprise déjà...



10 On suppose que des observations y sont distribuées suivant une loi exponentielle, conditionnellement aux variables explicatives, avec un lien logarithmique

$$f(y_i) = \frac{\exp[-y_i/\theta_i]}{\theta_i} \text{ avec } \log(\theta_i) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{2,i}^2 + \beta_4 x_{3,i}.$$

On nous donne les observations suivantes, $\boldsymbol{y}$, $\boldsymbol{X}$ et $\widehat{\boldsymbol{\beta}}$

$$\boldsymbol{y} = \begin{pmatrix} 14.8 \\ 137.6 \\ 0.4 \\ 38.3 \\ \vdots \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} 1 & 1 & 2.7 & 7.29 & 1 \\ 1 & 0 & 0.6 & 0.36 & 1 \\ 1 & 1 & 2.9 & 8.41 & 1 \\ 1 & 1 & 3.0 & 9.00 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix} \text{ et } \widehat{\boldsymbol{\beta}} = \begin{pmatrix} 2.99 \\ -0.27 \\ -0.67 \\ 0.16 \\ 0.91 \end{pmatrix}.$$

Donnez la valeur du résidu de Pearson pour la seconde observation $\widehat{\varepsilon}_2$

A) moins de 2

B) moins de -2

B) entre -2 et -1

C) entre -1 et +1

D) entre +1 et +2

E) plus de +2

Il s'agissait d'une question de l'examen CAS - S d'automne 2017 (adaptée). On a un lien logarithmique donc la prévision est

$$\log \widehat{\mu}_2 = \boldsymbol{x}_2^\intercal \widehat{\boldsymbol{\beta}} = 2.99 + 0.6 \cdot (-0.67) + 0.36 \cdot 0.16 + 0.91 = 3.5556$$

de telle sorte que $\widehat{\mu}_2 = \exp[3.5556] \sim 35.0088$.

On peut ensuite se souvenir que la loi exponentielle est un cas particulier de lois Gamma, donc la fonction variance est en $\mathsf{V}(\mu) = \mu^2$. Sinon on peut refaire le calcul.

Considérons une loi exponentielle de moyenne $\mu = 1/\lambda$,

$$f(y, \lambda) = \lambda e^{-\lambda y} = \exp[-\lambda y + \log \lambda]$$

Comme on l'a vu en cours, θ est ici $-\lambda$, $\phi = 1$, et on recommait $b(\theta) = -\log\lambda = -\log(-\theta)$. On peut alors calculer les deux premiers moments :

$$\mathbb{E}[Y] = b'(\theta) = \frac{-1}{\theta} \text{ et } \mathsf{Var}[Y] = b''(\theta) = \frac{1}{\theta^2}.$$

On vérifie que ces valeurs correspondent à celles que l'on connaît

$$\mathbb{E}[Y] = \frac{-1}{\theta} = \frac{-1}{-\lambda} = \mu = \text{ et } \mathsf{Var}[Y] = \frac{1}{\theta^2} = \frac{1}{\lambda^2} = \mu^2.$$

Cette dernière expression nous donne la forme de la fonction variance, au passage : $\mathsf{V}(\mu) = \mu^2$.

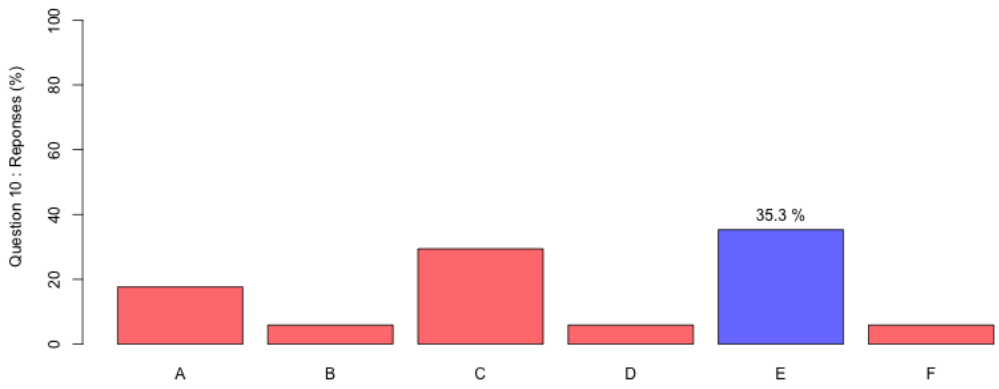
Bref, on peut maintenant calculer le résidu de Pearson,

$$\widehat{\varepsilon}^{\mathsf{Pearson}} = \frac{y - \widehat{\mu}}{\sqrt{\widehat{\mu}^2}} = \frac{y - \widehat{\mu}}{\widehat{\mu}}$$

Aussi, pour le cas qui nous intéresse,

$$\widehat{\varepsilon}_2^{\mathsf{Pearson}} = \frac{y_2 - \widehat{\mu}_2}{\widehat{\mu}_2} = \frac{137.2 - 35.0088}{35.0088} \sim 2.9190,$$

ce qui correspond à la réponse E.



11 On teste plusieurs modèles sur le même jeu de données, et on note la log-vraisemblance optimale

- seulement avec la constante : $y = \beta_0 + \varepsilon$, $\log\mathcal{L}(\widehat{\boldsymbol{\beta}}) = -1126.91$
- sans terme croisé : $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, $\log\mathcal{L}(\widehat{\boldsymbol{\beta}}) = -1122.41$
- modèle global : $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$, $\log\mathcal{L}(\widehat{\boldsymbol{\beta}}) = -1121.91$

On veut tester $H_0 = \beta_1 = \beta_2 = \beta_3 = 0$ avec un seuil $\alpha = 5\%$, en utilisant le test du rapport de vraisemblance. Quelle affirmation est juste ?

A) la statistique de test vaut 1, et l'hypothèse H_0 ne peut être rejetée

B) la statistique de test vaut 9, et l'hypothèse H_0 ne peut être rejetée

C) la statistique de test vaut 10, et l'hypothèse H_0 ne peut être rejetée

D) la statistique de test vaut 9, et l'hypothèse H_0 doit être rejetée

E) la statistique de test vaut 10, et l'hypothèse H_0 doit être rejetée

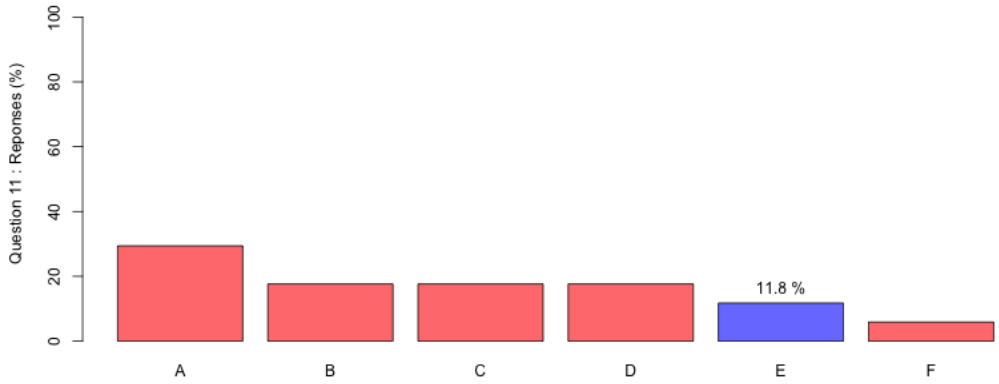
Examen SRM (sample). On revient un instant sur le modèle Gaussien (qui techniquement est aussi un GLM) parce que j'ai présenté le test du rapport de vraissemblance dans ce contexte. Sous l'hypothèse H_0 on a le premier modèle, avec juste la constante. L'hypothèse alternative (H_1) est le troisième. Le test du rapport de vraisemblance est basé sur la statistique

$$T = 2\big(\log \mathcal{L}_1 - \log \mathcal{L}_0\big) = 2\big(-1121.91 - (-1126.91)\big) = 10$$

On a maintenant le choix entre C et E, sur l'interprétation. Le nombre de degré de liberté est la différence entre les nombres de variables explicatives.

On a ici 3 degrés de liberté. On va alors chercher dans la table le quantile la loi du $\chi^2(3)$ à un niveau de 95%, soit 7.815. C'est la valeur supérieure de la région d'acceptation, et comme 10¿7.815, on doit donc **rejeter** H_0, ce qui correspond à la réponse E. Je suis supris par le faible nombre de bonnes réponses

En fait, on peut montrer que la p-value est de l'ordre de 1.85% .



