STT5100 - Automne 2019 - Examen Final (GLM)

Arthur Charpentier

Examen **A**

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

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

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

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

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

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

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

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

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

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

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

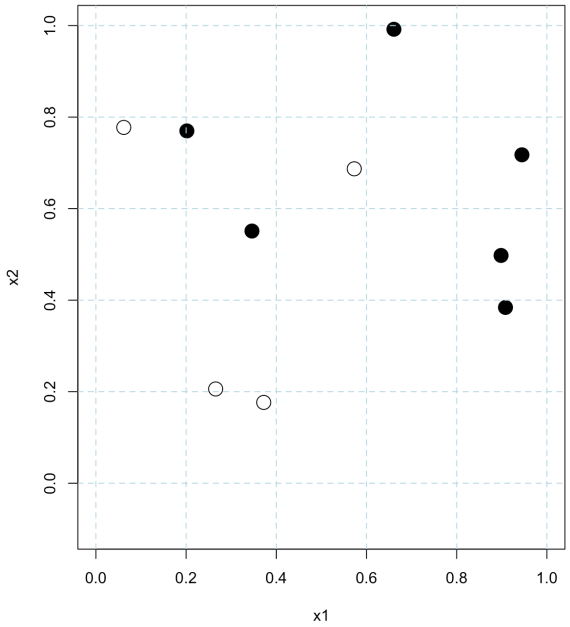
On appelle logit la fonction $(0,1) \mapsto \mathbb{R}$ définie par $\text{logit}(p) = \log(p/(1-p))$. Et la *prime pure* désigne l'espérance mathématique de la perte.

Code permanent :

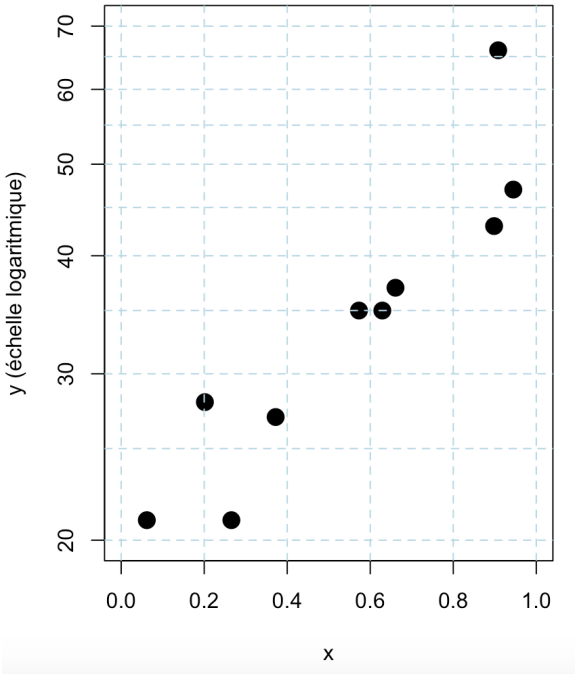
Sujet : A

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

question 11 :



question 17 :



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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

A) moins de 85

B) entre 85 et 95

C) entre 95 et 105

D) entre 105 et 115

E) plus de 115

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

A) moins de 40%

B) entre 40% et 50%

C) entre 50% et 60%

D) entre 60% et 80%

E) plus de 80%

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

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

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

(3) un modèle saturé a une déviance nulle

A) aucune affirmation n'est juste

B) (1) et (2) sont justes

C) (1) et (3) sont justes

D) (2) et (3) sont justes

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

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

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

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

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

A) moins de 5%

B) entre 5% et 10%

C) entre 10% et 15%

D) entre 15% et 20%

E) plus de 20%

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

A) μ

B) μ^2

C) $1/\mu$

D) 1

E) e^{μ}

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

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

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

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

et le modèle (2)

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

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.2665     0.4266   5.313 4.89e-07 ***
x             0.1974     0.7035   0.266     0.79
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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

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

A) -6

B) -2.25

C) 0

D) 0.75

E) 6.75

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

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

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

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

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

A) moins de 200

B) entre 200 et 225

C) entre 225 et 250

D) entre 250 et 275

E) plus de 275

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

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

Coefficients
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.663
youth1+        0.132
old2+         -0.031
```

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

A) moins de 1%

B) entre 1% et 2%

C) entre 2% et 3%

D) entre 3% et 4%

E) plus de 4%

11 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)    4.542      3.254   1.396    0.163
x1            -5.070      3.246  -1.562    0.118
x2            -4.539      4.224  -1.074    0.283
```

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

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

L'énoncé suivant est relatif aux questions 12 et 13

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

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

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

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

On dispose de la sortie suivante

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

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

A) 0.2

B) 1.2

C) 2.4

D) 4.8

E) 6.0

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

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

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

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

A) (i) seulement

B) (ii) seulement

C) (i) et (ii)

