EXAMEN INTRA, ACT2040, HIVER 2013

Les calculatrices[1] sont autorisées, mais pas les téléphones 'intelligents' (qui devront être rangés pendant toute la durée de l'épreuve). Tous les documents sont interdits.

Dans les feuilles qui suivent, il y a 34 questions,

- 19 questions générales sur la régression logistique (11 questions), et la régression de Poisson (8 questions)
- 15 questions portant sur la modélisation du nombre d'aventures extra-conjugales hétérosexuelles (sorties en annexes)

Précisions sur la notation

Pour chaque question à choix multiple, quatre réponses sont proposées, une seule est valide, et vous ne devez en retenir qu'une (au maximum),

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

Aucune justification n'est demandée.

Pour chaque question qui n'est pas à choix multiple, une courte réponse est demandée. Elle doit tenir dans l'espace réservé. Vous gagnez 1 point si la réponse est claire, et correcte.

Votre note finale est le total des points (sur 34). Je ne récupère que la feuille séparée (comportant votre nom et l'espace pour mettre les réponses).

Petit complément

Dans le tableau ci-dessous figurent quelques valeurs tirées de la Table de la loi normale centrée réduite. Il s'agit de valeurs liées aux quantiles, au sens où

$$\mathbb{P}(Z > z_p) = p \text{ où } Z \sim \mathcal{N}(0,1).$$

p	20%	15%	10%	5%	2.5%	1%	0.5%	0.1%
z_p	0.8416	1.036	1.28	1.645	1.960	2.326	2.576	3.090

[1] BA-35, BA II Plus, TI-30X, TI-30Xa, TI-30XIIS et TI-30XIIB (cf plan de cours).

1. Compréhension générale du cours

On obtenu la sortie suivante suite à une estimation de modèles logistiques. Les questions **1** à **8** portent sur cette sortie

```
> regH = glm(Y~(X1=="H")+X2,family=binomial(link="logit"))
> summary(regH)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.70444    0.56621 ■■■■■ ■■■■■ ■■
X1 == "H"TRUE -3.06984    0.50396  -6.091 1.12e-09 ***
X2            -0.02074    0.01105  -1.876   0.0607 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Number of Fisher Scoring iterations: 6

> regF = glm(Y~(X1=="F")+X2,family=binomial(link="logit"))
> summary(regF)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    ■■■■■   ■■■■■  -3.567 0.000361 ***
X1 == "F"TRUE  3.06984    0.50396   6.091 1.12e-09 ***
X2            -0.02074    0.01105  -1.876 0.060693 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Number of Fisher Scoring iterations: 6
```

Question 1. On souhaite faire une prévision pour `X1="H"` et `X2=25` avec le modèle `regH`. Quelle prévision pour $\mathbb{E}(Y|X_1 =$`"H"`$, X_2 =$`25`$)$ feriez vous ?

A. -2.8838
B. 0.5296
C. 0.0529
D. 0.2883

Question 2. Que deviendrait la prévision si `X1` prenait non plus la valeur `"H"` mais la valeur `"F"` (on a toujours `X2=25`) ?

A. 0.5463
B. 0.1860
C. 0.0546
D. 0.0180

Question 3. Dans la sortie `regH`, qu'a été effacé dans la 3ème colonne (`z value`) ?

A. 1.244
B. 0.704
C. 2.197
D. -3.567

Question 4. Dans la sortie `regH`, qu'a été effacé dans la 4ème colonne (`Pr(>|z|)`) ?

A. 0.03%
B. 89.25%
C. 21.50%
D. 10.75%

Question 5. Dans la sortie `regH`, qu'a été effacé dans la 4ème colonne (sans nom) ?

A. `***`
B. `*`
C. `.`
D. rien

Question 6. Dans la sortie `regF`, qu'a été effacé dans la 1ère colonne (`Estimate`) ?

A. 2.3654
B. -0.70444
C. 0.70444
D. -2.3654

Question 7. Dans la sortie `regF`, qu'a été effacé dans la 2ème colonne (`Std. Error`) ?