12 Un actuaire ajuste deux modèles GLMs, M_1 et M_2 afin de prévoir la probabilité qu'une personne achète une couverture assurantielle. On nous donne les informations suivantes

+ modèle M_1, variables explicatives
  - prix offert (*offered price*)
  - nombre de véhicules (*number of vehicles*)
  - âge de l'assuré principal (*age of primary insured*)
  - information sur la possession passée (*prior insurance carier*)

  nombre de degrés de libertés utilisés : 10

  log-vraisemblance $-11,565$
+ modèle M_2, variables explicatives

- prix offert (*offered price*)
- nombre de véhicules (*number of vehicles*)
- âge de l'assuré principal (*age of primary insured*)
- genre de l'assuré principal (*gender of primary insured*)
- score de crédit de l'assuré principal (*credit score of primary insured*)

nombre de degrés de libertés utilisés : 8

log-vraisemblance $-11,562$

On veut savoir lequel des deux modèles est le meilleur. Quelle est la meilleure stratégie à adopter ?

A) Utiliser un test de rapport de vraisemblance

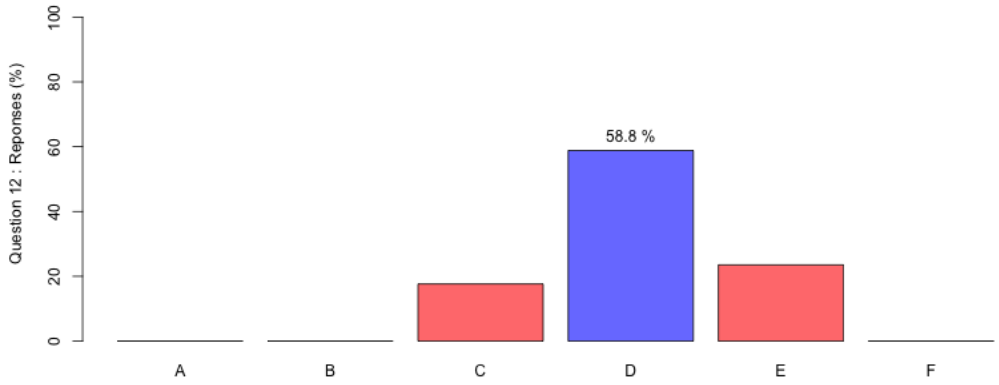
B) Utiliser un test de Fisher

C) Calculer et comparer les déviances des deux modèles

D) Calculer et comparer les AIC des deux modèles

E) Calculer et comparer les tests du chi-deux des deux modèles

Cette question est extraite de l'examen CAS MAS-I (d'automne 2018). Ici les deux modèles ne sont pas imbriqué car chacun prossède une variable explicative que l'autre n'a pas. La statistique de Fisher suppose que les modèles soient imbriqués, donc on ne va pas l'utiliser (ce qui exclue la réponse B). Les deux modèles n'ont pas les mêmes nombres de degrés de liberté, donc on va aussi exclure A et C. D serait la réponse la plus naturelle, parce que c'est justement ce pour quoi le AIC est défini : comparer des modèles de même type, avec des ensembles de variables explicatives différents. Le test du chi-deux est utilisé pour les données groupées donc on va exclure E. Il fallait donc répondre D.



13 On essaye de modéliser la probabilité qu'une famille reste assuré l'an prochain (y est ici `retention`) à l'aide de trois variables explicatives, l'ancienneté `Tenure` qui est une une variable catégorielle prenant 2 modalités (`05Y` (entre 0 et 5 ans) et `5Y` (plus de 5 ans)), la variation de la prime `PriorRate` qui est une une variable catégorielle prenant 3 modalités (`Less0` (moins de 0% - i.e. une baisse), `010` (entre 0 et 10) et `More10` (plus de 10%)), et enfin `AmountInsurance` qui est une une variable numérique (correspondant au montant en '000 de dollars).

```
Call:
glm(retention ~ Tenure + PriorRate + AmountInsurance, family = binomial(link = "probit"))

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)          0.4270
Tenure5Y             0.1320
PriorRateLess0       0.0160
PriorRateMore10     -0.0920
AmountInsurance      0.0015
```

Estimer la probabilité de rétention pour un assuré ayant 4 ans d'ancienneté, ayant eu +12% d'augmentation et pour une police d'un montant de $ 225,000 ?

A) moins de 60%

B) entre 60% et 70%

C) entre 70% et 80%

D) entre 80% et 90%

E) au moins 90%

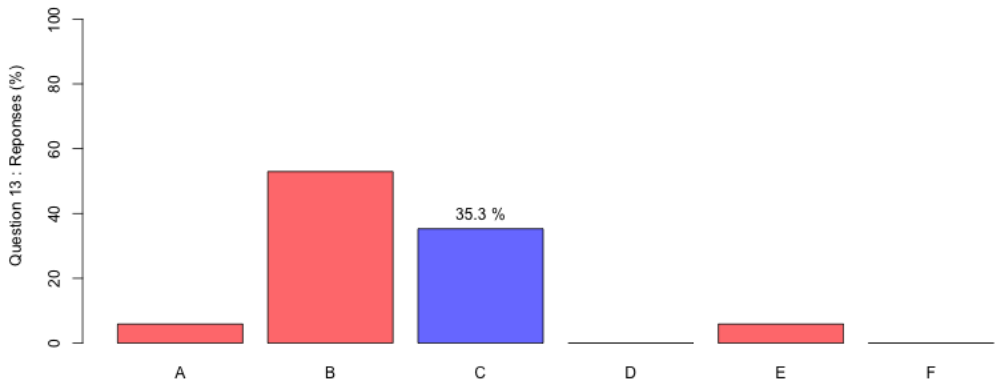
En utilisant la notation des GLM,

$$g(\widehat{p}) = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot 0 + \widehat{\beta}_2 \cdot 0 + \widehat{\beta}_3 \cdot 1 + \widehat{\beta}_4 \cdot \frac{225,000}{1000} = 0.4270 - 0.0920 + 0.0015 \cdot 225 = 0.6725$$

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. J'ai l'impression que le nombre élevé de B pourrait venir d'un oubli de l'utilisation de la fonction de lien inverse, pour revenir sur une probabilité...



14 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

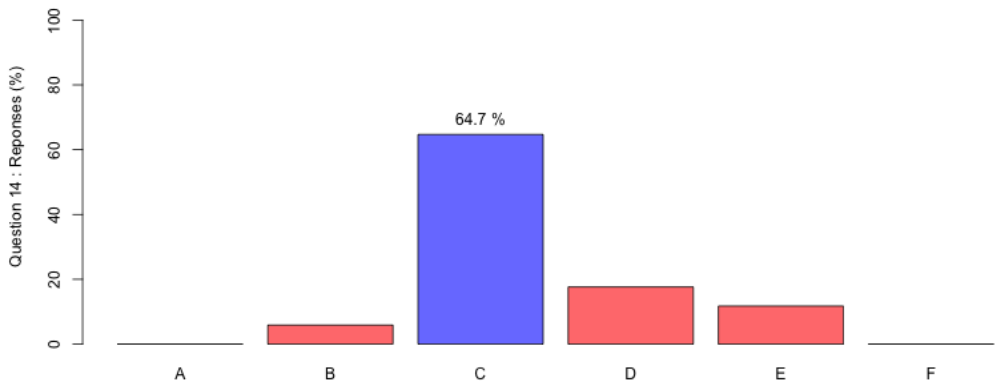
B) transformer la prévision : toute prévision inférieure à 0 devient 0, et toute prévision supérieure à 1 devient 1

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. $p \mapsto \log(p/(1-p))$. Autrment dit, on transforme 0 en $-\infty$ et 1 en $+\infty$. On retient donc C.



