STT5100 - Automne 2018 - Examen Final

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, toute sortie avant midi est autorisée, et définitive.

Dans les feuilles qui suivent, il y a 30 questions relatives au cours sur les modèles linéaires généralisés (incluant la régression logistique, et Poisson). Pour chaque question, quatre ou cinq réponses sont proposées, une seule est valide, et vous ne devez en retenir qu'une (au maximum),

- 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. 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 réponse est à la toute fin du document : merci de décrocher la feuille et de ne rendre que cette dernière, après avoir indiqué votre code permanent en haut. La-dite feuille contient une page avec les réponses à cocher, et au dos, 2 *graphiques* à compléter (pour les questions 5 et 6).

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.

Formulaire

Pour la loi normale, centrée et réduite ou une loi de Student, on utilisera 1.96 comme valeur du quantile à 97.5%, et 1.64 pour le quantile à 95%.

On notera $x \mapsto \mathbf{1}_A(x)$ la fonction indicatrice vérifiant $\mathbf{1}_A(x) = 0$ si $x \notin A$ et $\mathbf{1}_A(x) = 1$ si $x \in A$. Par extention, si $A = \{a\}$, on notera $x \mapsto \mathbf{1}_a(x)$ la fonction qui vérifie $\mathbf{1}_a(x) = 0$ si $x \neq a$ et $\mathbf{1}_a(x) = 1$ si $x = a$. Dans ce dernier cas, on pourra aussi utiliser la notation $\mathbf{1}(x = a) = 0$.

1 Comment est estimé un modèle logistique

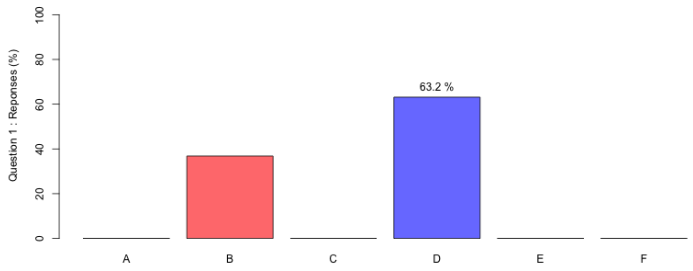
A) par la méthode des moindres carrés sur les y_i

B) par la méthode des moindres carrés sur une transformation logistique de observations y_i

C) par la méthode des moments sur le logarithme des y_i

D) par maximum de vraisemblance car les y_i suivent une loi binomiale conditionnellement aux $\boldsymbol{x}_i$

C'est le cours, je ne pensais pas revenir dessus ici... Que dire ? à part répéter ce que j'ai dit 20 fois dans la seconde partie du cours : (1) dans les GLM (et donc *a fortiori* pour la régression Bernoulli), on ne transforme pas les variables, on suppose que l'espérance est une fonction (que l'on se donne, la fonction lien) d'une combinaison linéaire des variables explicatives (en un sens, on transforme l'espérance) ! (2) la réponse B n'a pas de sens : en effet, la transformation logistique, c'est $\log[y/(1-y)]$, mais comme les observations sont dans l'ensemble $\{0,1\}$ - qui n'est composé que de deux éléments, $\{0\}$ et $\{1\}$, cf cours d'introduction aux probabilités - les valeurs que prend $y \mapsto \log[y/(1-y)]$ sont $\log(0) = -\infty$ et $\log(1/0) = +\infty$. Bref, on aura du mal à faire un modèle par moindres carrés sur les variables transformées !



2 Dans un modèle logistique, si $\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2$ vaut 0, quelle sera la prévision pour $\mathbb{P}[Y = 1|x_1, x_2]$?

A) 0.25

B) 0.5

C) 0.75

D) 1

E) aucune des réponses proposées

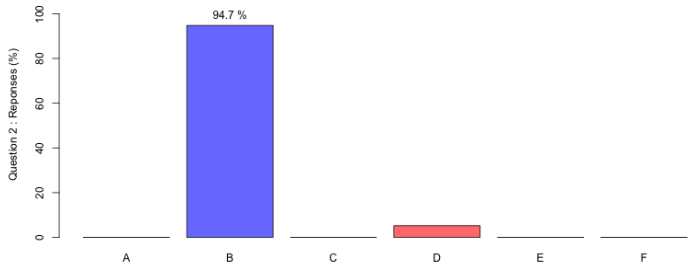
Dans un modèle logistique,

$$\mathbb{P}[Y = 1|x_1, x_2] = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}$$

donc si $\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2$ vaut 0,

$$\mathbb{P}[Y = 1|x_1, x_2] = \frac{\exp(0)}{1 + \exp(0)} = \frac{1}{1+1} = \frac{1}{2}$$

qui est la réponse B.



3 On a obtenu la sortie de régression suivante

```
Call:
glm(formula = cancer ~ genre, family = binomial("logit"), data = df)

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.30060    0.08382 -15.517  < 2e-16 ***
genreF              ???      0.11026                  ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

On sait que les femmes ont deux fois plus de chances d'avoir un cancer que les hommes. Donnez un ordre de grandeur pour la valeur manquante

A) -0.452

B) 0.452

C) 0.693

D) -0.693

Toutes mes excuses mais aucune réponse n'était valide ici...On a ici un modèle de la forme (si $Y = 1$ quand la personne a un cancer)

$$\mathbb{P}[Y = 1|x] = \frac{\exp(\beta_0 + \beta_1 \mathbf{1}_{\mathsf{F}}(x))}{1 + \exp(\beta_0 + \beta_1 \mathbf{1}_{\mathsf{F}}(x))} = \frac{\exp(-1.3 + \beta_1 \mathbf{1}_{\mathsf{F}}(x))}{1 + \exp(-1.3 + \beta_1 \mathbf{1}_{\mathsf{F}}(x))}$$

Or on nous dit que

$$\mathbb{P}[Y = 1|x = \mathsf{F}] = 2 \times \mathbb{P}[Y = 1|x = \mathsf{H}]$$

autrement dit

$$\frac{\exp(-1.3 + \beta_1 \cdot 1)}{1 + \exp(-1.3 + \beta_1 \cdot 1)} = 2 \times \frac{\exp(-1.3 + \beta_1 \cdot 0)}{1 + \exp(-1.3 + \beta_1 \cdot 0)} = 2 \cdot \frac{\exp(-1.3)}{1 + \exp(-1.3)} = 0.428$$

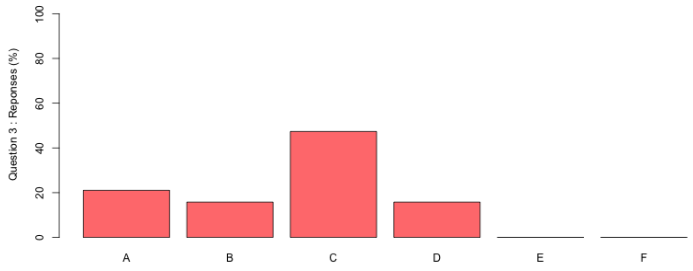
Or on sait que si

$$\frac{\exp(x)}{1 + \exp(x)} = y \text{ alors } x = \log\left(\frac{y}{1 - y}\right)$$

(on a suffisement manipulé cette relation dans le cours) donc

$$(-1.3 + \beta_1) = \log \frac{0.428}{1 - 0.428} = -0.29$$

donc $\beta_1 = -0.29 + 1.300 \sim 1$. Qui n'était pas proposé ici.



J'ai l'impression qu'il s'agit d'une typo : si on suppose que les femmes ont deux fois **moins** de chances d'avoir un cancer que les hommes, avec la réponse D, on avait un score d'environ -2, or $\exp(-2)/(1 + \exp(-2)) \sim 0.11$ qui est environ la moitié de $\exp(-1.3)/(1 + \exp(-1.3)) \sim 0.21$. Pour la note finale, j'ai retiré cette question, et noté sur 29 !

4 On a autant d'hommes que de femmes dans l'échantillon

```
Call:
glm(formula = cancer ~ 0 + genre, family = binomial("logit"), data = df)

Coefficients:
         Estimate Std. Error z value Pr(>|z|)
genreH   -1.30060    0.08382  -15.52   <2e-16 ***
genreF   -0.93642    0.07163  -13.07   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Donner un ordre de grandeur du ratio $\mathbb{P}[\text{cancer}|x = \texttt{H}]/\mathbb{P}[\text{cancer}|x = \texttt{F}]$

A) -0.847

B) 0.847

C) 2.601

D) -2.601

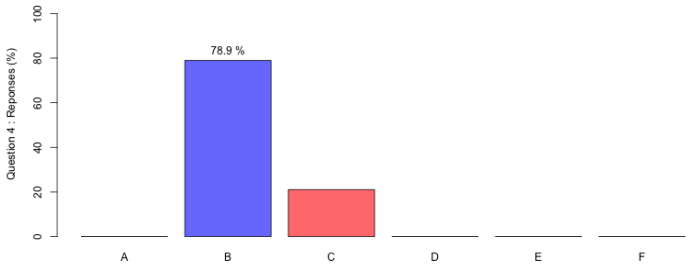
On peut utiliser les calculs précédants. On nous demande

$$\frac{\mathbb{P}[\text{cancer}|x = \texttt{H}]}{\mathbb{P}[\text{cancer}|x = \texttt{F}]} = \frac{\exp[\widehat{\beta}_{\texttt{H}}]}{1+\exp[\widehat{\beta}_{\texttt{H}}]} \cdot \frac{1+\exp[\widehat{\beta}_{\texttt{F}}]}{\exp[\widehat{\beta}_{\texttt{F}}]}$$

soit

$$\frac{\exp[-1.30060]}{1+\exp[-1.30060]} \cdot \frac{1+\exp[-0.93642]}{\exp[-0.93642]} \sim 0.8246159$$

qui correspond à la réponse B (on demandait un ordre de grandeur...)



5 On considere la sortie de regression suivante, avec une variable binaire $y \in \{0,1\}$ et deux variables explicatives continues, x_1 et x_2 à valeurs dans l'intervalle $[0,1]$.