A. 1.50799
B. -0.66313
C. 0.66313
D. 0.19748

Question 8. Dans la zone réservée sur la feuille de réponse, expliquez la phrase

```
Number of Fisher Scoring iterations: 6
```

Question 9. Lors de la modélisation à l'aide d'une régression logistique, on a prédit 0.314 pour $\mathbb{E}(Y|X_1, X_2)$. Que peut-on prédire pour $\text{Var}(Y|X_1, X_2)$

A. 0.215
B. 0.314
C. 0.457
D. on ne peut pas faire de prévision, il manque des informations

On obtenu la sortie suivante suite à une estimation de d'un modèle logistique (sur une autre variable d'intérêt).

```
> reg=glm(Z~0+X1+X2,family=binomial(link="logit"),data=base)
> summary(reg)

Coefficients:
    Estimate Std. Error z value Pr(>|z|)
X1F -6.14096    0.77525  -7.921 2.35e-15 ***
X1H -5.46391    0.69995  -7.806 5.90e-15 ***
X2   0.09192    0.01207   7.614 2.66e-14 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)
```

On a alors tenté de représenter graphiquement les deux prédictions,

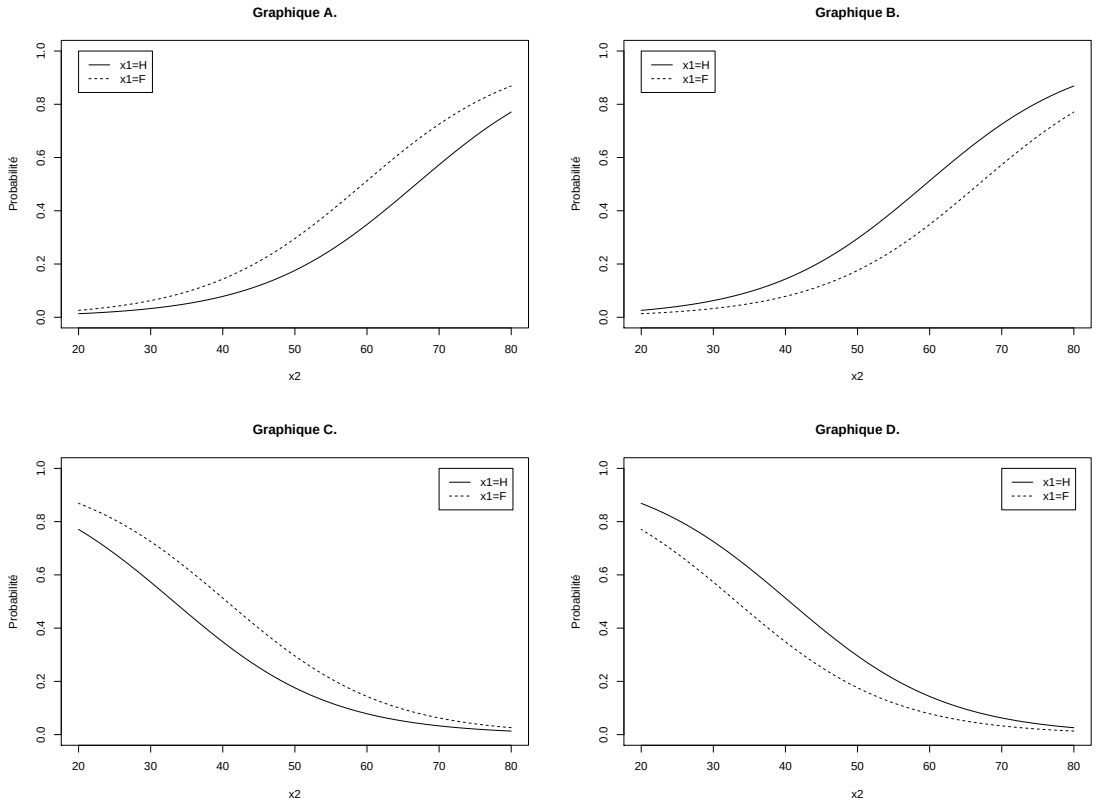
$$x_2 \mapsto \mathbb{E}(Z|X_1 = x_1, X_2 = x_2)$$

lorsque x_1 vaut `H` (en trait continu) et lorsque x_1 vaut `F` (en trait pointillé)

```
> ZH=predict(reg, newdata=data.frame(X1="H", X2=seq(20,80)), type="response")
> ZF=predict(reg, newdata=data.frame(X1="F", X2=seq(20,80)), type="response")
> plot(seq(20,80),ZH,type="l",lty=1)          # trait continu
> lines(seq(20,80),ZF,lty=2)                  # trait pointille
```

Question 10. Quel graphique ci-dessous correspond à ce qui a été demandé ?