15 On nous donne les informations suivante : la probabilité de renouvellement $p(x)$ est modélisée par une régression logistique, sur la constante et une unique variable explicative (X). De plus $\widehat{\beta}_0 = 5$ et $\widehat{\beta}_1 = -0.65$. Calculez la cote ($\mathbb{P}[Y = 1|X = x]/\mathbb{P}[Y = 0|X = x]$) pour $x = 5$

A) moins de 2

B) entre 2 et 4

C) entre 4 et 6

D) entre 6 et 8

E) plus de 8

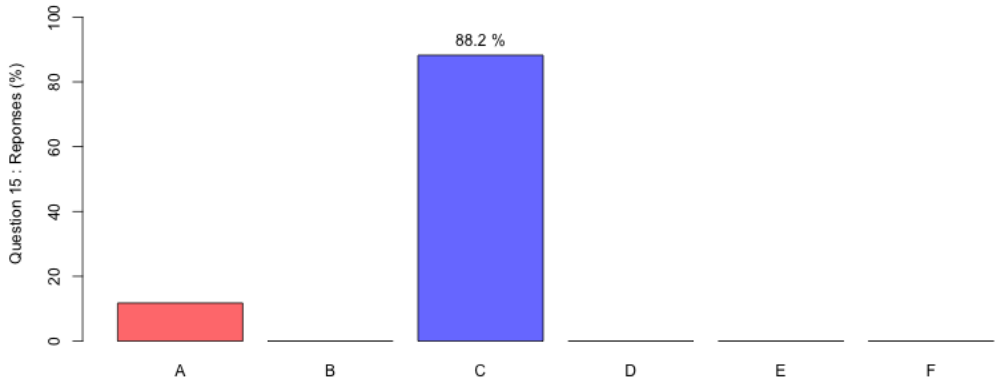
Sujet CAS-MAS-I du printemps 2018. On (en fait ce n'était pas dans l'énoncé de la CAS) nous rappelle que la cote est $\mathbb{P}[Y=1|X=x]/\mathbb{P}[Y=0|X=x]$. Or on a vu en cours que le modèle logististique revient à supposer que le logarithme de la cote est une combinaison linéaire des variables explicatives. Autrement dit

$$\log \frac{\mathbb{P}[Y=1|X=x]}{\mathbb{P}[Y=0|X=x]} = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot x \text{ soit } \underbrace{\frac{\mathbb{P}[Y=1|X=x]}{\mathbb{P}[Y=0|X=x]}}_{\text{cote}} = \exp\left[\widehat{\beta}_0 + \widehat{\beta}_1 \cdot x\right]$$

or ici $x=5$, donc la cote est

$$\exp\left[\widehat{\beta}_0 + \widehat{\beta}_1 \cdot x\right] = \exp\left[5 - 0.65 \cdot 5\right] = \exp[1.75] \sim 5.754$$

ce qui correspond à la réponse C.



16 On estime un modèle de Poisson pour une variable Y avec une constante et une unique variable explicative X, avec n observations. On suppose que $\log \mathbb{E}[Y|X=x] = \beta_0 + \beta_1 x$. On note $\widehat{\varepsilon}_i$ le résidu brut associé à la $i$ème observation. On a les trois affirmations suivantes

1. $\sum_{i=1}^{n} \widehat{\varepsilon}_i = 0$
2. $\sum_{i=1}^{n} x_i \widehat{\varepsilon}_i = 0$
3. La variance estimée de $\widehat{\beta}_1$ est $S_{2,2}$ où $S = \sum_{i=1}^{n} \widehat{\mu}_i \boldsymbol{x}_i \boldsymbol{x}_i^{\mathsf{T}}$

Lesquelles sont vraies

A) aucune

B) 1 et 2 seulement

C) 1 et 3 seulement

D) 2 et 3 seulement

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

Exercice ACTEC pour l'examen SRM (mai 2019). On avait vu que, pour une régression GLM avec le lien canonique, les conditions du premier ordre s'écrivaient

$$\sum_{i=1}^{n} \underbrace{(y_i - \widehat{\mu}_i)}_{=\widehat{\varepsilon}_i} \cdot \boldsymbol{x}_i = \mathbf{0}, \text{ ou encore } \sum_{i=1}^{n} \widehat{\varepsilon}_i \cdot \begin{pmatrix} 1 \\ x_i \end{pmatrix} = \sum_{i=1}^{n} \begin{pmatrix} \widehat{\varepsilon}_i \\ \widehat{\varepsilon}_i \cdot x_i \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

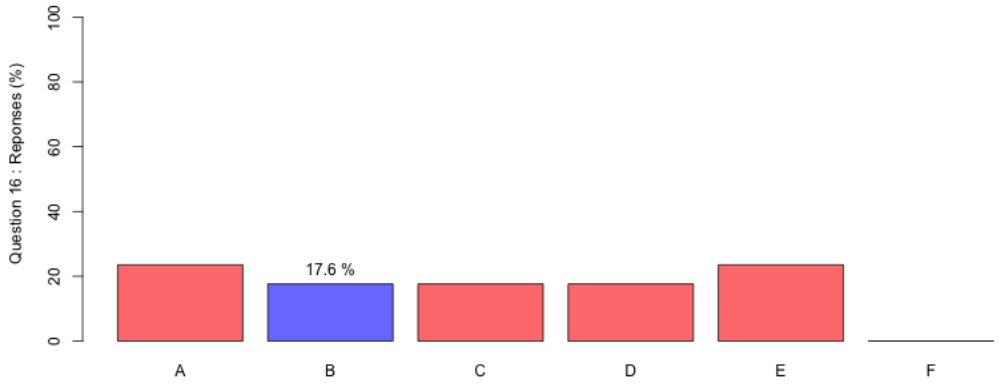
Les deux conditions sont alors

$$\sum_{i=1}^{n} \widehat{\varepsilon}_i = 0 \text{ et } \sum_{i=1}^{n} \widehat{\varepsilon}_i \cdot x_i = 0$$

qui correspondent aux affirmations 1 et 2, respectivement. Pour la troisième affirmation, $S = \sum_{i=1}^{n} \widehat{\mu}_i \boldsymbol{x}_i \boldsymbol{x}_i^{\mathsf{T}}$ est obtenue par la relation

$$S = \left. \frac{\partial^2 \log \mathcal{L}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathsf{T}}} \right|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}} = \sum_{i=1}^{n} \widehat{\mu}_i \boldsymbol{x}_i \boldsymbol{x}_i^{\mathsf{T}}$$

qui correspond alors à l'information de Fisher, $I(\widehat{\boldsymbol{\beta}})$. Or on sait que pour l'estimateur du maximum de vraisemblance, si n est suffisement grand, $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}$suit une loi normale $\mathcal{N}(\mathbf{0}, I(\widehat{\boldsymbol{\beta}})^{-1})$, ou dit autrement, S est l'inverse de la matrice de variance ! L'affirmation est fausse ! (pour s'en convaincre, on notera que les termes de S vont devenir infiniment grand quand n devient très grand, or la matrice de variance d'un estimateur a souvent tendance à décroitre ave n...). Donc 3 est fausse. Il fallait donc retenir la réponse B. La question était particulièrement vicieuse... mais c'était un exercice de l'examen SRM



17 On dispose de la sortie de régression suivante

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

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    5.421      0.228
GenderF       -0.557      0.217
Age            0.107      0.002
```

Calculez la prévision pour y pour une femme de 22 ans

A) moins de 1,000

B) entre 1,000 et 1,500

C) entre 1,500 et 2,000

D) entre 2,000 et 2,500

E) plus de 2,000

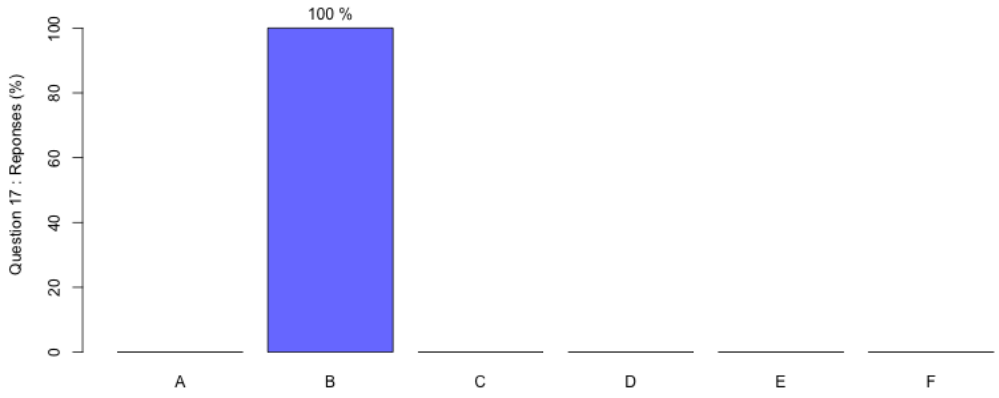
Il s'agissait d'une question de l'examen CAS-S du printemps 2017. Vue la sortie, l'âge semble être une variable continue, donc

$$\log \widehat{\mu} = \widehat{\beta}_0 + \widehat{\beta}_1 \mathbf{1}(\text{genre} = F) + \widehat{\beta}_2 \cdot x$$

où x désigne l'âge. Donc pour une femme ($\mathbf{1}(\text{genre} = F) = 1$) de 22 ans ($x = 22$), on obtient

$$\widehat{\mu} = \exp\left(5.421 - 0.557 + 0.107 \cdot 22\right) = \exp(7.218) \sim 1,363$$

ce qui correspond à la réponse B.



18 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) $1/\mu$

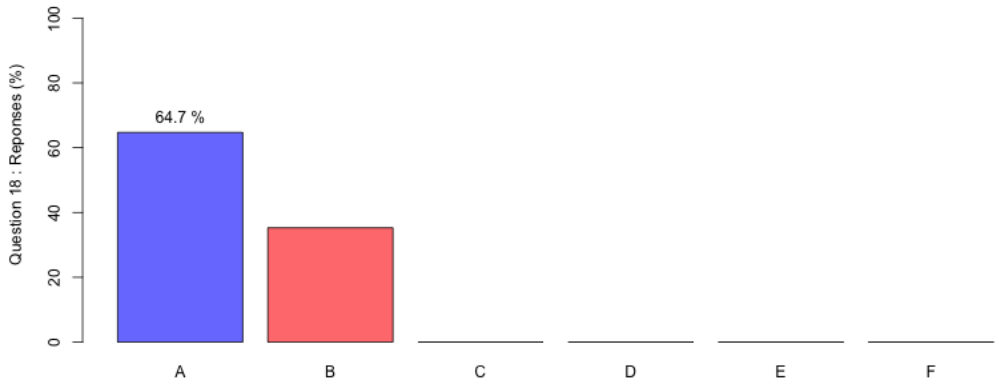
D) 1

E) e^{μ}

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

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

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



19 On souhaite retrouver les valeurs la sortie de régression suivante (en fait, on cherche juste `???`), avec une variable d'intérêt (`y`), une variable catégorielle (`gender`) prenant deux modalités (`Male` et `Female`), et une autre variable catégorielle (`territory`) prenant deux modalités (`Q` et `R`),