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

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.5659     0.1071  -5.283 1.27e-07 ***
x1            0.8571     0.1379   6.215 5.14e-10 ***
x2           -0.2674     0.1390  -1.923   0.0545 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Sur la Figure 1, représentez la région des points (x_1, x_2) pour lesquels $\mathbb{P}[Y = 1|x_1, x_2] > \mathbb{P}[Y = 0|x_1, x_2]$.

Commençons par chercher la frontière, c'est à dire l'ensemble des points (x_1, x_2) tels que $\mathbb{P}[Y = 1|x_1, x_2]$=$\mathbb{P}[Y = 0|x_1, x_2]$. Comme on l'a vu en cours, les GLM sont fondamentalement des modèles linéaires. La frontière sera une droite. En fait, on a vu en cours que les courbes d'iso-densité sont des droites qui sont parallèles entre elles. On se lance ! Si $\mathbb{P}[Y = 1|x_1, x_2]$=$\mathbb{P}[Y = 0|x_1, x_2]$ c'est que $\mathbb{P}[Y = 1|x_1, x_2]$=$1/2$, soit

$$\mathbb{P}[Y = 1|x_1, x_2] = \frac{\exp(\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2)}{1 + \exp(\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2)} = \frac{1}{2}$$

mais ce n'est peut-être pas la formule la plus simple à utiliser. De manière équivalente, on veut (on inverse la fonction logistique)

$$\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 = \log \frac{1/2}{1 - 1/2} = \log 1 = 0$$

On cherche donc l'ensemble des points (x_1, x_2) tels que $\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 = 0$. On reconnaît une droite. Si on veut la tracer dans le repère (x_1, x_2), il s'agit de la droite

$$x_2 = -\frac{\widehat{\beta}_0 + \widehat{\beta}_1 x_1}{\widehat{\beta}_2} = \frac{-0.5659 + 0.8571 \cdot x_1}{0.2674} = -2.113 + 3.2053 \cdot x_1$$

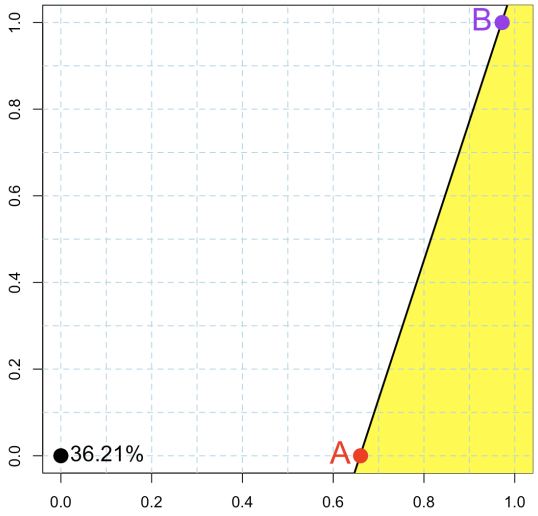
Pour la tracer, on peut (plus simplement) chercher deux points par lesquelles passe la droite. On peut regarder les valeurs de x_1 pour lesquelles x_2 vaut 0 (axe du bas) et 1 (axe du haut). Pour $x_2 = 0$, on doit avoir $-0.5659 + 0.8571 \cdot x_1 = 0$, et comme $0.5659/0.8571 \sim 0.660249$ la droite passe par A $= (0.66, 0)$. Pour $x_2 = 1$, on doit avoir $-0.5659 + 0.8571 \cdot x_1 = 0.2674$,

et comme $(0.2674 + 0.5659)/0.8571 \sim 0.9722319$ la droite passe par B $= (0.97, 1)$.

Mais la question était de trouvait l région où la probabilité de valoir 1 excède la probabilité de valoir 0. Le plus simple est de prendre un point quelconque, simple, et de voir dans quelle région il se trouve. Le plus simple (pour les calculs) est probablement le point $(0, 0)$. Or pour ce point, la probabilité prédite est

$$\mathbb{P}[Y = 1|0, 0] = \frac{\exp(\widehat{\beta}_0 + \widehat{\beta}_1 \cdot 0 + \widehat{\beta}_2 \cdot 0)}{1 + \exp(\widehat{\beta}_0 + \widehat{\beta}_1 \cdot 0 + \widehat{\beta}_2 x \cdot 0)} = \frac{\exp(\widehat{\beta}_0)}{1 + \exp(\widehat{\beta}_0)} = 0.3621834$$

(en réalité, on n'a pas vraiment besoin de faire le calcul : comme $\widehat{\beta}_0 < 0$, on sait que la probabilité est plus petite que $1/2$) Autrement dit, $(0, 0)$ est dans la partie où la probabilité de valoir 0 est la plus grande ! C'est donc la partie inférieure (en jaune sur le dessin) qui correspond au cas où $\mathbb{P}[Y = 1|x_1, x_2] > \mathbb{P}[Y = 0|x_1, x_2]$.



Pour avoir 1 point, il fallait colorier la région en bas à droite, avec une frontière linéaire - c'est un modèle linéaire, c'est la moindre des choses que je voulais voir. Certains se sont contenter de calculer pour quelques valeurs, et on eu $1/2$ point (une copie montrait une vingtaine de points en bas à droite, mais sans la frontière).

6 Sur les mêmes données, on décide de couper x_1 et x_2 en classes.

```
Call:
glm(formula = y ~ (x1 > 0.8) + (x2 > 0.65), family = binomial("logit"),
    data = df)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.30533    0.05398  -5.656 1.55e-08 ***
x1 > 0.8TRUE   0.41028    0.09646   4.253 2.11e-05 ***
x2 > 0.65TRUE -0.14849    0.08380  -1.772   0.0764 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Sur la Figure 2, représentez la région des points (x_1, x_2) pour lesquels $\mathbb{P}[Y = 1|x_1, x_2] > \mathbb{P}[Y = 0|x_1, x_2]$.

Ici, on a tranformé nos variables continues en catégories. Aussi, x_1 devient $\{x_1 \leqslant 0.8\}$ et $\{x_1 > 0.8\}$, alors que x_2 devient $\{x_2 \leqslant 0.65\}$ et $\{x_2 > 0.65\}$. Autrement dit, on régresse ici sur 2 variables, ayant chacun deux modalités. On a alors 4 possibilités (correspondant au produit cartésien des deux variables) :

- $\{x_1 \leqslant 0.8\} \cap \{x_2 \leqslant 0.65\}$ - qui contient $(0, 0)$ (coin sud-ouest)
- $\{x_1 \leqslant 0.8\} \cap \{x_2 > 0.65\}$ - qui contient $(0, 1)$ (coin nord-ouest)
- $\{x_1 > 0.8\} \cap \{x_2 \leqslant 0.65\}$ - qui contient $(1, 0)$ (coin sud-est)
- $\{x_1 > 0.8\} \cap \{x_2 > 0.65\}$ - qui contient $(1, 1)$ (coin nord-est)

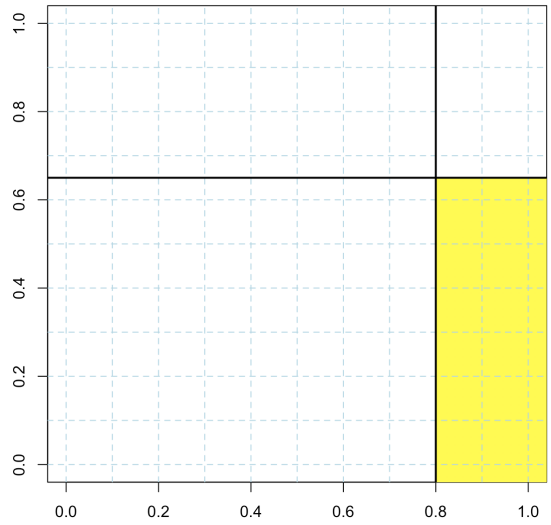
Oui, j'ai rajouté 4 points qui sont dans ces quatre ensembles parce qu'ici, pour répondre à la question, il va valoir calculer les probabilités dans les quatre ensembles, et les mettre sur la figure. Les probabilités sont alors

- $\{x_1 \leqslant 0.8\} \cap \{x_2 \leqslant 0.65\}$: $\dfrac{\exp(\widehat{\beta}_0)}{1+\exp(\widehat{\beta}_0)}$
- $\{x_1 \leqslant 0.8\} \cap \{x_2 > 0.65\}$: $\dfrac{\exp(\widehat{\beta}_0+\widehat{\beta}_2)}{1+\exp(\widehat{\beta}_0+\widehat{\beta}_2)}$
- $\{x_1 > 0.8\} \cap \{x_2 \leqslant 0.65\}$: $\dfrac{\exp(\widehat{\beta}_0+\widehat{\beta}_1)}{1+\exp(\widehat{\beta}_0+\widehat{\beta}_1)}$
- $\{x_1 > 0.8\} \cap \{x_2 > 0.65\}$: $\dfrac{\exp(\widehat{\beta}_0+\widehat{\beta}_1+\widehat{\beta}_2)}{1+\exp(\widehat{\beta}_0+\widehat{\beta}_1+\widehat{\beta}_2)}$

Ca semble compliqué, mais en fait, on peut faire plus simple : la question est juste '*est-ce que dans cette ensemble, la probabilité de valoir* 1 *dépasse 1/2* ?' Mais là encore, on sait que la probabilité dépasse $1/2$ si la combinaison linéaire (la partie à l'intérieure de l'exponentielle) est positive) ! Donc inutile de faire les calculs !!!

- (coin sud-ouest) : $\widehat{\beta}_0 = -0.30 < 0$ donc $\mathbb{P}[Y = 1|x_1, x_2] < 1/2$
- (coin nord-ouest): $\widehat{\beta}_0 + \widehat{\beta}_2 = -0.45 < 0$ donc $\mathbb{P}[Y = 1|x_1, x_2] < 1/2$
- (coin sud-est): $\widehat{\beta}_0 + \widehat{\beta}_1 = 0.10 > 0$ donc $\mathbb{P}[Y = 1|x_1, x_2] > 1/2$
- (coin nord-est): $\widehat{\beta}_0 + \widehat{\beta}_1 + \widehat{\beta}_2 = -0.04 < 0$ donc $\mathbb{P}[Y = 1|x_1, x_2] < 1/2$

Seule la partie en bas à droite donne une probabilité de valoir 1 plus grande (ce qui est cohérent avec la question précédante)



Pour avoir 1 point, il fallait colorier la région en bas à droite, avec les deux droites - une horizontale et l'autre verticale - bien placées : 0.8 en x_1 et 0.65 en x_2

7 Considérons une variable y prenant les valeurs $\{A, B\}$. La régression de $\mathbf{1}_A(y)$ sur x_1 et x_2 donne

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

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.9524     0.2890   3.296 0.000982 ***
x1           -0.4728     0.1454  -3.252 0.001147 **
x2           -0.9676     0.4826  -2.005 0.044957 *
```