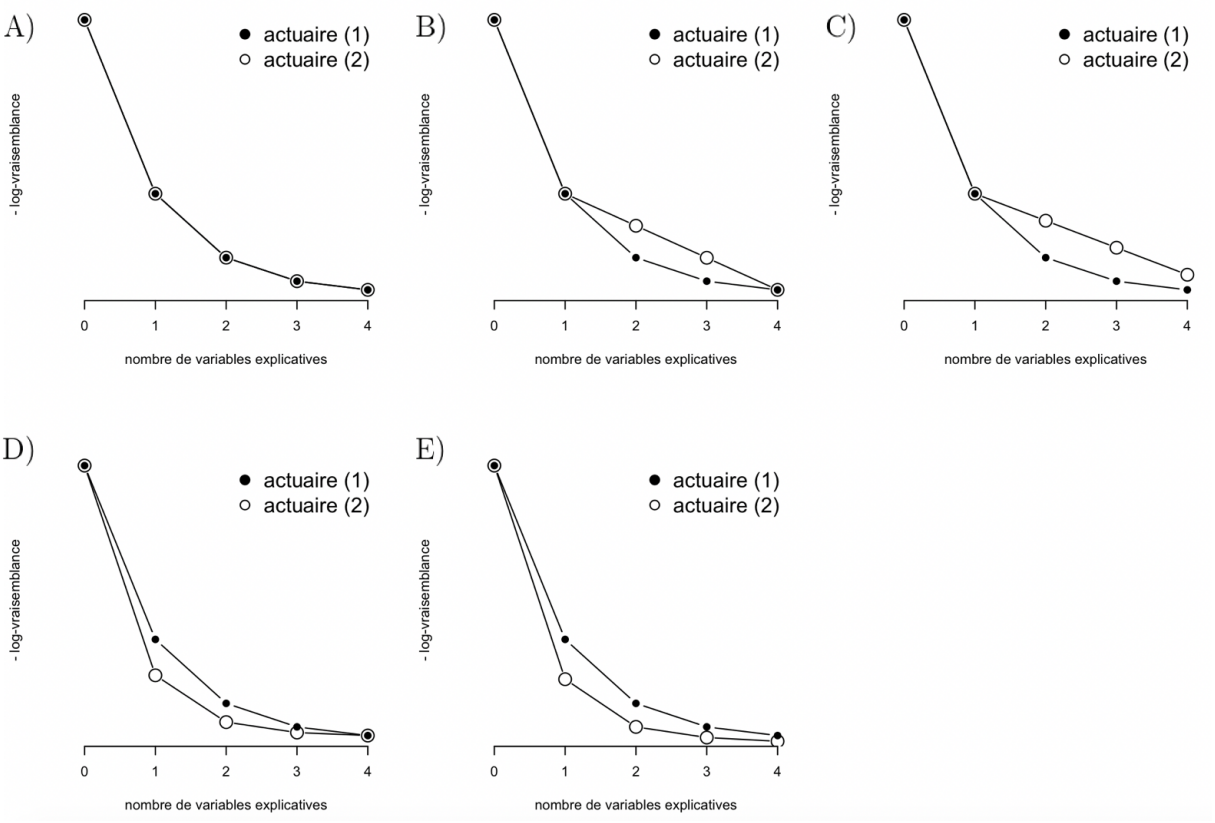
D) (i) et (iii)

E) (ii) et (iii)

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

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

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



15 On nous donne la densité suivante pour Y,

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

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

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

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

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

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

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

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

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

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

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

A) (i) seulement

B) (ii) seulement

C) (iii) seulement

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

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

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

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

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

(Dispersion parameter for poisson family taken to be 1)

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

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

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

A) 1.1

B) 1.3

C) 1.5

D) 1.7

E) 1.9

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

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

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

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

A) moins de 60%

B) entre 60% et 70%

C entre 70% et 80%

D) entre 80% et 90%

E) au moins 90%

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

A) 20θ

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

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

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

E) $20\theta^2$

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

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

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

A) 40%

B) 50%

C) 60%

D) 70%

E) 80%

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

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

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

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

On dispose de la sortie de régression suivante

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

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

(Dispersion parameter for quasipoisson family taken to be ? )

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

Number of Fisher Scoring iterations: 5
```

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

A) moins de 0.9

B) entre 0.9 et 1

C) exactement 1

D) entre 1 et 2

E) plus de 2

23 On dispose de la sortie de régression suivante

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

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

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

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

A) moins de 55

B) entre 55 et 60

C) entre 60 et 65

D) entre 65 et 70

E) plus de 70

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

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

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

A) 46%

B) 50%

C) 54%

D) 58%

E) 62%

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

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

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

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

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

E) aucune des propositions

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

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

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

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

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

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

A) moins de 35%

B) entre 35% et 45%

C) entre 45% et 55%

D) entre 55% et 65%

E) plus de 65%

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

A) moins de -5

B) entre -5 et -4.5

C) entre -4.5 et -4

D) entre -4 et -3.5

E) plus de -3.5

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

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

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

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

(Dispersion parameter for poisson family taken to be 1)

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

et le modèle (13) est

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

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

(Dispersion parameter for poisson family taken to be 1)

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

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

A) moins de 0

B) entre 0 et 10

C) entre 10 et 20

D) entre 20 et 50

E) plus de 50

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

A) $[62.3, 63.9]$

B) $[60, 66]$

C) $[55, 70]$

D) $[48, 80]$

E) $[30, 100]$

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

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

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

A) moins de 0.9

B) entre 0.9 et 1.1

C) entre 1.1 et 2.0

D) entre 2 et 10

E) plus de 10