```
Call:
glm(formula = y ~ gender + territory + gender*territory, family = ********(link = "log"))

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                    ***
genderFemale                   ***
territoryR                     ***
genderFemale*territoryR        ???
```

On nous donne les prévisions suivantes :

		territory	
		Q	R
gender	Male	148	545
	Female	446	4,024

Déterminez $\widehat{\beta}_3$ associée à la variable croisée.

A) moins de 0.85

B) entre 0.85 et 0.95

C) entre 0.95 et 1.05

D) entre 1.05 et 1.15

E) plus de 1.15

Il s'agissait d'une question de l'examen CAS-MAS I du printemps 2018. Ici $x_1 = \mathbf{1}(\text{gender} = \text{'Female'})$ et $x_1 = \mathbf{1}(\text{territory} = \text{'R'})$ (la troisième variable étant juste le produit des deux, $x_3 = x_1 \cdot x_2$). On a en lien logarithmique, donc

$$\log \mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 \text{ soit } \log \widehat{\mu} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \widehat{\beta}_3 x_1 x_2$$

Autrement dit, si on reprend le tableau de prévision, on devrait avoir

		territory	
		Q	R
gender	Male	$\exp(\widehat{\beta}_0)$	$\exp(\widehat{\beta}_0+\widehat{\beta}_2)$
	Female	$\exp(\widehat{\beta}_0+\widehat{\beta}_1)$	$\exp(\widehat{\beta}_0+\widehat{\beta}_1+\widehat{\beta}_2+\widehat{\beta}_3)$

Si on prend les logarithmes dans les tableaux,

		territory	
		Q	R
gender	Male	$\widehat{\beta}_0$	$\widehat{\beta}_0+\widehat{\beta}_2$
	Female	$\widehat{\beta}_0+\widehat{\beta}_1$	$\widehat{\beta}_0+\widehat{\beta}_1+\widehat{\beta}_2+\widehat{\beta}_3$

$=$

		territory	
		Q	R
gender	Male	4.9972	6.30078
	Female	6.10032	8.30003

On reconnaît un système linéaire (relativement simple) de quatre équations à quatre inconnues. On nous demande (juste) $\widehat{\beta}_3$, or

$$\widehat{\beta}_3 = \underbrace{(\widehat{\beta}_0+\widehat{\beta}_1+\widehat{\beta}_2+\widehat{\beta}_3)}_{8.30003} - \underbrace{(\widehat{\beta}_0+\widehat{\beta}_1)}_{6.10032} - \big(\underbrace{(\widehat{\beta}_0+\widehat{\beta}_2)}_{6.30078} - \underbrace{\widehat{\beta}_0}_{4.9972}\big)$$

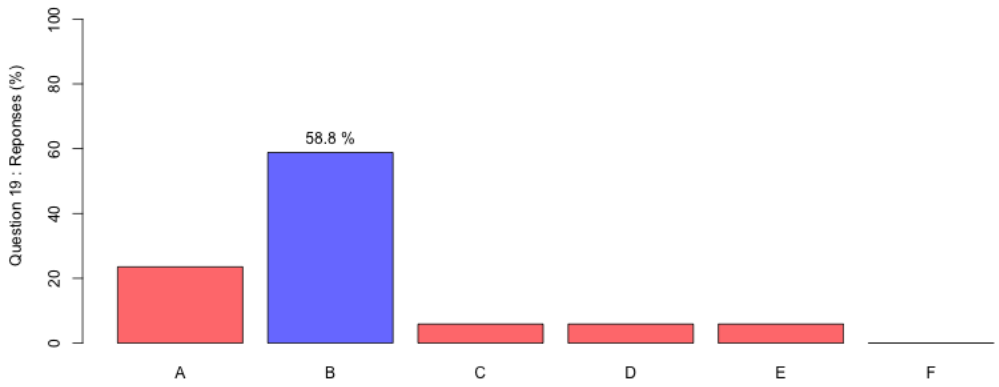
soit

$$\widehat{\beta}_3 = 8.30003 - 6.10032 - 6.30078 + 4.9972 \sim 0.8961$$

Si on veut une expression plus simple (et juste), on note que

$$\widehat{\beta}_3 = \log(4024) - \log(545) - \log(446) + \log(148) = \log\frac{4024\cdot 148}{545\cdot 446} = \log\frac{297776}{121535}$$

Bref, $\widehat{\beta}_3 \sim 0.8961$ ce qui correspond à la réponse B.



20 On nous donne le tableau de données suivant

i	x	y
1	34	2
2	38	1
3	45	0
4	25	3
5	21	3
total	163	9

On suppose que les variables Y_i sont, conditionnellement à X_i indépendantes, et suivent une loi de Poisson de moyenne $\mu_i = \beta x_i$.

Donnez l'écart type (asymptotique) de $\widehat{\beta}$ (estimé par maximum de vraisemblance)

A) moins de 0.015

B) entre 0.015 et 0.020

C) entre 0.020 et 0.025

D) entre 0.025 et 0.030

E) plus de 0.030

On a ici une question pour l'examen 4 de la SOA de l'automne 2004. On a ici 5 observations. Comme je l'ai dit en cours, on en fait pas de régression quand on a 5 observations ! Mais c'est typique des exemans de la SOA. On peut regarder, l'idée est de faire un simple calcul. En effet, la vraisemblance et log-vraisemblance sont ici

$$\mathcal{L} = \prod_{i=1}^{5} \frac{e^{-\beta x_i}(\beta x_i)^{y_i}}{y_i!} \text{ soit } \log \mathcal{L} = -\beta \sum_{i=1}^{5} x_i + \log \beta \sum_{i=1}^{5} y_i$$

à une constante près ! Aussi, l'équation normale (comme on a un seul paramètre, elle est unique)

$$\left.\frac{\partial \log \mathcal{L}}{\partial \beta}\right|_{\beta=\widehat{\beta}} = -\sum_{i=1}^{5} x_i + \frac{1}{\widehat{\beta}} \sum_{i=1}^{5} y_i = 0$$

de telle sorte que le maximum de la (log-)vraisemblance est atteint lorsque

$$\widehat{\beta} = \frac{\sum_{i=1}^{5} y_i}{\sum_{i=1}^{5} x_i} = \frac{9}{163}$$

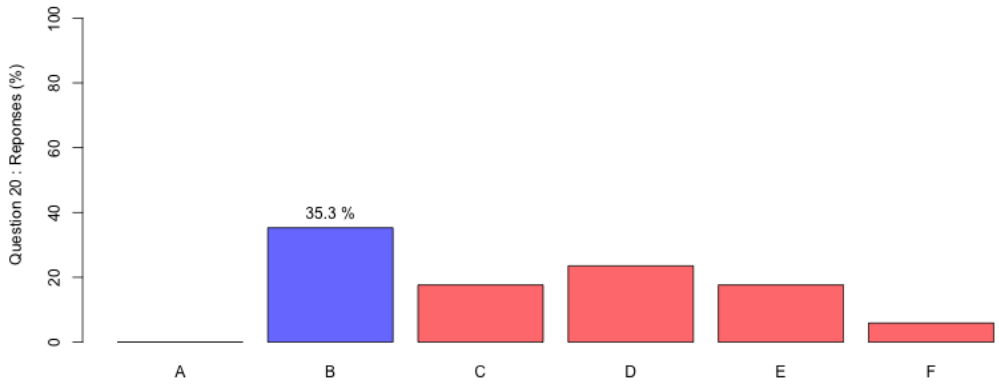
L'information de Fisher associée (ce n'est pas une matrice ici, comme on a un unique paramètre) est

$$I(\beta) = -\mathbb{E}\left[\frac{\partial^2 \log \mathcal{L}}{\partial \beta^2}\right] = -\mathbb{E}\left[\frac{\partial}{\partial \beta}\left(-\sum_{i=1}^{5} X_i + \frac{1}{\beta}\sum_{i=1}^{5} Y_i\right)\right] = -\mathbb{E}\left[\frac{-1}{\beta^2}\sum_{i=1}^{5} Y_i\right] = \frac{1}{\beta^2}\sum_{i=1}^{5}\mathbb{E}(Y_i)$$

donc

$$I(\beta) = \frac{1}{\beta^2}\sum_{i=1}^{5}\mathbb{E}(Y_i) = \frac{1}{\beta^2}\sum_{i=1}^{5}\beta x_i = \frac{1}{\beta}\sum_{i=1}^{5} x_i = \frac{163}{\beta}$$

et la version estimée est alors $I(\widehat{\beta}) = \dfrac{163}{\widehat{\beta}}$. L'écart-type estimé de $\widehat{\beta}$ est ici la racine carrée de l'inverse de $I(\widehat{\beta})$, soit $\dfrac{\widehat{\beta}}{163}$ soit numériquement $\dfrac{9}{163^2} = \dfrac{3}{163} \sim 0.0184$ qui est compris entre 0.015 et 0.020, ce qui correspond à la réponse B.



21 On dispose de la sortie de régression suivante

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

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)            2.100
zone1                  7.678
zone2                  4.227
zone3                  1.336
zone5                  1.734
classConvertible       1.200
classCoupe             1.300
classTruck             1.406
classMinivan           1.875
classUtility           2.000
AgeYouth               2.000
AgeOld                 1.800
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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

Calculer la variance de la charge sinistre pour une observation dans la zone 4, pour un véhicule de la classe Utility et un conducteur d'âge intermédiaire.

A) moins de 55

B) entre 55 et 60

C) entre 60 et 65

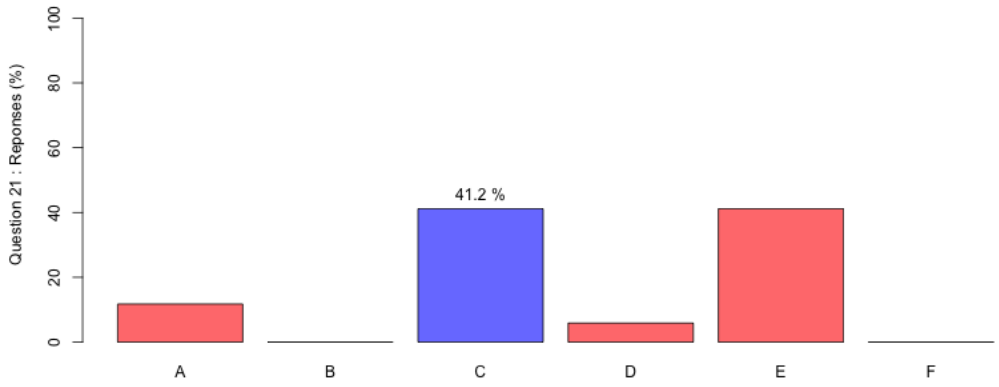
D) entre 65 et 70

E) plus de 70

C'est une question de l'examen CAS-S de printemps 2016. **pour information, pour l'âge, on avait 'youth' et 'old', intermédiaire est ici la modalité de référence** On a un lien log, ce qui signifie que

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

de telle sorte que $\widehat{\mu} = \exp[4.1] \sim 60.34$ qui correspond à la réponse C.



22 On utilise un modèle GLM-Tweedie pour modéliser la charge annuelle par police, avec deux variables explicatives : le genre (homme ou femme) et le lieu de résidence, ou localisation (urbain ou rural, en ville ou à la campagne). On retient un modèle Tweedie de paramètre $p = 1.5$ avec $\phi = 1$. Un lien log est utilisé.