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

La régression de $\mathbf{1}_B(y)$ sur x_1 et x_2 donne

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

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    ???
x1             ???
x2             ???
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Donnez la valeur de $\widehat{\beta}_0$ (estimateur de la constante) dans la seconde régression

A) 0.9524

B) 1.0498

C) -0.9524

D) 0.0476

Pour obtenir le résultat, un rapide calcul préalable :

$$\frac{e^u}{1+e^u} = \frac{e^u}{e^u \cdot e^{-u} + e^u} = \frac{1}{e^{-u}+1} = \frac{1}{1+e^{-u}}.$$

et de manière symmétrique

$$\frac{1}{1+e^u} = \frac{e^{-u}}{1+e^{-u}}.$$

Maintenant on se lance : on a ici pour la seconde équation

$$\mathbb{P}[Y = \mathsf{B}|x_1, x_2] = \frac{\exp(\widehat{\beta}_0 + \widehat{\beta}_1 \cdot x_1 + \widehat{\beta}_2 \cdot x_2)}{1 + \exp(\widehat{\beta}_0 + \widehat{\beta}_1 \cdot x_1 + \widehat{\beta}_2 x \cdot x_2)}$$

La première équation nous disait que

$$\mathbb{P}[Y = \mathsf{A}|x_1, x_2] = \frac{\exp(\widehat{\alpha}_0 + \widehat{\alpha}_1 \cdot x_1 + \widehat{\alpha}_2 \cdot x_2)}{1 + \exp(\widehat{\alpha}_0 + \widehat{\alpha}_1 \cdot x_1 + \widehat{\alpha}_2 x \cdot x_2)}$$

Or on sait que $\mathbb{P}[Y = \mathsf{B}|x_1, x_2] = 1 - \mathbb{P}[Y = \mathsf{A}|x_1, x_2]$, autrement dit

$$\frac{\exp(\widehat{\beta}_0 + \widehat{\beta}_1 \cdot x_1 + \widehat{\beta}_2 \cdot x_2)}{1 + \exp(\widehat{\beta}_0 + \widehat{\beta}_1 \cdot x_1 + \widehat{\beta}_2 x \cdot x_2)} = \frac{1}{1 + \exp(\widehat{\alpha}_0 + \widehat{\alpha}_1 \cdot x_1 + \widehat{\alpha}_2 x \cdot x_2)}$$

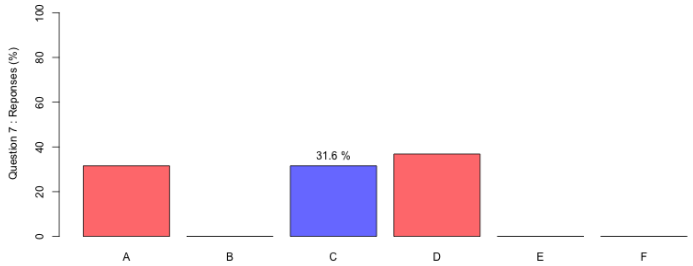
donc en utilisant l'expression montrée en préalable

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

Par identification (ou en utilisant l'inverse de la loi logistique), on en déduit que

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

Aussi, $\widehat{\beta}_0$ vaut ici -0.9524 ce qui correspond à la réponse C.



8 Sur la même sortie que la question précédante, donnez la valeur de $\widehat{\beta}_1$ (estimateur associé a la variable x_1) dans la seconde régression

A) -0.4728

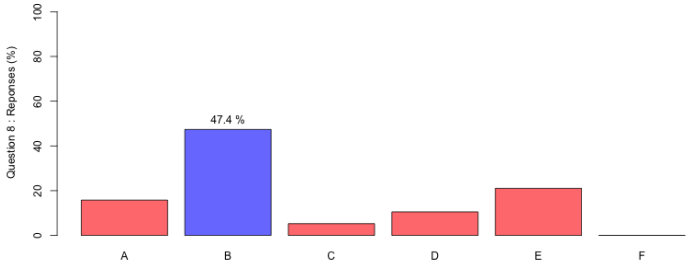
B) 0.4728

C) 2.1151

D) 0.5272

E) on ne peut pas savoir

en utilisant le calcul prédédant, on a la réponse B automatiquement.



9 Soit N une variable de comptage (d'accidents de la route), et on considère deux variables explicatives (possibles) $X_1 \in \{\mathsf{male}, \mathsf{female}\}$ et $X_2 \in \{\mathsf{urban}, \mathsf{mid\text{-}urban}, \mathsf{rural}\}$. Soit E le nombre d'individus par classe. On a observé les données suivantes (présentée ici sous forme de tableau de contingence)

E	men	women
urban	100	100
mid-urban	100	100
rural	100	100

N	men	women
urban	10	8
mid-urban	13	10
rural	7	5

On considère une régression de Poisson $(N|X_1, X_2, E) \sim \mathcal{P}(E \cdot \lambda_{x_1,x_2})$ où

$$\lambda_{x_1,x_2} = \exp\left[\gamma_0 + \alpha_{\mathsf{w}}\mathbf{1}(x_1 = \mathsf{woman}) + \beta_{\mathsf{r}}\mathbf{1}(x_2 = \mathsf{rural}) + \beta_{\mathsf{u}}\,\mathbf{1}(x_2 = \mathsf{urban})\right]$$

En utilisant les données individuelles, et en faisant une régression sous R, on obtient la sortie suivante

```
> summary(reg)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -2.0388     0.2407  -8.470   <2e-16 ***
woman           ???
rural         -0.6506     0.3561  -1.827   0.0677 .
urban         -0.2451     0.3147  -0.779   0.4360
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Quelle est la valeur de $\widehat{\alpha}_{\mathsf{w}}$ dans la sortie précédante ?

A) -0.3582

B) -0.7814

C) -0.2657

D) 0.1426

La question était un peu difficile (mais il faut bosser un minimum pour avoir plus de 95%, non ?) Et encore, on va le voir, le calcul est incroyablement simple, si on a compris comme fonctionne la régression de Poisson. On avait vu en cours que la méthode des marges était équivalente à une régression de Poisson sur les facteurs - avec un lien logarithmique, et donc un modèle multiplicatif. Aussi,

$$\sum_i E_{i,j} \cdot A_i B_j = \sum_i N_{i,j} \text{ et } \sum_j E_{i,j} \cdot A_i B_j = \sum_j N_{i,j}$$

Or ici, on nous donne (presque) les valeurs de $A = (A_{\mathsf{m}}, A_{\mathsf{v}})$ et $A = (B_{\mathsf{mu}}, B_{\mathsf{r}}, B_{\mathsf{u}})$. En effet, $A = (e^{\gamma_0}, e^{\gamma_0+\alpha_{\mathsf{w}}})$ et $B = (1, e^{\beta_{\mathsf{r}}}, e^{\beta_{\mathsf{u}}})$. En fait, comme le problème initial n'est pas identifiable, on peut choisir la normalisation que l'on souhaite. On peut aussi écrire $A = (1, e^{\alpha_{\mathsf{w}}})$ et $B = (e^{\gamma_0}, e^{\gamma_0+\beta_{\mathsf{r}}}, e^{\gamma_0+\beta_{\mathsf{u}}})$ - ce dernier étant proportionnel à $(1, e^{\beta_{\mathsf{r}}}, e^{\beta_{\mathsf{u}}})$. Le point ici est que l'on connait le vecteur A. Or on a vu en cours que si on connait A, on connait B (et inversement... d'où l'idée de l'algorithme itératif pour trouver le point fixe). Ici, si on remplace par les valeurs numérique, on sait que

$$B = (1, e^{\alpha_{\mathsf{r}}}, e^{\alpha_{\mathsf{u}}}) = (1, 0.522, 0.783)$$

De nos premières équations, on peut déduire que

$$B_j \sum_i E_{i,j} \cdot A_i = \sum_i N_{i,j} \text{ et } A_i \sum_j E_{i,j} \cdot B_j = \sum_j N_{i,j}$$

en particulier

$$A_i = \frac{\sum_j N_{i,j}}{\sum_j E_{i,j} \cdot B_j}$$

soit

$$A_{\mathsf{m}} = \frac{\sum_{j\in\{\mathsf{u},\mathsf{mu},\mathsf{r}\}} N_{\mathsf{m},j}}{\sum_{j\in\{\mathsf{u},\mathsf{mu},\mathsf{r}\}} E_{\mathsf{m},j} \cdot B_j} = \frac{10+13+7}{100 \cdot (1+0.522+0.783)}$$

alors que

$$A_{\mathsf{w}} = \frac{\sum_{j\in\{\mathsf{u},\mathsf{mu},\mathsf{r}\}} N_{\mathsf{m},j}}{\sum_{j\in\{\mathsf{u},\mathsf{mu},\mathsf{r}\}} E_{\mathsf{m},j} \cdot B_j} = \frac{8+10+5}{100 \cdot (1+0.522+0.783)}$$

Autrement dit, on remarque que l'on a une relation très simple ! (on aurait pu la voir avant, cela dit)

$$\frac{A_i}{A_{i'}} = \frac{\sum_j N_{i,j}}{\sum_j N_{i,j'}}$$

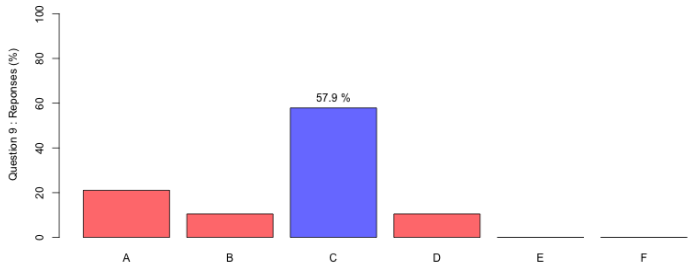
et si i' est la modalité de réfŕence,

$$\frac{A_i}{A_{i'}} = \exp \beta_{j'} \frac{\sum_j N_{i,j}}{\sum_j N_{i,j'}}$$

Aussi, ici

$$\exp(\widehat{\alpha}_{\mathsf{w}}) = \frac{8+10+5}{10+13+7} = \frac{23}{30} \sim 0.7666$$

et donc $\widehat{\alpha}_{\mathsf{w}} = \log 23 - \log 30 = -0.2657$ qui est la réponse C.



10 On suppose que la matrice de variance-covariance pour la sortie précédante a été obtenue