A. le graphique A.
B. le graphique B.
C. le graphique C.
D. le graphique D.



Question 11. Différentes écritures pour la régression de Poisson sont proposées

(i) $N_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ avec ε_i i.i.d. de loi $\mathcal{P}(\lambda)$

(ii) $N_i = \exp[\beta_0 + \beta_1 X_i] + \varepsilon_i$ avec ε_i i.i.d. de loi $\mathcal{P}(\lambda)$

(iii) $N_i = \exp[\beta_0 + \beta_1 X_i + \varepsilon_i]$ avec ε_i i.i.d. de loi log-Poisson

(iv) $\log N_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ avec ε_i i.i.d. de loi log-Poisson

Quelle(s) écriture correspond à ce qui a été appelé 'régression log-Poisson' dans le cours

A. (i)

B. (ii)

C. (iii) et (iv)

D. aucune des formes proposées

On obtenu la sortie suivante suite à une estimation de régressions de Poisson (les variables `X1` et `X2` sont les mêmes que dans les questions précédantes). Les questions **12** à **19** portent sur cette sortie

```
> mean(base[["N"]])
[1] ■■■■■
> mean(base[X1=="H","N"])
[1] 0.2361111
> mean(base[X1=="F","N"])
[1] 0.1145833
> sum(base[["X1"]]=="H")
[1] 144
> sum(base[["X1"]]=="F")
[1] 96
> reg=glm(N~(X1=="H"),family=poisson(link="log"),data=base)
> summary(reg)
Coefficients:
              Estimate Std. Error  z value   Pr(>|z|)
(Intercept)    ■■■■■     0.3015  ■■■■■   ■■■■■  ■■
X1 == "H"TRUE  ■■■■■     0.3469  ■■■■■   ■■■■■  ■■
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
(Dispersion parameter for poisson family taken to be 1)
> regq=glm(Y~(X1=="H"),family=quasipoisson(link="log"),data=base)
> summary(regq)
Coefficients:
              Estimate  Std. Error   z value  Pr(>|z|)
(Intercept)    ■■■■■    ■■■■■  ■■■■■  3.45e-11  ■■
X1 == "H"TRUE  ■■■■■    ■■■■■  ■■■■■    0.0449  ■■
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
(Dispersion parameter for quasipoisson family taken to be 1.068463)
```

Question 12. Quelle est la première valeur manquante (calcul de `> mean(base[["N"]])`)

A. 0.1631
B. 0.1875
C. 0.1753
D. on n'a pas assez d'éléments pour conclure

Question 13. Dans la sortie de régression `reg`, quelle est la valeur de `estimate` associée au coefficient `(Intercept)`

A. 0.114
B. 0.175
C. -1.740
D. -2.166

Question 14. Dans la sortie de régression `reg`, quelle est la valeur de `estimate` associée au coefficient `X1 == "H"TRUE`

A. 0.121
B. 0.236
C. -2.112
D. 0.723

Question 15. Dans la sortie de régression, quels coefficients sont significativements non nuls (avec une probabilité de 95%)

A. aucun
B. `(Intercept)` seulement
C. `X1 == "H"TRUE` seulement
D. `(Intercept)` et `X1 == "H"TRUE`

Question 16. Dans la sortie de régression `regq`, quelle est la valeur de `estimate` associée au coefficient `(Intercept)`

A. 0.114
B. -2.313
C. -1.740
D. -2.166

Question 17. Dans la sortie de régression `regq`, quelle est la valeur de `estimate` associée au coefficient `X1 == "H"TRUE`