```
Call:
glm(formula = y ~ zone + class + age, family = tweedie(var.power=1.5,link.power=0))

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)            2.32
LocationRural         -0.64
GenderMale             0.76
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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

Calculer la variance de la charge anuuelle pour un homme habitant à la campagne.

A) moins de 25

B) entre 25 et 100

C) entre 100 et 175

D) entre 175 et 250

E) lus de 250

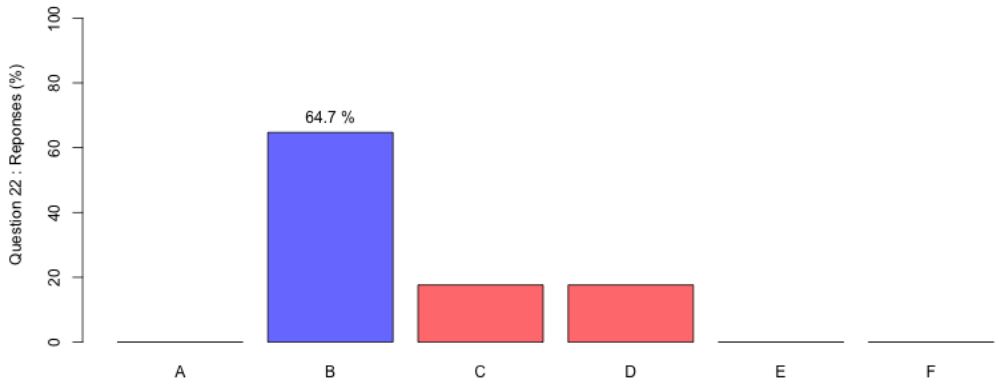
On utilise ici un lien logarithmique, donc, comme on veut une prévision pour un homme ($\mathbf{1}(\text{gender} = \text{'Male'}) = 1$) et qui vit à la campagne ($\mathbf{1}(\text{location} = \text{'Rural'}) = 1$)

$$\log \widehat{\mu} = \widehat{\beta}_0 + \widehat{\beta}_1 + \widehat{\beta}_2 = 2.32 - 0.64 + 0.76 = 2.44 \text{ soit } \widehat{\mu} = \exp[2.44]$$

On a un modèle Tweedie, correspondant à une fonction variance de type puissance, $\mathsf{V}(\mu) = 1 \cdot \mu^{1.5}$, donc ici, la variance (estimée) est

$$\widehat{\mu}^{1.5} = \big(\exp[2.44]\big)^{1.5} = \exp[3.66] \sim 38.86$$

ce qui correspond à la réponse B.



23 Une entreprise envisage de changer sa tarification, en rajoutant quelques variables explicatives. On dispose des statistiques suivantes :

	ancien modèle	nouveau modèle
log-vraisemblance	-750	-737.5
déviance	500	475
nombre de paramètres	10	15
nombre d'observations	1,000,000	1,000,000

On entend les trois arguments suivants

1. Le nouveau modèle est meilleur sur la base du AIC
2. Le nouveau modèle est meilleur sur la base du BIC
3. Le BIC est préférable au AIC quand on a beaucoup d'observations (comme ici)

Quelles sont les affirmations correctes ?

A) 1 seulement

B) 2 seulement

C) 3 seulement

D) 1, 2 et 3

E) aucune des propositions A, B, C ou D

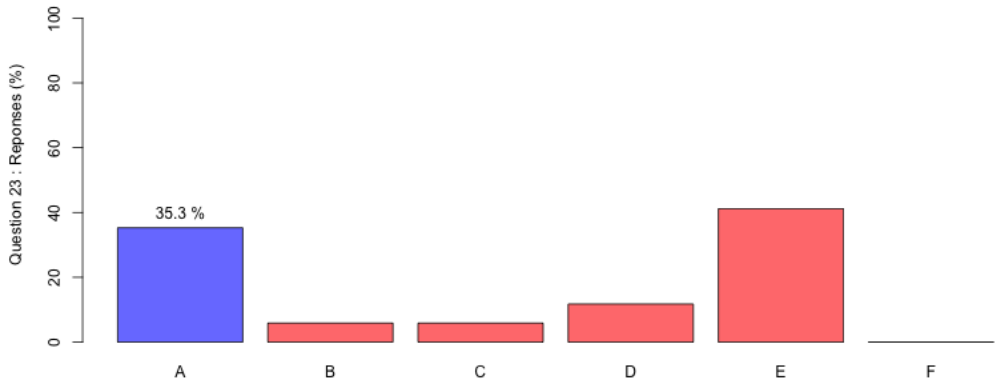
C'était une question de l'examen 8 de la CAS de l'automne 2016. Pour rappel, $AIC = 2k - 2\log\mathcal{L}(\widehat{\beta})$ et $BIC = k\log n - 2\log\mathcal{L}(\widehat{\beta})$, alors que la déviance s'écrit $D = 2\big(\log\mathcal{L}_{\text{saturée}} - \log\mathcal{L}(\widehat{\beta})\big)$. Regardons les diverses affirmations.