```
> vcov(reg)
```

	(Intercept)	X1woman	X2rural	X2urban
(Intercept)	0.05794367	-3.333333e-02	-4.347826e-02	-4.347826e-02
X1woman	-0.03333333	7.681159e-02	-7.803914e-18	-2.675628e-18
X2rural	-0.04347826	-7.803914e-18	1.268116e-01	4.347826e-02
X2urban	-0.04347826	-2.675628e-18	4.347826e-02	9.903381e-02

Donner un intervalle de confiance à 95% pour $\gamma_0 + \beta_{\mathsf{u}}$

A) $[-4.28; -0.28]$

B) $[-2.33; -2.23]$

C) $[-2.51; -1.55]$

D) $[-2.81; -1.75]$

On l'avait revu au dernier cours (en fait, c'est la même chose que pour les modèles linéaires). On utilise une hypothèse de normalité du vecteur des

estimateurs paramètres, de telle sorte que tout combinaison linéaire sera aussi Gaussienne. Aussi, pour avoir un intervalle de confiance pour $\gamma_0 + \beta_u$, il suffit d'avoir son estimateur, et sa variance. En effet, l'intervalle à 95% sécrit

$$\left[(\widehat{\gamma}_0 + \widehat{\beta}_u\ ) \pm 1.96\sqrt{\mathsf{Var}[\widehat{\gamma}_0 + \widehat{\beta}_u\ ]}\right]$$

Le terme de gauche est simplement $-2.0388 - 0.2451 = -2.2839$. Pour le terme de droite, il faut un minimum de calculs, puisque

$$\mathsf{Var}[\widehat{\gamma}_0 + \widehat{\beta}_u\ ] = \mathsf{Var}[\widehat{\gamma}_0] + 2\mathsf{Cov}[\widehat{\gamma}_0, \widehat{\beta}_u\ ] + \mathsf{Var}[\widehat{\beta}_u\ ]$$

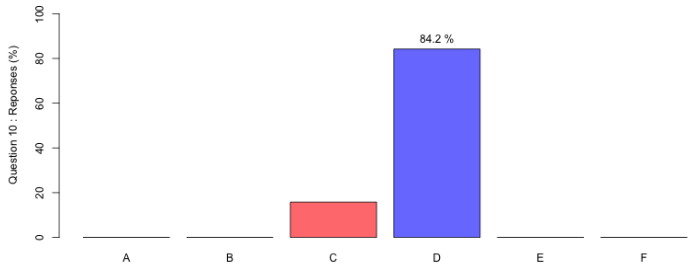
donc, compte tenu des valeurs données dans la matrice de variance-covariance,

$$\mathsf{Var}[\widehat{\gamma}_0 + \widehat{\beta}_u\ ] = 0.058 + 2 \cdot (-0.043) + 0.099 = 0.07$$

Aussi, l'intervalle de confiance s'écrit

$$\left[\ -2.284 \pm 1.96\sqrt{0.07}\right] = \left[\ -2.284 \pm 0.518\right] = [-2.81; -1.75]$$

qui est la réponse D.



11 Combien d'accidents pour les hommes vivant en ville (*man* et *urban*) doit-on espérer avoir, avec 95% de chance, quand 100 personnes sont dans le portefeuille de l'assureur

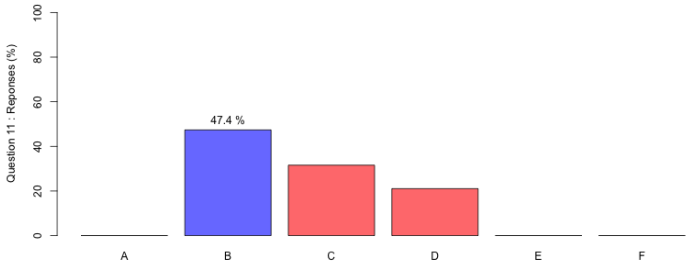
A) $[5.12; 95.29]$

B) $[5.66; 14.74]$

C) $[7.44; 17.40]$

D) $[5.52; 17.52]$

Une fois qu'on a notre intervalle de confiance sur le score, on a l'intervalle de confiance sur son exponentielle - car l'exponentielle est une fonction croissante. Aussi, l'intervalle de confiance à 95% **pour la fréquence d'accident** est $[\exp(-2.81); \exp(-1.75)]$ soit $[0.0566; 0.1474]$ (attention aux erreurs d'arrondis). Comme on a 100 personnes, on a alor s comme intervalle $[5.66; 14.74]$.



12 On observe les statistiques de décès lors de voyages en avions suivantes (les nombres de passagers sont ici en millions de miles parcourus)

	annee	deces	passagers
1	1976	734	386300
2	1977	516	430000
3	1978	754	502700
4	1979	877	548100
5	1980	814	581400
6	1981	362	603300
7	1982	764	587700
8	1983	809	622300
9	1984	223	743300
10	1985	1066	710700

Un premier modèle est considéré, avec une loi binomiale, $D_i \sim \mathcal{B}(E_i, p_i)$, où D est le nombre de décès, et E le nombre de passagers, et p est une transformation logistique d'une fonction linéaire de l'année

```
Call:
glm(formula = cbind(deces, passagers - deces) ~ annee, family = binomial,
    data = df)

Coefficients:
```

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 113.312570   8.425622   13.45   <2e-16 ***
annee        -0.060597   0.004254    ???      ???
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

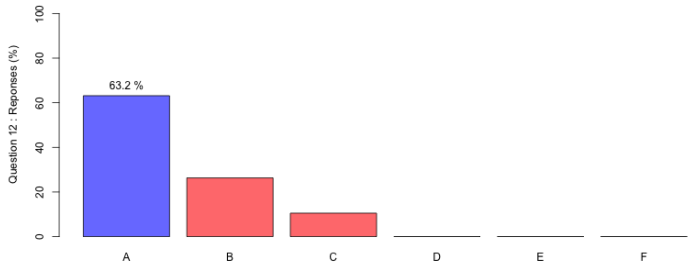
On lit dans un journal que la baisse du nombre de décès (relativement à la hausse du nombre de passagers) n'est pas significative. Qu'en pensez vous ?

A) C'est faux, la baisse est significative

B) C'est vrai, la baisse n'est pas statistiquement significative

C) On n'a pas assez d'observations pour conclure dans un sens ou dans l'autre

Pour trancher, on utilise la statistique z qui permet de tester la significativité de la variable. En l'occurence, la valeur manquante est ici $z = \widehat{\beta}/\text{se}(\widehat{\beta})$ soit ici $-0.060597/0.004254 = -14.244$ qui est encore plus grande (en valeur absolue) que la valeur au dessus dans la sortie... donc la p-value est inférieure à 2e16. Bref, la variable est significative, très très largement ! La réponse A est ici correcte.



13 On tente comme alternative une régression de Poisson $D_i \sim \mathcal{P}(E_i \cdot \lambda_i)$, où λ est une transformation exponentielle d'une fonction linéaire de l'année,

```
Call:
glm(formula = deces ~ annee + offset(log(passagers)), family = poisson,
    data = df)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-22.342   -3.886    3.351    4.778   14.240

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 113.162703   8.420250   13.44   <2e-16 ***
annee        -0.060522   0.004252  -14.23   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Cette sortie donne un modèle très proche du précédant. Parmi les quatre explications suivantes, **laquelle est fausse**

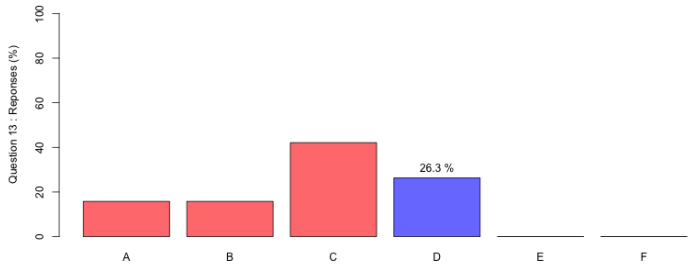
A) Si $Y \sim \mathcal{B}(n,p)$ avec n très grand, et p petit, alors $Y \approx \mathcal{P}(np)$

B) La probabilité de mourir est faible, et $\log(p/(1-p)) \approx \log(p)$

C) La probabilité de mourir est faible, et $e^x/(1+e^x) \approx e^x$

D) Le nombre d'observations est important, et asymptotiquement, la régression de Poisson et la régression logistique sont équivalentes

La première est vraie, c'est la loi des petits nombres. Je renvoie à un cours de probabilité pour la preuve (sinon en gros, on pose $\lambda = np$, et on utilise que $(1-\lambda/n)^n \sim e^{-\lambda}$). Pour les deux suivantes, je les avais évoquées en cours. Pour moi, c'est la dernière, qui n'avait aucun rapport avec le sujet, qu'il fallait rejeter.



