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
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

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

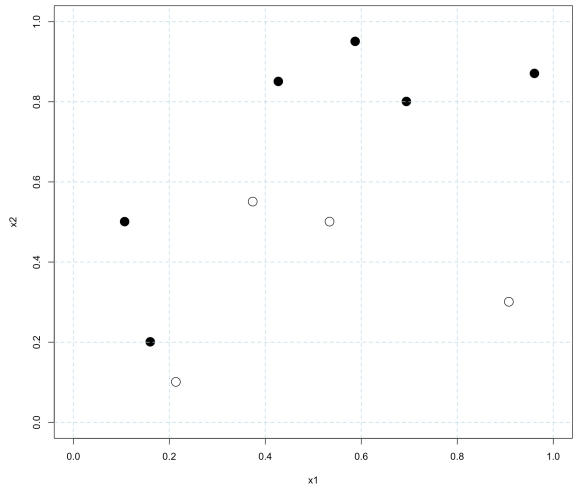
Règlement no 18 *Tout acte de plagiat, fraude, copiage, tricherie, falsification de document ou création d'un faux document commis par une candidate, un candidat, une étudiante, un étudiant, de même que toute participation à ces actes ou tentative de les commettre, à l'occasion d'un examen, d'un travail ou d'un stage faisant l'objet d'une évaluation ou dans toute autre circonstance, constitue une infraction au sens de ce règlement.* (a. 2.1)
https://r18.uqam.ca/

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 :



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

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

C) entre 0 et 1

D) entre 1 et 2

E) plus que 2

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

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

```
Call:
glm(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

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

```
Call:
glm(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

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

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(y ~ x1 + x2, family = binomial(link = "logit"))

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

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

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

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

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

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

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

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

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 1

B) entre 1 et 2

C) entre 2 et 3

D) entre 3 et 4

E) plus de 4

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

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

13 On essaye de modéliser la probabilité qu'une famille reste assuré l'an prochain (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%

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

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

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 $\log \mu = \log \mathbb{E}[Y|X = x] = \beta_0 + \beta_1 x$. 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 $\widehat{\beta}_2$ 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

17 On dispose de la sortie de régression suivante

```
Call:
glm(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

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}

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(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 ($\widehat{\mu}$) 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

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

21 On dispose de la sortie de régression suivante

```
Call:
glm(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

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(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) plus de 250

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

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

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

26 On dispose de la sortie de régression suivante

```
Call:
glm(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

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)

28 On dispose d'un premier modèle GLM, et on 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

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(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

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

```
Call:
glm(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

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