L'affirmation 1 est juste. En effet, le AIC vaut $2{\cdot}10 - 2{\cdot}(-750) = 1,520$ pour l'ancien modèle et $2{\cdot}15 - 2{\cdot}(-737.5) = 1,505$ pour le nouveau modèle, qui est plus faible. Or plus faible signifie *meilleur*.

L'affirmation 2 est fausse. En effet, le BIC vaut $\log(1,000,000)\cdot 10 - 2\cdot(-750) = 1,638.16$ pour l'ancien modèle et $\log(1,000,000)\cdot 15 - 2\cdot(-737.5) = 1,682.16$ pour le nouveau modèle, qui est plus élevé. Or (là aussi), c'est plus faible qui signifie *meilleur*.

L'affirmation 3 est fausse. Dans la forme originale, l'affirmation "*BIC is more reliable*". Avec le AIC, on a une pénalisation en $2k$ (où k est le nombre de variables explicatives), alors que pour la pénalisation est en $k\log(n)$. Ici, on a 1 million d'observations, donc une pénalisation en $13.815\cdot k$: rajouter une variable avec le BIC donne la même pénalisation qu'en rajouter 7 avec le AIC ! Dit autrement, le BIC va nous pousser à rejeter des variables explicatives, même si elle ont un fort pouvoir prédictif. Donc on aura tendance à ne pas utiliser le BIC en (très) grande dimension (comme évoqué en cours).

Si on résume, seule 1 est juste, autrement dit, la bonne réponse est A.



24 Quelle affirmation parmi les suivantes est vraie

A) le modèle avec le plus grand AIC est le meilleur, quand on compare des modèles

B) le modèle avec le plus grand BIC est le meilleur, quand on compare des modèles

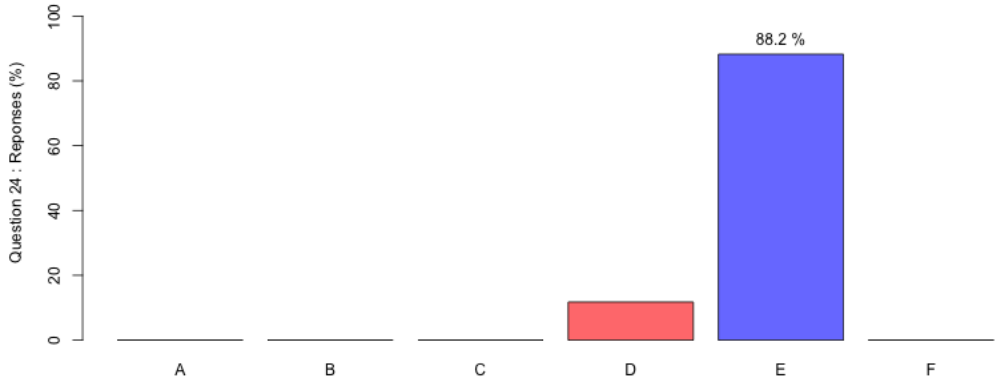
C) le modèle avec la plus grande déviance est le meilleur

D) toutes choses étant égales par ailleurs, si le nombre d'observations est grand, le AIC pénalise davantage les variables que l'on rajoute que le BIC

E) toutes choses étant égales par ailleurs, si le nombre d'observations est grand, le BIC pénalise davantage les variables que l'on rajoute que le AIC

Pour les trois premiers points, il faut chercher les trois plus *petites* valeurs. Pour rappel, $AIC = 2k - 2\log\mathcal{L}(\widehat{\beta})$ et $BIC = k\log n - 2\log\mathcal{L}(\widehat{\beta})$: dans les deux cas, maximiser la log-vraisemblance signifier minimiser les critères

AIC ou BIC. Pour la déviance, on a $D = 2\big(\log \mathcal{L}_{\text{saturée}} - \log \mathcal{L}(\widehat{\beta})\big)$, et là aussi, maximiser la log-vraisemblance signifier minimiser la déviance (on peut aussi se souvenir que la déviance est une distance au modèle parfait... on la veut la plus petite possible). Reste à choisir entre D et E. Comme on vient de le revoir, avec le AIC, on a une pénalisation en $2k$ (où k est le nombre de variables explicatives), alors que pour le BIC la pénalisation est en $k\log(n)$. Donc "*BIC pénalise davantage les variables que l'on rajoute que le AIC*" (je reprends les termes de l'affirmation E) signifie $k\log n > 2k$ ce qui est obtenu dès lors que $n > e^2 = 7.38$. Autrement dit, si on a plus de 7 observation, oui, l'indice BIC péalise davantage le rajout de variable (on le verra concrètement dans la question 27). Bref, l'affirmation E est valide.



25 On dispose de la sortie suivante

i	x	y	$\widehat{\mu}$
1	-1	2	?
2	0	?	3.838434
3	+1	?	7.080783

On suppose qu'une régression de Poison a été estimée avec un lien logarithmique, $\log(\mu) = \beta_0 + \beta_1 x$.

Donnez le résidu de Pearson pour la première observation, $\widehat{\varepsilon}_1$

A) moins de -0.10

B) entre -0.10 et -0.05

C) entre -0.05 et 0

D) entre 0 et +0.05

E) plus de +0.05

Avec les deux valeurs donc on dispose, on devrait pouvoir retrouver les estimateurs $\widehat{\beta}_0$ et $\widehat{\beta}_1$:

$$\log \widehat{\mu}_2 = \log 3.838434 = 1.345 = \widehat{\beta}_0 + \widehat{\beta}_1 x_2 = \widehat{\beta}_0$$

et

$$\log \widehat{\mu}_3 = \log 7.080783 = 1.957 = \widehat{\beta}_0 + \widehat{\beta}_1 x_3 = \widehat{\beta}_0 + \widehat{\beta}_1$$

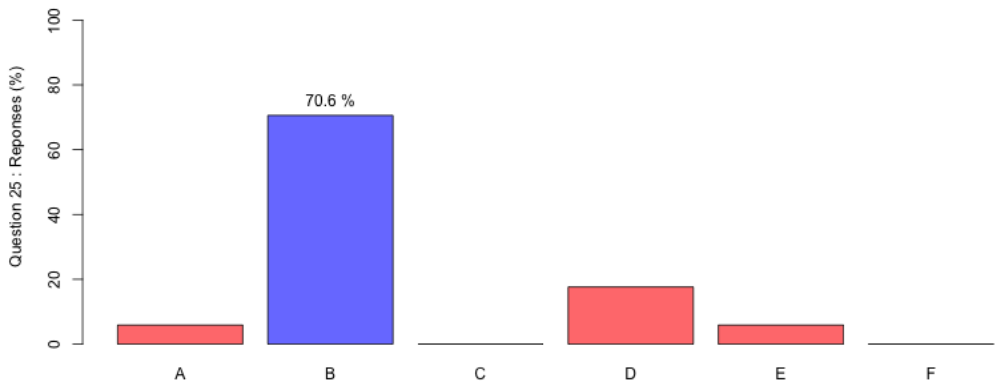
De la première équation, on déduit $\widehat{\beta}_0 = 1.345$ et de la seconde $\widehat{\beta}_1 = 0.612$. On peut alors utiliser ces deux valeurs pour avoir la prévision $\widehat{\mu}_1$,

$$\widehat{\mu}_1 = \exp[\widehat{\beta}_0 + \widehat{\beta}_1 x_1] = \exp[\widehat{\beta}_0 - \widehat{\beta}_1] = \exp[1.345 - 0.612] \sim 2.0813$$

On le résidu de Pearson pour un modèle de Poisson est

$$\widehat{\varepsilon}_1 = \frac{y_1 - \widehat{\mu}_1}{\sqrt{\widehat{\mu}_1}} = \frac{2 - 2.0813}{\sqrt{2.0813}} \sim -0.0563$$

ce qui correspond à la réponse B.