Compte tenu du fort taux de réponse pour C (relativement aux autres), je vais expliquer pourquoi B et C sont liés ici ! Dans le premier modèle, on disait que $D_i \sim \mathcal{B}(E_i, p_i)$ alors que pour le second $D_i \sim \mathcal{P}(E_i \cdot \lambda_i)$, avec respectivement

$$p_i = \frac{e^{\alpha_0+\alpha_1 x_i}}{1+e^{\alpha_0+\alpha_1 x_i}} \text{ et } \lambda_i = e^{\beta_0+\beta_1 x_i}$$

où x désigne l'année. Or on sait que la loi de Poisson $\mathcal{P}(\lambda)$ est une approximation de la loi Binomiale $\mathcal{B}(e,p)$ avec $\lambda = ep$ lorsque e est grand, et p est faible (c'est a question A, et je renvoie au cours d'introduction aux probabilités pour une preuve). p est ici la probabilité de mourir. Et elle est faible, comment l'affirment B et C. En effet, si on utilise le premier modele, comme $x > 1975$,

$$p < \frac{e^{113.31-0.060597\cdot 1975}}{1+e^{113.31-0.060597\cdot 1975}} \sim 0.001715207$$

Aussi ici, si on regarde la prévision, on peut écrire avec le premier modèle

$$\mathbb{E}[D|E,x] = E \cdot \frac{e^{\alpha_0+\alpha_1 x}}{1+e^{\alpha_0+\alpha_1 x}}$$

et avec le second

$$\mathbb{E}[D|E,x] = E \cdot e^{\beta_0+\beta_1 x}$$

On peut alors conclure que $\alpha_0 + \alpha_1 x \sim \beta_0 + \beta_1 x$ (et donc $\alpha_0 \sim \beta_0$ et $\alpha_1 \sim \beta_1$) en utilisant le fait que ,

$$\frac{e^z}{1+e^z} \approx e^z \text{ lorsque } z \to -\infty$$

(et c'est le cas ici car $\alpha_0 + \alpha_1 x$ et $\beta_0 + \beta_1 x$ sont négatifs, et importants (en valeur absolue)). Donc C est juste...

14 Qu'est-ce que la *surdispersion* dans un modèle ?

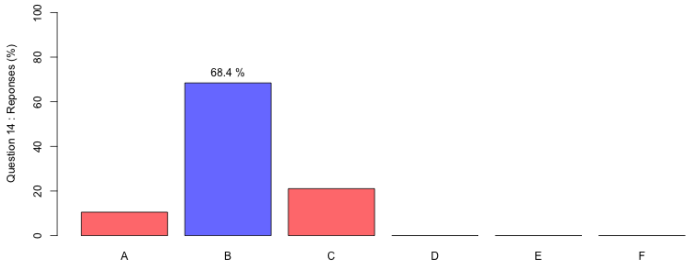
A) c'est lorsque la variance de $\widehat{Y}$ est plus grande que la variance de Y

B) c'est lorsque $\text{Var}[Y|X] > \mathbb{E}[Y|X]$

C) c'est lorsque $\text{Var}[Y|X] > \text{Var}[Y]$

D) c'est comme la notion d'homoscédasticité dans les modèles linéaires, mais pour les GLM

Je renvoie au cours... c'est la définition !



15 question 29 (examen CAS)

29. You are given the following information for a fitted GLM:

Response variable	Occurrence of Accidents
Response distribution	Binomial
Link	Logit

Parameter	df	$\hat{\beta}$
Intercept	1	x
Driver's Age	2	
1	1	0.288
2	1	0.064
3	0	0
Area	2	
A	1	-0.036
B	1	0.053
C	0	0
Vehicle Body	2	
Bus	1	1.136
Other	1	-0.371
Sedan	0	0

The probability of a driver in age group 2, from area C and with vehicle body type Other, having an accident is 0.22.
Calculate the odds ratio of the driver in age group 3, from area C and with vehicle body type Sedan having an accident.
A. Less than 0.200
B. At least 0.200, but less than 0.250
C. At least 0.250, but less than 0.300
D. At least 0.300, but less than 0.350
E. At least 0.350

La bonne réponse était la réponse E.

Ici, il manque la constante, $\widehat{\beta}$. Mais on nous donne une information : la probabilité d'avoir un accident dans le groupe d'âge 2, dans la zone C et pour un véhicule Other est de 22%. Aussi,

$$\frac{\exp(\widehat{\beta}+0.064+0-0.371)}{1+\exp(\widehat{\beta}+0.064+0-0.371)}=0.22$$

ou encore (comme auparavant, ce n'est pas la formule la plus simple à utiliser

$$\widehat{\beta}+0.064+0-0.371=\log\frac{0.22}{1-0.22}=-1.265666$$

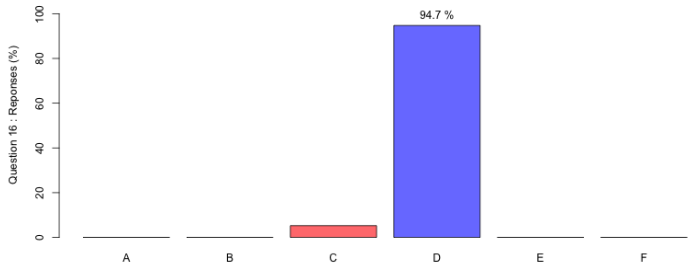
soit $\widehat{\beta}=-0.9587$. À partir de là, on peut faire le calcul demandé : la probabilité est ici

$$p=\frac{\exp(-0.9587+0+0+0)}{1+\exp(-0.9587+0+0+0)}$$

ou (en fait c'est la question posée) la cote (le *odds-ratio*) est ici

$$\frac{p}{1-p} = \exp(-0.9587 + 0 + 0 + 0) = 0.3834$$

qui est la réponse E.



16 question 30 (examen CAS)

30. You are given the following information for a fitted GLM:

Response variable	Occurrence of Accidents
Response distribution	Binomial
Link	Logit

Parameter		df	$\hat{\beta}$	se
Intercept		1	-2.358	0.048
Area		2		
	Suburban	0	0.000	
	Urban	1	0.905	0.062
	Rural	1	-1.129	0.151

Calculate the modeled probability of an Urban driver having an accident.

A. Less than 0.01
B. At least 0.01, but less than 0.05
C. At least 0.05, but less than 0.10
D. At least 0.10, but less than 0.20
E. At least 0.20

La bonne réponse était la réponse D.

On utilise ici tout simplement la formule

$$\mathbb{P}[Y = 1 | X = \text{Urban}] = \frac{\exp(\widehat{\beta}_0 + \widehat{\beta}_{\text{u}})}{1 + \exp(\widehat{\beta}_0 + \widehat{\beta}_{\text{u}})} = \frac{\exp[-2.358 + 0.905]}{1 + \exp[-2.358 + 0.905]} = 0.190$$

qui est entre 10% et 20%, correspondant à la réponse D.

17 question 40 (examen CAS)

40. You are given the following results from a fitted GLM on the frequency of accidents:

Parameter	df	$\hat{\beta}$	se
Intercept	1	-11.2141	0.1826
Location	1		
Rural	0	0.0000	
City	1	1.0874	0.3162

Calculate the Wald statistic for testing the null hypothesis of $\beta_{city} = 0$.

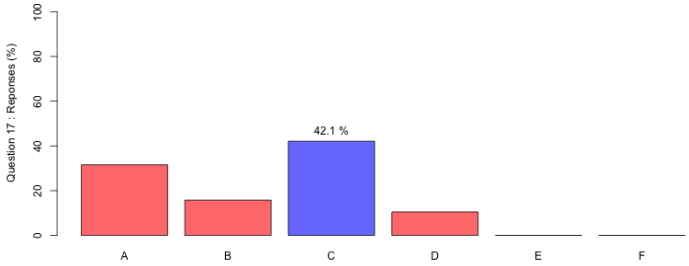
A. Less than 10.5
B. At least 10.5, but less than 11.5
C. At least 11.5, but less than 12.5
D. At least 12.5, but less than 13.5
E. At least 13.5

La bonne réponse était la réponse C.

J'avais évoqué en cours le test de Wald. En effet, on avait vu le test de Student (ici le z-test), qui propose de regarder le ratio $\widehat{\beta}/\sqrt{\text{Var}[\widehat{\beta}]}$, qui doit suivre (asymptotiquement) une loi normale centrée et réduite si $H_0 : \beta = 0$ est vraie. Une alternative est d'utiliser le test de Wald, qui nous dit que le ratio $\widehat{\beta}^2/\text{Var}[\widehat{\beta}]$ doit suivre (asymptotiquement) une du chi-deux à un degré de liberté si $H_0 : \beta = 0$ est vraie. Rien de surprenant, car on sait que si $Z \sim \mathcal{N}(0,1)$, alors $Z^2 \sim \chi^2(1)$. L'avantage ici c'est que la statistique est positive, c'est une distance : on rejette H_0 si la statistique de test est grande (dans le test de Student, on rejette si *la valeur absolue* de la statistique est grande). Bref, ici on regarde $\widehat{\beta}^2/\text{Var}[\widehat{\beta}]$ - où la variance (pour rappel) est le carré de l'écart-type. Aussi,

$$W = \frac{\widehat{\beta}^2}{\text{se}[\widehat{\beta}]^2} = \frac{1.0874^2}{0.3162^2} = 11.826$$

qui est la réponse C. Si on va un peu plus loin, en comparant avec le quantile d'une loi du chi-deux, on peut noter que cela correspond à une p-value de 0.06%. Si on utilise la statistique z (classique), on obtient environ 3.5. Or la probabilté qu'une loi normale dépasse 3.5 est de l'ordre de 0.03% - qui est la moitié de 0.06% (l'histoire de la valeur absolue...).