A. -2.112
B. -0.252
C. 0.252
D. 0.723

Question 18. Dans la sortie de régression `regq`, quelle est la valeur de `Std. Error` associée au coefficient `(Intercept)`

A. 0.114
B. 0.311
C. 0.301
D. 0.322

Question 19. Dans la sortie de régression `regq`, quelle est la valeur de `Std. Error` associée au coefficient `X1 == "H"TRUE`

A. 0.335
B. 0.347
C. 0.370
D. 0.358

2. Analyse des données

Question 20. A l'aide de la **sortie 1**, si on souhaite modéliser Y, le nombre d'aventures extraconjugales dans l'année suivant une loi de Poisson, $\mathcal{P}(\lambda)$, que vaudrait $\widehat{\lambda}$, estimateur du maximum de vraisemblance de λ ?

A. 1.879
B. -0.461
C. 0.631
D. 0.750

Question 21. A l'aide du modèle prédédant, donner un estimateur de $\mathbb{P}(Y > 0)$, la probabilité qu'une personne (prise au hasard) ait une - ou plusieurs - aventure(s) extraconjugale(s) dans l'année

A. 80.1%
B. 19.9%
C. 24.9%
D. 46.7%

Question 22. Toujours à l'aide de la **sortie 1**, si on souhaite modéliser Y_0, le fait qu'une personne ait eu - ou pas ($Y_0 = 1$ si elle en a eu, $Y_0 = 0$ si elle n'en a pas eu) - des aventures extraconjugales dans l'année suivant une loi de Bernoulli, $\mathcal{B}(\pi)$, que vaudrait $\widehat{\pi}$, estimateur du maximum de vraisemblance de π ?

A. 80.1%
B. 19.9%
C. 24.9%
D. 46.7%

Question 23. A l'aide du modèle `regbernoulli` de la **sortie 2**, prédire la probabilité qu'un homme marié depuis 10 ans ait une aventure extraconjugale (ou plus) dans l'année.

A. 2.36%
B. 23.4%
C. 42.5%
D. 99.5%

Question 24. A l'aide du modèle `regpoisson` de la **sortie 2**, prédire la probabilité qu'un homme marié depuis 10 ans ait une aventure extraconjugale (ou plus) dans l'année.

A. 1.76%
B. 23.4%
C. 42.5%
D. 98.7%

Question 25. Dans le modèle `regpoisson2` de la **sortie 3**, peut-on éliminer la variable `SEXE` ?

A. non, car seuls les hommes ont été pris en compte (variable `SEXEH`), il faudrait faire la même régression en prenant en compte les femmes (variable `SEXEF`)
B. non, car la variable était significativement non nulle dans les précédantes régressions (sortie 2)
C. oui, car la modalité homme (variable `SEXEH`) n'est pas significativement différente de la modalité femme (correspondant ici la modalité de référence)
D. oui, car comme une homme trompe sa femme avec une autre femme, et qu'une femme trompe son mari avec un autre homme: l'effet de la variable `SEXE` s'annule.

Question 26. Dans le modèle `regpoisson2` de la **sortie 3**, donnez un intervalle de confiance à 95% pour β associé à la variable `EDUCATION` ?

A. [0.098;0.151]
B. [-0.199;0.449]
C. [1.074;1.193]
D. [0.072;0.177]

Question 27. Dans la **sortie 4**, plusieurs variables sont utilisées dans la régression, à savoir `YEARMARRIAGE`, `AGE`, `RELIGIOUS`, `EDUCATION` et `SATISFACTION`, toutes prises ici comme des variables numériques. On considère une personne de 35 ans (`AGE=35`), mariée depuis 10 ans (`YEARMARRIAGE=10`), athée (`RELIGIOUS=1`) et de degré d'éducation élevé (`EDUCATION=20`). On ne connait pas son degré de satisfaction dans son mariage (`SATISFACTION`) qui peut varier de 1 à 5. Quel peut être l'intervalle pour λ pour cette personne,

A. [1.07;4.84]
B. [0.84;1.07]
C. [1.07;18.44]
D. [0.84;12.24]

Question 28. Dans la **sortie 5**, on se demande si la variable `SATISFACTION` peut être considérée comme une variable continue, ou s'il faut la prendre en compte comme facteur. Qu'en pensez vous (une seule réponse est autorisée) ?