26 On dispose de la sortie de régression suivante

```
Call:
glm(formula = y ~ age + zone + vehicle, family = binomial(link = "logit"))

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)          ?
age1             0.288
age2             0.064
zoneA           -0.036
zoneB            0.053
vehicleBus       1.136
vehicleCar      -0.371
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

La probabilité qu'un conducteur du groupe d'âge `2`, de la zone `C` et conduisant un véhicle de type `Car` ait un accident est de 22%. Calculer la cote d'un conducteur du groupe d'âge `3`, de la zone `C` et conduisant un véhicle de type `Other`, associée à la survenance d'un accident.

A) moins de 0.2

B) entre 0.2 et 0.25

C) entre 0.25 et 0.30

D) entre 0.30 et 0.35

E) plus de 0.35

Il s'agissait d'une question de l'examen S de la CAS, du printemps 2016. On va utiliser la probabilité qu'on nous donne

$$\frac{e^{\boldsymbol{x}^\intercal\widehat{\boldsymbol{\beta}}}}{1+e^{\boldsymbol{x}^\intercal\widehat{\boldsymbol{\beta}}}}=\frac{1}{1+e^{-\boldsymbol{x}^\intercal\widehat{\boldsymbol{\beta}}}}=22\% \text{ ou } \log\frac{22\%}{1-22\%}\sim -1.2657=\boldsymbol{x}^\intercal\widehat{\boldsymbol{\beta}}$$

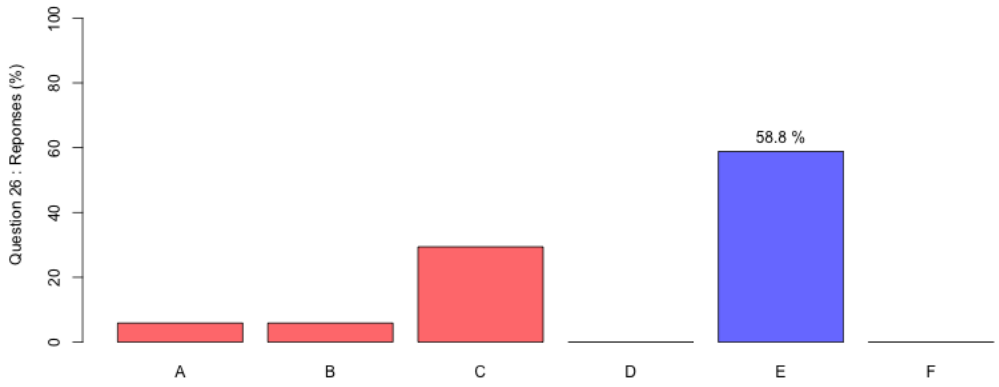
avec ici

$$\boldsymbol{x}^\intercal\widehat{\boldsymbol{\beta}}=x+0.064-0.371 \text{ soit } x=-1.2657-0.064+0.371\sim -0.9587$$

Maintenant, on peut faire nos calculs : la cote estimée pour ce nouvel individu est

$$\frac{\widehat{p}}{1-\widehat{p}}=e^{\boldsymbol{x}^\intercal\widehat{\boldsymbol{\beta}}}=\exp\big(-0.9587\big)\sim 0.3834$$

(on a juste la constante car cet individu correspond à toutes les modalités de référence). Ce qui correspond à la réponse E.



27 Un actuaire veut estimer la probabilité qu'un sinistre survienne, à l'aide d'un modèle logistique. Il a un historique de $n = 1000$ observations, incluant

- y la survenance (ou non) d'un sinistre
- x_1 le coût de la maison (en '000 \$)
- x_2 l'âge de la maison

Cinq modèles sont estimés,

	modèle	scaled deviance
A)	$y \sim 1 + x_1$	1085.0
B)	$y \sim 1 + x_1 + x_2$	1084.8
C)	$y \sim 1 + x_1 + (x_1 \cdot x_2)$	1083.0
D)	$y \sim 1 + x_1 + (x_1^2) + (x_1^3)$	1081.9
E)	$y \sim 1 + x_1 + (x_1^2) + (x_1^3) + (x_1^4)$	1081.6

En utilisant comme critère le BIC, quel modèle doit-on choisir ?

A) modèle (A)

B) modèle (B)

C) modèle (C)

D) modèle (D)

E) modèle (E)

C'était une question du Sample Exem pour l'examen S de la CAS. Rappelons que le BIC, le critère Bayésien ou de Schwarz, est défini par

$$BIC = -2\log \mathcal{L}(\widehat{\boldsymbol{\mu}}) + k\log(n)$$

où la notation $\log \mathcal{L}(\widehat{\boldsymbol{\mu}})$ montre qu'on calcule la log-vraisemblance au maximum, c'est à dire en $\widehat{\boldsymbol{\mu}}$. Mais ce n'est pas ce qu'on nous donne ici ! On en avait parlé en cours... L'astuce est d'écrire

$$BIC = -2\log \mathcal{L}(\boldsymbol{y}) + \underbrace{2\big(\log \mathcal{L}(\boldsymbol{y}) - \log \mathcal{L}(\widehat{\boldsymbol{\mu}})\big)}_{=\text{scaled deviance}} + k\log(n)$$

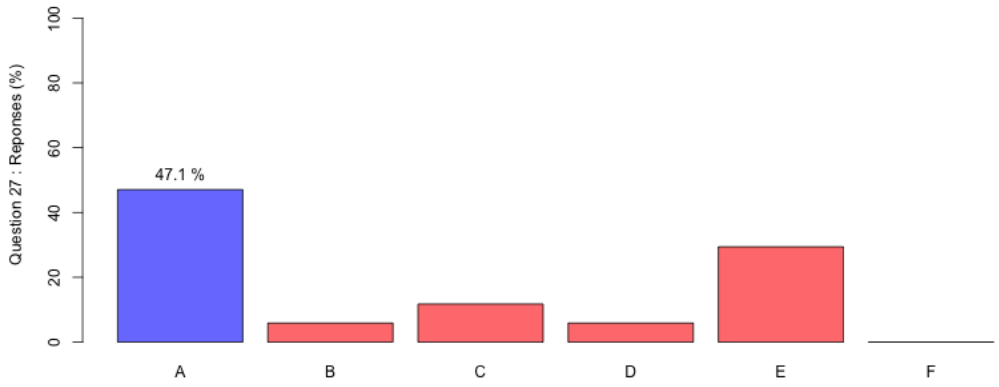
On pourrait avoir l'impression qu'on n'a pas trop avancé... sauf que le premier terme est le même, quel que soit le modèle estimé, puisqu'on ne change pas la loi et le lien ! On peut écrire pour simplfier

$$BIC = -2\log \mathcal{L}(\boldsymbol{y}) + \underbrace{\text{scaled deviance} + k\log(n)}_{=BIC'}$$

où BIC' est juste une translation (avec la même valeur) du BIC. Bref, comparer les BIC, c'est équivalent à comparer les BIC'. On cherche le modèle qui a la plus petit BIC' ! Or ici

	modèle	scaled deviance$+k\log n$
A)	$y \sim 1 + x_1$	$1085.0 + 2\log(1000) \sim \mathbf{1098.82}$
B)	$y \sim 1 + x_1 + x_2$	$1084.8 + 3\log(1000) \sim 1105.52$
C)	$y \sim 1 + x_1 + (x_1 \cdot x_2)$	$1083.0 + 3\log(1000) \sim 1103.72$
D)	$y \sim 1 + x_1 + (x_1^2) + (x_1^3)$	$1081.9 + 4\log(1000) \sim 1109.53$
E)	$y \sim 1 + x_1 + (x_1^2) + (x_1^3) + (x_1^4)$	$1081.6 + 5\log(1000) \sim 1116.14$

On cherche ici le plus petit BIC', on va donc retenir le modèle A. C'est assez cohérent avec le fait que le critère BIC donne une pénalité très importante (en $\log n$)



28 On dispose d'un premier modèle GLM, et regarde ce qui se passe en rajoutant une variable catégorielle.

- le premier modèle avait 15 paramètres
- en rajoutant cette variable la *scale deviance* varie de -53
- en rajoutant cette variable le AIC varie de -47
- en rajoutant cette variable le BIC varie de -32

Combien d'observations y-a-t-il dans la base de données

A) moins de 200

B) entre 200 et 500

C) entre 500 et 700

D) entre 700 et 1,000

E) plus de 1,000

On en avait parlé dans la question précédante :

$$AIC = -2\log\mathcal{L}(\boldsymbol{y}) + \underbrace{2\big(\log\mathcal{L}(\boldsymbol{y}) - \log\mathcal{L}(\widehat{\boldsymbol{\mu}})\big)}_{=\text{scaled deviance}} + 2k = -2\log\mathcal{L}(\boldsymbol{y}) + \underbrace{\text{scaled deviance} + 2k}$$

et

$$BIC = -2\log\mathcal{L}(\boldsymbol{y}) + \underbrace{2\big(\log\mathcal{L}(\boldsymbol{y}) - \log\mathcal{L}(\widehat{\boldsymbol{\mu}})\big)}_{=\text{scaled deviance}} + k\log(n) = -2\log\mathcal{L}(\boldsymbol{y}) + \underbrace{\text{scaled deviance} + k\log(n)}$$

Si on va un peu plus loin, on parle de changement de modèle seulement en terme de variables explicatives, et pas de forme (même loi et même fonction lien), donc partout on a le même $-2\log\mathcal{L}(\boldsymbol{y})$. Aussi,

$$\Delta AIC = \Delta\text{scaled deviance} + 2\Delta k$$

et

$$\Delta BIC = \Delta\text{scaled deviance} + \Delta k \cdot \log(n)$$

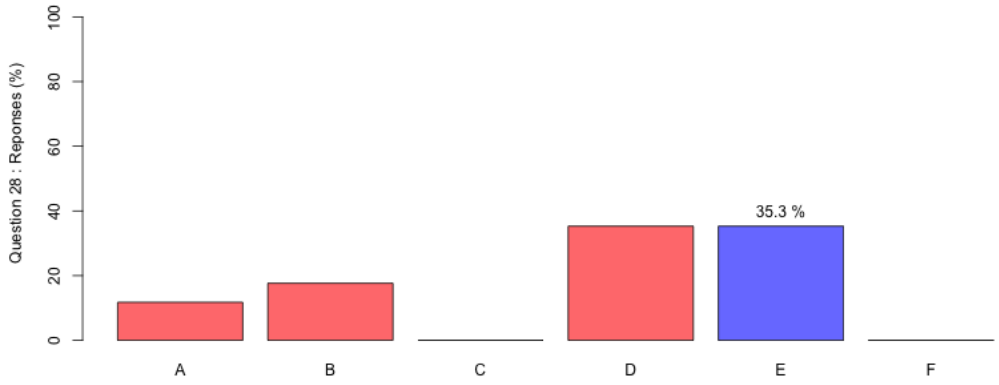
Le premier est intéressant car on en déduit que

$$\Delta AIC = (-47) = \Delta\text{scaled deviance} + 2\Delta k = (-53) + 2\Delta k \text{ soit } \Delta k = 3$$

Le nombre de paramètres ajouté est alors 3 (ce qui correspondrait - a priori - à un modèle avec 4 variables explicatives). Mais peu importe en fait. Si on remet $\Delta k = 3$ dans la seconde équation, on obtient

$$\Delta BIC = (-32) = \Delta\text{scaled deviance} + \Delta k \cdot \log(n) = (-53) + 3\log(n),$$

donc $n = \exp([53-32]/3) = \exp[7] = 1096.63$ ce qui correspond à la réponse E.



29 On nous donne la sortie de régression suivante, pour modéliser le défaut ($y = 1$) d'un emprunteur, à l'aide de deux variables explicatives

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

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -0.1286
x1                0.3321
x2               -0.4607
```