18 question 31 (examen CAS)

31. You are given the following information for a fitted GLM:

Response variable	Claim size	
Response distribution	Gamma	
Link	Log	
Dispersion Parameter	1	
Parameter	df	$\hat{\beta}$
Intercept	1	2.100

Zone		4	
	1	1	7.678
	2	1	4.227
	3	1	1.336
	4	0	0.000
	5	1	1.734
Vehicle Class		6	
	Convertible	1	1.200
	Coupe	1	1.300
	Sedan	0	0.000
	Truck	1	1.406
	Minivan	1	1.875
	Station wagon	1	2.000
	Utility	1	2.500
Driver Age		2	
	Youth	1	2.000
	Middle age	0	0.000
	Old	1	1.800

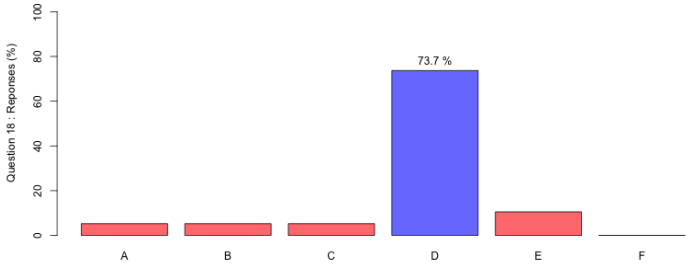
Calculate the predicted claim size for an observation from Zone 3, with Vehicle Class Truck and Driver Age Old.
A. Less than 650
B. At least 650, but less than 700
C. At least 700, but less than 750
D. At least 750, but less than 800
E. At least 800

La bonne réponse était la réponse D.

Là encore, on applique simplement la formule du cours, qui est simple puisqu'on nous dit que le lien est un lien logarithmique ! Comme on a compris l'idée, je vais vite : la valeur moyenne prédite est ici

$$\exp[2.100 + 1.336 + 1.406 + 1.800] = 766.63$$

qui est comprise entre 750 et 800, c'est à dire la réponse D.



19 question 34 (pardon, 32) (examen CAS)

32. You are given the following information for a fitted GLM:

Response variable	Claim size	
Response distribution	Gamma	
Link	Log	
Dispersion parameter	1	
Parameter	df	$\hat{\beta}$
Intercept	1	2.100
Zone	4	
1	1	7.678
2	1	4.227
3	1	1.336
4	0	0.000
5	1	1.734
Vehicle Class	6	
Convertible	1	1.200
Coupe.	1	1.300
Sedan	0	0.000
Truck	1	1.406
Minivan	1	1.875
Station wagon	1	2.000
Utility	1	2.500
Driver Age	2	
Youth	1	2.000
Middle age	0	0.000
Old	1	1.800

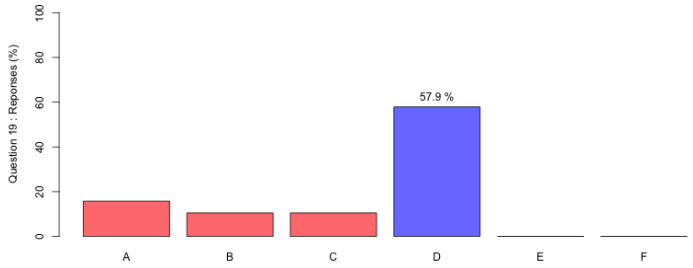
Calculate the variance of a claim size for an observation from Zone 4,
with Vehicle Class Sedan and Driver Age Middle age
A. Less than 55
B. At least 55, but less than 60
C. At least 60, but less than 65
D. At least 65, but less than 70
E. At least 70

La bonne réponse était la réponse D.

Pour la prévision (question d'avant), seule la fonction lien était importante. Ici, on nous interroge sur la variance, donc la loi sera importante. Pour une loi Gamma, la fonction variance est quadratique, ou plus précisément, $\text{Var}[Y|X] = \phi \cdot \mathbb{E}[Y|X]^2$. Ici, on nous dit que le paramètre de dispersion (en fait le paramètre de nuisance) vaut ici 1, donc ici $\text{Var}[Y|X] = \mathbb{E}[Y|X]^2$. Et pour calculer le terme droite, comme pour la question précédante, on note que $\mathbb{E}[Y|X] = \exp[2.1]$ puisque le lien est - là encore - logarithmique, et toutes les autres informations correspondent aux modalités de référence. Bref,

$$\text{Var}[Y|X] = \mathbb{E}[Y|X]^2 = \exp[2 \cdot 2.1] = 66.686$$

qui est compris entre 65 et 70, c'est donc la réponde D.



20 question 32 (pardon, 34) (examen CAS)

34. Determine which of the following statements are true.
I. The deviance is useful for testing the significance of explanatory variables in nested models.
II. The deviance for normal distributions is proportional to the residual sum of squares.
III. The deviance is defined as a measure of distance between the saturated and fitted model.
A. I only B. II only C. III only D. All but III E. All

La bonne réponse était la réponse E.

Toutes sont vraies ! On avait démontré II. au dernier cours, et on avait évoqué III. tout au long de la second partie du cours. Le premier point avait été survolé, mais évoqué. Pour aller un peu plus loin, je peux mettre deux pages du livre de Piet de Jong et Gillian Heller, *Generalized Linear Models for Insurance Data*

$$\frac{\partial \ell}{\partial \theta_i} = \frac{y_i - \mu_i}{\phi} = \frac{y_i - \dot{a}(\theta_i)}{\phi} ,$$

the MLE of θ_i under the saturated model is $\check{\theta}_i$, where $\dot{a}(\check{\theta}_i) = y_i$. Thus each fitted value is equal to the observation and the saturated model fits perfectly.

The value of the saturated log-likelihood is

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

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:

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

- When the model provides a good fit, then $\hat{\ell}$ is expected to be close to (but not greater than) $\check{\ell}$. A large value of the deviance indicates a badly fitting model.
- The size of Δ is assessed relative to the χ^2_{n-p} distribution (Dobson 2002). This is the approximate sampling distribution of the deviance, assuming the fitted model is correct and n is large. The expected value of the deviance is $n - p$, and typically the deviance divided by its degrees of freedom $n - p$ is examined: a value much greater than one indicates a poorly fitting model.
- A direct calculation shows that for the exponential family

$$\Delta = 2 \sum_{i=1}^{n} \left\{ \frac{y_i(\check{\theta}_i - \hat{\theta}_i) - a(\check{\theta}_i) + a(\hat{\theta}_i)}{\phi} \right\} , \qquad (5.11)$$

  since terms involving $c(y_i, \phi)$ cancel out, and where $\check{\theta}_i$ and $\hat{\theta}_i$ are such that $\dot{a}(\check{\theta}_i) = y_i$ and $g\{\dot{a}(\hat{\theta}_i)\} = x_i'\hat{\beta}$, respectively.
- When ϕ is unknown and estimated, then the χ^2_{n-p} distribution for the deviance is compromised. In the case of the Poisson distribution, $\phi = 1$ and the χ^2 approximation is useful. In the case of the normal distribution, when σ^2 is known then the χ^2 distribution of the deviance is exact; however, when σ^2 is estimated then we cannot rely on the deviance being χ^2 distributed. Several authors, for example (McCullagh and Nelder 1989, pp. 119, 122) caution against using the deviance as an overall goodness of fit measure in general, as its approximate χ^2_{n-p} distribution depends on assumptions which are frequently not tenable. However, the deviance is useful for testing the significance of explanatory variables in nested models: see Section 5.8.

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.)

Deviance for well-known distributions. Table 5.2 gives the expressions for the deviance for the exponential family distributions discussed previously. The derivation of the deviance expressions for the normal and Poisson distributions are illustrated and the others are left as exercises.

Normal. In this case

$$a(\theta) = \frac{1}{2}\theta^2 , \qquad \dot{a}(\theta) = \theta , \qquad \check{\theta}_i = y_i , \qquad \hat{\theta}_i = \hat{\mu}_i .$$

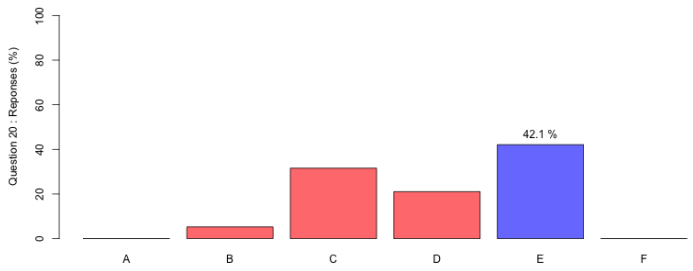
Hence each term in the sum (5.11) is, apart from the divisor $\phi = \sigma^2$

$$y_i(y_i - \hat{\mu}_i) - \frac{1}{2}\left(y_i^2 - \hat{\mu}_i^2\right) = \frac{1}{2}(y_i - \hat{\mu}_i)^2 .$$

Thus the deviance is as given in Table 5.2 and is proportional to the residual sum of squares.

Poisson. In this case $\phi = 1$ and

$$a(\theta) = \mathrm{e}^{\theta} , \qquad \dot{a}(\theta) = \mathrm{e}^{\theta} , \qquad \check{\theta}_i = \ln y_i , \qquad \hat{\theta}_i = \ln \hat{\mu}_i .$$



21 question 37 (examen CAS)

37. Determine which of the following GLM selection considerations is true.
A. The model with the largest AIC is always the best model in model selection process.
B. The model with the largest BIC is always the best model in model selection process.
C. The model with the largest deviance is always the best model in model selection process.
D. Other things equal, when the number of observations > 1000, AIC penalizes more for the number of parameters used in the model than BIC.
E. Other things equal, when number of observations > 1000, BIC penalizes more for the number of parameters used in the model than AIC.

La bonne réponse était la réponse E.

Faisons le dans l'ordre...

Le modèle avec le plus petit AIC est généralement le meilleur modèle dans le processus de sélection, toutes choses étant égales par ailleurs. L'énoncé A n'est pas vrai.

Le modèle avec le plus petit BIC est généralement le meilleur modèle dans le processus de sélection , toutes choses étant égales par ailleurs. L'énoncé B n'est pas vrai.

Le modèle présentant la plus petite déviance est généralement le meilleur modèle dans le processus de sélection, toutes choses étant égales par ailleurs. L'énoncé C n'est pas vrai.

Revenons maintenant aux définitions

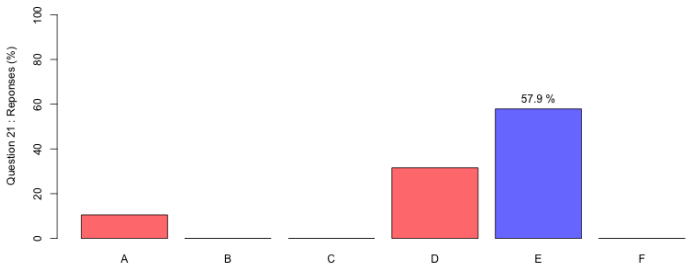
$$AIC = -2 \times \text{log-vraisemblance maximale} + 2 \times \text{nombre de paramètres}$$

et

$$BIC = -2 \times \text{log-vraisemblance maximale} + \log(n) \times \text{nombre de paramètres.}$$

Autrement dit, on pénalise la log vraisemblance, avec comme pénalité pour AIC $2\times$(nombre de paramètres) et pénalité pour BIC (nombre de paramètres) $\times \log$(nombre d'observations). Les deux pénalités sont égales si $\log(n) = 2$, c'est à dire $n \sim e^2 = 7.4$. Aussi, toutes choses égales par ailleurs, lorsque le nombre d'observations dépasse 8, le BIC pénalise davantage que l'AIC. L'énoncé E est donc vrai.

En fait, les énoncés D et E sont opposés, il était fort probable que l'un d'eux soit vrai, et la stratégie optimale (si on ne connaissait pas la réponse) était de tirer au hasard entre les deux.



22 question 41 (examen CAS)

41. A Poisson regression model with log link is used to estimate the number of diabetes deaths. The parameter estimates for the model are:

Response variable	Number of Diabetes Deaths
Response distribution	Poisson
Link	Log

Parameter	df	$\hat{\beta}$	p-value
Intercept	1	-15.000	<0.0001
Gender: Female	1	-1.200	<0.0001
Gender: Male	1	0.000	
Age	1	0.150	<0.0001
Age^2	1	0.004	<0.0001
Age × Gender: Female	1	0.012	<0.0001
Age × Gender: Male	0	0.000	

Calculate the expected number of deaths for a population of 100,000 females age 25.
A. Less than 3
B. At least 3, but less than 5
C. At least 5, but less than 7
D. At least 7, but less than 9
E. At least 9

La bonne réponse était la réponse C.

Comme rappelé au début de la section sur la régression de Poisson, l'espérance dans un population de taille 100,000 sera 100,000 fois la valeur prédite,

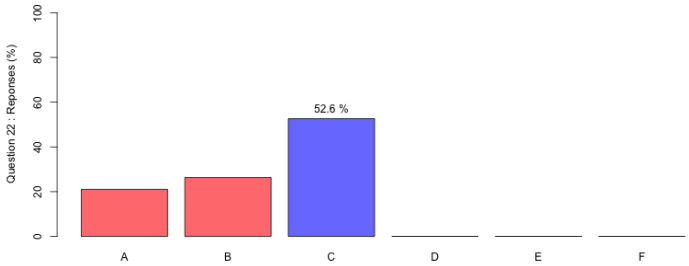
autrement dit

$100,000 \times exp[-15 - 1.2 + 0.15 \cdot x + 0.004 \cdot x^2 + 0.012 \cdot x]$ pour un âge x

soit

$100,000 \times \exp[-15 - 1.2 + 0.15 \cdot 25 + 0.004 \cdot 25^2 + 0.012 \cdot 25] = 100,000 \times \exp[-9.65] = 6.44$

soit plus que 5, mais moins que 7, qui est la réponse C.



23 question 31 (examen CAS)

31.

Given the following information:

- Y is a random variable in the exponential family

$$f(y) = c(y, \phi) * \exp\left[\frac{y\theta - a(\theta)}{\phi}\right]$$

- $a(\theta) = -\sqrt{-2\theta}$
- $\theta = -0.3$
- $\phi = 1.6$

Calculate $E(Y)$.

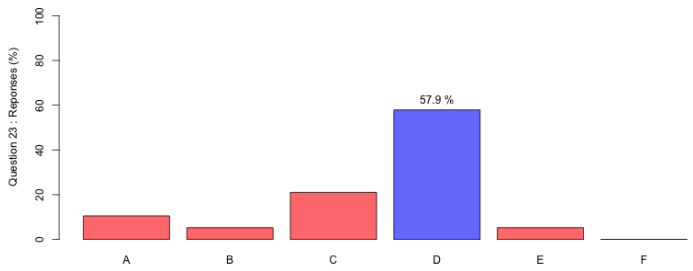
A. Less than -1
B. At least -1, but less than 0
C. At least 0, but less than 1
D. At least 1, but less than 2
E. At least 2

La bonne réponse était la réponse D.

Petite subtilité ici : la fonction notée $b(\cdot)$ dans le cours est ici notée $a(\cdot)$. Mais on ne va pas se laisser démonter pour autant ! Ici, a est en puissance $1/2$, qui correspond à la loi inverse Gaussienne... Bon, une fois qu'on a dit ça, on n'a pas dit grand chose. Par contre, en présentant la famille exponentielle, j'avais mentionné une propriété importante (plus qu'une en fait) : $\mathbb{E}[Y] = a'(\theta)$. Ici, $a(\theta) = -(-2\theta)^{1/2}$ donc $a(\theta) = -(-2) \cdot (1/2) \cdot (-2\theta)^{-1/2}$. Or on nous dit que $\theta = -0.3$ donc en replaçant

$$\mathbb{E}[Y] = a'(-0.3) = \frac{1}{\sqrt{0.6}} = 1.291$$

qui est compris entre 1 et 2, ce qui correspond à la réponse D.



24 question 36 (examen CAS)

36.

You are given the following information for two potential logistic models used to predict the occurrence of a claim:

- Model 1: (AIC = 262.68)

Parameter	$\hat{\beta}$
(Intercept)	-3.264
Vehicle Value ($000s)	0.212
Gender-Female	0.000
Gender-Male	0.727

- Model 2: (AIC = 263.39)

Parameter	$\hat{\beta}$
(Intercept)	-2.894
Gender-Female	0.000
Gender-Male	0.727

- AIC is used to select the most appropriate model.

Calculate the probability of a claim for a male policyholder with a vehicle valued $12,000 by using the selected model.

A. Less than 0.15
B. At least 0.15, but less than 0.30
C. At least 0.30, but less than 0.45
D. At least 0.45, but less than 0.60
E. At least 0.60

La bonne réponse était la réponse D.

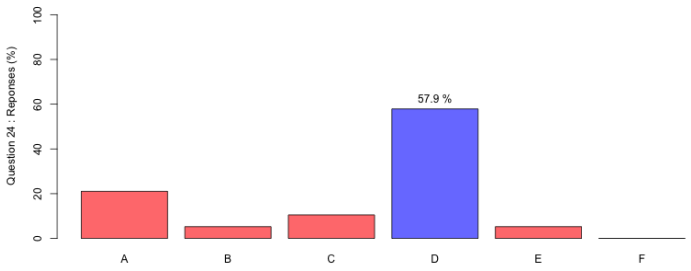
Comme on l'a vu longuement dans la première partie du cours, l'utilisation du AIC comme critère de choix de modèle conduit à retenir celui avec le plus faible AIC. Autrement dit, on retient le modèle 1. La cote est ici

$$\exp[\boldsymbol{x}^{\mathsf{T}}\boldsymbol{\beta}] = \exp[-3.264 + 12 \cdot 0.212 + 0.727] = \exp[0.007] = 1.007$$

On est ici très proche de 1. Dans ce cas, la probabilité sera proche de $1/2$ (cf second question de l'examen). Plus précisément, la probabilité est ici

$$\frac{\exp[0.007]}{1 + \exp[0.007]} = 50.2\%$$

qui est compris entre 45% et 60%, qui est la réponse D.



25 question 32 (examen CAS)

32. A GLM is used to model claim size. You are given the following information about the model:

- Claim size follows a Gamma distribution.
- Log is the selected link function.
- Scale parameter is estimated to be 2.
- Model Output:

Variable	$\hat{\beta}$
(Intercept)	2.32
Location - Urban	0.00
Location - Rural	-0.64
Gender - Female	0.00
Gender - Male	0.76

Calculate the variance of the predicted claim size for a rural male.

A. Less than 25
B. At least 25, but less than 100
C. At least 100, but less than 175
D. At least 175, but less than 250
E. At least 250

La bonne réponse était la réponse E.

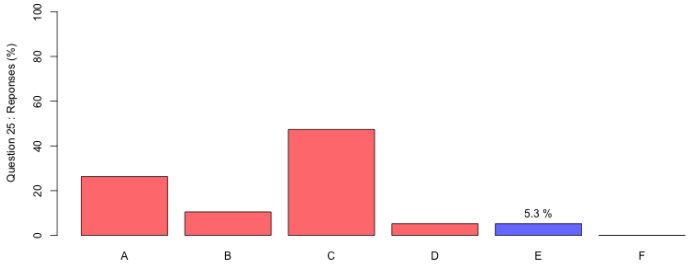
Là encore (je renvoie à la question 19) comme on parle de variance, on a besoin de deux termes : l'espérance sera obtenue facilement avec le lien logarithmique, et la variance sera obtenue avec la fonction variance de la loi Gamma - variance quadratique. On se lance ! Pour l'espérance on a

$$\mathbb{E}[Y|\boldsymbol{x}]\exp[2.32-0.64+0.76]=11.473$$

et on utilise ensuite le fait que

$$\text{Var}[Y|\boldsymbol{x}]=\phi\cdot\mathbb{E}[Y|\boldsymbol{x}]^2=2\cdot 11.473^2=263.3$$

car ici on nous parle de *scale parameter* de 2 (personnellement, je comprends ça comme la valeur de ϕ), qui est supérieur à 250, et donc correspond à la réponse E.



Il y a eu beaucoup (relativement) de réponses C. Je pense que le *scale parameter* avait été oublié.

26 question 31 (examen CAS)

31. Within the context of Generalized Linear Models, suppose that y has an exponential distribution with probability density function expressed as:

$$f(y) = \frac{1}{\mu}\ \exp(-\frac{y}{\mu}); \text{ for } y > 0$$

Determine the variance of y in terms of μ.

A. $1/\mu$ B. $\sqrt{\mu}$ C. μ D. μ^2

E. Cannot be determined from the given information

La bonne réponse était la réponse D.