Prédire la probabilité d'avoir un défaut si $x_1 = x_2 = -1$ (on retiendra la valeur la plus proche)

A) 0.27

B) 0.36

C) 0.50

D) 0.63

E) 0.73

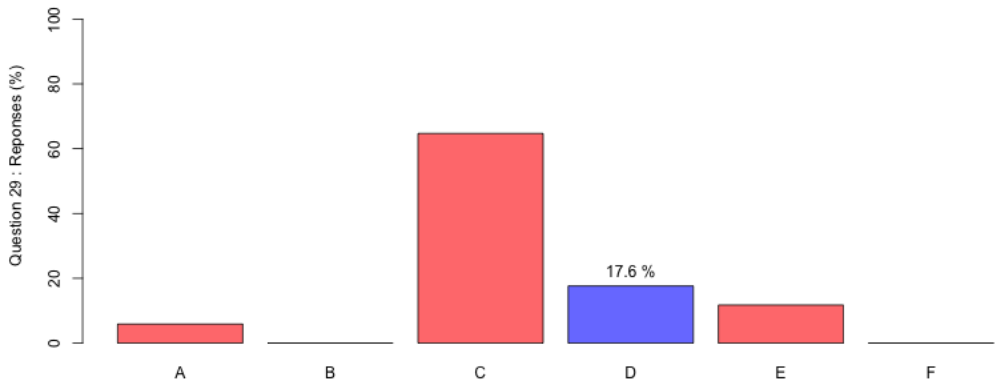
Exercice ACTEX. En remplçant, le score est

$$\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot (-1) + \widehat{\beta}_2 \cdot (-1) = -0.1286 - 0.3321 + 0.4607 = 0$$

La fonction de lien est ici la fonction loglog, i.e. $g(p) = \log(-\log(1-p))$. Pour rappel (cf cours), on ne peut pas utiliser $\log(\log(p))$ car comme $p < 1$, $\log(p) < 0$. On aurait pu avoir $\log(-\log(p))$, mais classiqueemnt, on veut que la fonction de lien soit croissante ! Donc on utilise $\log(-\log(1-p))$. Ici la prévision est alors $g^{-1}(0)$ soit ici

$$1 - \exp(-\exp(-\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2)) = 1 - \exp[-1] = 0.6321$$

. On a donc ici la réponse D. J'ai l'impression que le fort taux de C vient d'une mauvaise fonction de lien (avec un lien probit ou logit, on aurait obtenu 0.5).



30 On utilise une régression de Poisson pour modéliser le nombre de décès par diabète. On obtient la sortie suivante

```
Call:
glm(formula = y ~ genre + age + I(age^2) + I(age*(genre=="F")), family = poisson(link = "log"))

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)        -15.0                     <.0001
genreF             -1.20                     <.0001
age                0.150                     <.0001
age^2              0.004                     <.0001
age*(genre=="F")   0.012                     <.0001
```

Calculer le nombre de décès attendu pour une population de 100,000 femmes de 25 ans.

A) moins de 3

B) entre 3 et 5

C) entre 5 et 7

D) entre 7 et 9

E) plus de 9

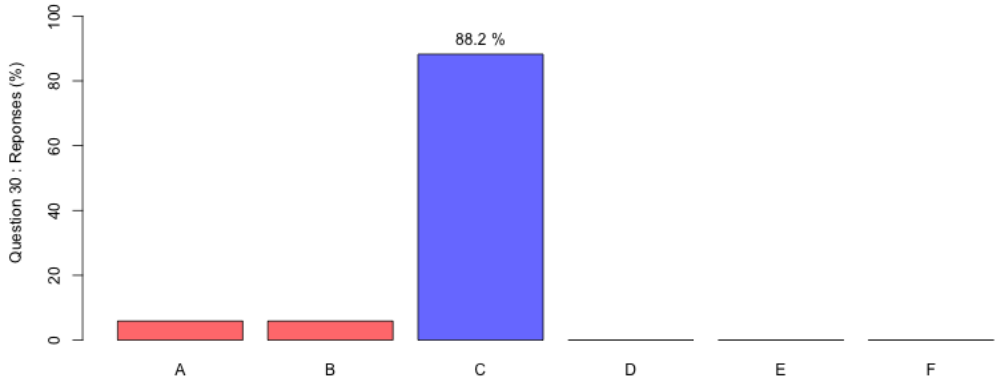
[CAS Examen S Printemps 2016] Dit de manière un peu rapide, pour une femme, le nombre moyen de décès attendu est

$$\widehat{\mu} = \exp[-15 - 1.2 + 0.15 \cdot 25 + 0.004 \cdot 25^2 + 0.012 \cdot 25] = \exp[-9.65] = 0.0000644256$$

Aussi, si on suppose que la survie des femmes est indépendantes, le nombre de décès parmi 100,000 femmes de 25 ans suit une loi de Poisson $\mathcal{P}(100,000 \cdot \widehat{\mu})$, de moyenne

$$100,000 \cdot \widehat{\mu} = 6.44$$

ce qui correspond à la réponse C.



31 (*bonus*) Sur les 30 questions précédantes, combien de bonnes réponses pensez vous avoir ?

Je suis toujours surpris que plusieurs étudiant(e)s ne répondent pas à cette question... Pourtant, rationnellement, il est intéressant de répondre, même n'importe quoi... Personnellement, cette question me permet aussi de mieux savoir s'il y a un soucis, qui pensait avoir réussi (ou raté) l'examen, et comparer à la note réelement obtenue...