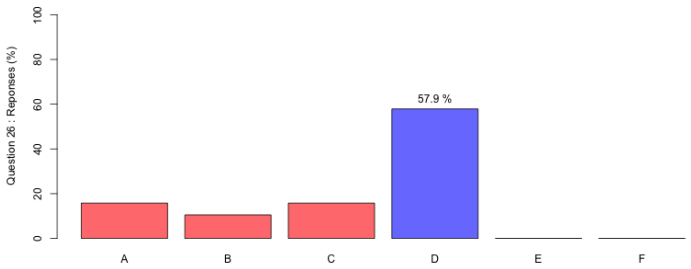
On reconnaît la loi exponentielle de moyenne μ... et de variance μ^2... Il faut vraiment refaire le calcul ? je pourrais renvoyer aux cours de probabilités de première année, et en fait presque tous les cours du programme. Si on veut se raccrocher au cours, on peut utiliser ce qu'on a vu sur la famille exponentielle :

$$f(y) = \exp\left(y \cdot \frac{-1}{\mu} - (-\log\frac{1}{\mu})\right) = \exp[y\theta - a(\theta)] \text{ avec } \theta = \frac{-1}{\mu}$$

et $a(\theta) = -\log(-\theta)$. Or on a vu que la variance d'une variable de densité f était

$$\text{Var}[Y] = \phi a''(\theta) = \frac{1}{\theta^2} = \mu^2$$

car ici $\phi = 1$. Bref, on a la réponse D. En fait, on avait vu en cours qu'une donction variance quadratique correspondant à une loi Gamma... qui est une généralisation de la loi exponentielle... Donc pas grande suprise ici.



27 question 33 (examen CAS)

33.

A distribution belongs to the exponential family if it can be written in the canonical form:

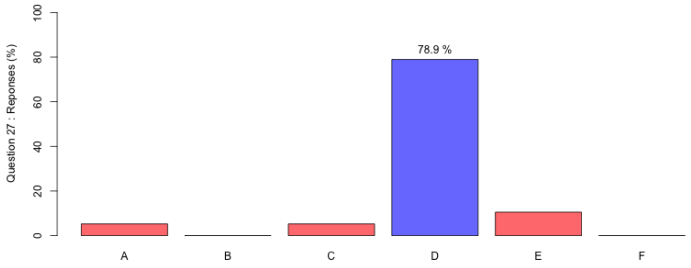
$$f(y;\theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)]$$

Determine the values of $a(y), b(\theta), c(\theta)$, and $d(y)$ for the Poisson distribution.

A. $a(y) = y, b(\theta) = \ln(\theta), c(\theta) = \theta$, and $d(y) = \ln(y!)$
B. $a(y) = -y, b(\theta) = -\ln(\theta), c(\theta) = \theta$, and $d(y) = \ln(y!)$
C. $a(y) = y, b(\theta) = \ln(\theta), c(\theta) = \theta$, and $d(y) = 1/(y!)$
D. $a(y) = y, b(\theta) = \ln(\theta), c(\theta) = -\theta$, and $d(y) = -\ln(y!)$
E. The answer is not given by (A), (B), (C), or (D)

La bonne réponse était la réponse D.

Je le refais ici ? On l'avait vu en cours... Pour rappel, la loi de Poisson, c'est $f(y) = \exp[-\lambda]\lambda^y/y!$. Aussi, D correspond à la loi de Poisson de paramètre $\lambda = \theta$.



28 question 37 (examen CAS)

37. You are given the outputs from two GLMs fitted to the same data from a trial of a new drug.

Model 1		Model 2	
Response variable	Number	Response variable	Number
Response distribution	Poisson	Response distribution	Negative binomial
Link	Log	Link	Log
AIC	273.877	AIC	164.880

Parameter	$\hat{\beta}$	s. e. ($\hat{\beta}$)	Parameter	$\hat{\beta}$	s. e. ($\hat{\beta}$)
Intercept	4.529	0.147	Intercept	4.526	0.595
Treatdrug			Treatdrug		
Placebo	0.000	0.000	Placebo	0.000	0.000
Drug	-1.359	0.118	Drug	-1.368	0.369
Age	-0.039	0.006	Age	-0.039	0.021

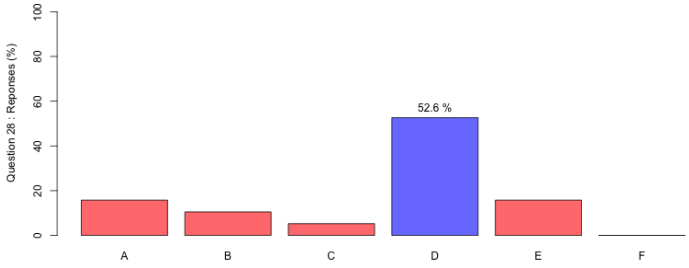
Determine which of the following statements is false using the Wald test.
A. Under Model 1, the Treatdrug coefficient has a p-value less than 0.01.
B. Under Model 1, the Age coefficient has a p-value less than 0.01.
C. Under Model 2, the Treatdrug coefficient has a p-value less than 0.01.
D. Under Model 2, the Age coefficient has a p-value less than 0.01.
E. Under both models, the intercept coefficient has a p-value of less than 0.01.

La bonne réponse était la réponse D.

J'avais évoqué dans une question précédante la statistique de Wald, qui est juste le carré de la statistique z utilisée pour un test de significativité. Comme je ne donnais pas la table de la loi du chi-deux, on va simplement faire un test z. Complétons ici la table comme une sortie de régression classique, pour nos trois paramètres, et nos deux modèles :

	$\widehat{\beta}$	$se(\widehat{\beta})$	z	p-value
Model I				
(intercept)	4.529	0.147	30.809	<0.01%
treat:drug	-1.359	0.118	-11.517	<0.01%
age	-0.039	0.006	-6.500	<0.01%
Model II				
(intercept)	4.526	0.595	7.607	<0.01%
treat:drug	-1.368	0.368	-3.717	0.02%
age	-0.039	0.021	-1.857	6.33%

Si la première page ne permettait pas d'avoir la dernière colonne, on notera quand même qu'avec une stastique de l'ordre -1.85 on est au delà d'une p-value à 5%. La bonne réponse était donc la réponse D.



29 question 37 (examen CAS)

37.

Let $Y_1, \ldots, Y_n$ be independent Poisson random variables, each with respective mean μ_i for $i = 1, 2, \ldots, n$, where:

$$\ln(\mu_i) = \begin{cases} \alpha, & \text{for } i = 1,2, \ldots, m \\ \beta, & \text{for } i = m+1, m+2, \ldots, n \end{cases}$$

The claims experience for a portfolio of insurance policies with $m = 50$ and $n = 100$ is:

$$\sum_{i=1}^{50} y_i = 563$$

$$\sum_{i=51}^{100} y_i = 1{,}261$$

Denote by $\hat{\alpha}$ and $\hat{\beta}$ the maximum likelihood estimates of α and β, respectively.

Calculate the ratio $\frac{\hat{\alpha}}{\hat{\beta}}$.

A. Less than 0.40
B. At least 0.40, but less than 0.60
C. At least 0.60, but less than 0.80
D. At least 0.80, but less than 1.00
E. At least 1.00

La bonne réponse était la réponse C.

Pour la distribution de Poisson, la méthode du maximum de vraisemblance coïncide avec la méthode des moments. Autrement dit, $\widehat{\lambda}$ est la moyenne empirique. Sur les $m = 50$ premières observations,

$$\widehat{\lambda}_1 = \frac{563}{50} = 11.26 \text{ donc } \widehat{\alpha} = \log[\widehat{\lambda}_1] = 2.421$$

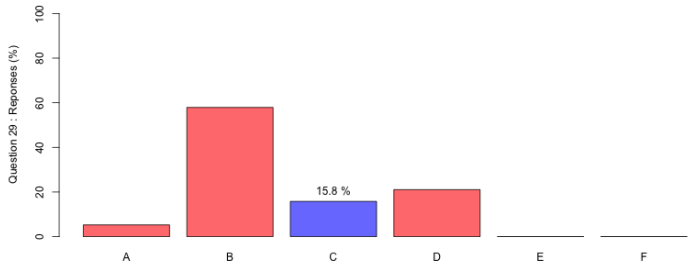
alors que pour les $n - m = 50$ suivantes,

$$\widehat{\lambda}_2 = \frac{1261}{50} = 25.22 \text{ donc } \widehat{\beta} = \log[\widehat{\lambda}_2] = 3.228$$

de telle sorte que

$$\frac{\widehat{\alpha}}{\widehat{\beta}} = \frac{2.421}{3.228} = 0.750$$

qui est supérieur à 0.6 mais infŕieur à 0.8, qui était la réponse C.



Je ne sais pas pourquoi autant de réponses B ont été données ici... si ce n'est que $563/1261 \sim 0.4464711$ est dans l'intervalle $[0.4; 0; 6]$. Mais ça n'a aucun rapport avec la question posée, qui était peu ambigüe.

30 question 32 (examen CAS)

32.

Given a family of distributions where the variance is related to the mean through a power function:

$$\text{Var}[Y] = \text{aE}[Y]^p$$

One can characterize members of the exponential family of distributions using this formula.

You are given the following statements on the value of p for a given distribution:

I. Normal (Gaussian) distribution, $p = 0$
II. Compound Poisson–gamma distribution, $1 < p < 2$
III. Inverse Gaussian distribution, $p = -1$

Determine which of the above statements are correct.

A. I only
B. I and II only
C. I and III only
D. II and III only
E. The answer is not given by (A), (B), (C), or (D)

La bonne réponse était la réponse B.

On l'avait vu en cours : pour la loi normale, on a un modèle homoscédastique, variance constante, donc $p = 0$. Le cas $p \in (0, 2)$ correspond au modèle Tweedie qui est une loi Poisson composée avec des sauts Gamma. On obtenait les cas limites avec Poisson ($p = 1$) et Gamma ($p = 2$). J'avais dit que le cas $p = 3$ correspond au cas Inverse Gaussien (on l'avait fait sur R au dernier cours). Autrement dit III. est fausse.