A. les réponses étant 1,2,3,4 et 5, la variable est une variable *numérique*, donc on peut la considérer comme numérique, et la prendre en tant que facteur n'a pas de sens.
B. les valeurs $\widehat{\beta} \cdot x$ (x désignant le niveau de satisfaction) sont dans les intervalles de confiance des $\widehat{\beta}_k^x$ (variables prises en tant que facteur) donc on peut considérer la variable comme numérique car cela réduit le nombre de variables dans le modèle
C. dans la régression `regpoisson3` trop de variables sont non-significatives: on ne peut pas utiliser des facteurs, il faut donc prendre la variable en tant que variable numérique
D. dans la régression `regpoisson3` on note que les facteurs ne sont pas monotones: il y a un effet non linéaire, et il ne faut pas utiliser la variable en tant que variable numérique mais en tant que facteur (ou alors il faudrait lisser la variable)

Question 29. Dans la **sortie 6**, on ajuste une régression quasi-Poisson (appelée `regqpoisson`) et une régression binomiale négative (appelée `regnegbin`). Pensez-vous que le coefficient de surdispersion φ soit strictement plus grand que 1 ?

A. oui car dans `regqpoisson`, $\widehat{\phi} = 3.5285$ et $3.5285 > 1$
B. oui car dans `regqpoisson`, $\widehat{\phi} = 3.5285$ et $3.5285 > 1.960$
C. oui car dans `regnegbin`, θ est significativement non nul ($0.1610/0.0232 > 1.960$) et le test construit pour le paramètre de surdispersion repose sur une modélisation binomiale négative
D. non car dans `regnegbin`, le paramètre de surdispersion (appelé `Dispersion parameter`) est 1

Question 30. On considère une personne de 35 ans (`AGE=35`), marié depuis 10 ans `YEARMARRIAGE=10`, athée (`RELIGIOUS=1`), de degré d'éducation élevé (`EDUCATION=20`) et qui se déclare heureuse en mariage (`SATISFACTION=5`). On notera $\boldsymbol{x}$ ces caractéristiques, et N le nombre d'aventures extraconjugales eu par cette personnes pendant l'année passée. A l'aide du modèle `regqpoisson`, donnez un estimateur pour $\text{Var}(N|\boldsymbol{X} = \boldsymbol{x})$,

A. 3.788
B. 2.016
C. 1.944
D. 10.26

Question 31. [*suite de la question 30*]. Pour cette même personne, à l'aide du modèle `regnegbin`, donnez un estimateur pour $\mathbb{E}(N|\boldsymbol{X} = \boldsymbol{x})$,

A. 0.65
B. 1.04
C. 1.07
D. 4.82

Question 32. Dans un modèle à inflation de zéros, si $p_i(\cdot)$ est une fonction de probabilité sur $\mathbb{N}$, on suppose que pour $\forall i = 1, \cdots, n$,

$$\mathbb{P}(Y_i = k) = \begin{cases} \pi_i + [1-\pi_i] \cdot p_i(0) \text{ si } k = 0, \\ [1-\pi_i] \cdot p_i(k) \text{ si } k = 1, 2, \cdots \end{cases} \tag{2.1}$$

On dispose d'observations (X_i, Y_i) indépendantes. Dans un modèle où π_i est identique pour tous (noté π), et où $p_i(\cdot)$ est donné par un modèle log-Poisson, écrire la vraisemblence $\mathcal{L}(\beta_0, \beta_1; \boldsymbol{X}, \boldsymbol{Y})$ dans l'espace réservé sur la feuille de réponses.

Question 33. Dans la **sortie 7**, on se demande si la surdispersion ne pourrait pas venir d'une sur-représentation des 0 dans notre base de données, qui pourrait être interprété comme un mensonge. Dans un modèle à inflation de zéro (équation (2.1)), où π_i serait considéré comme constant (noté π), quel estimateur pour π suggèreriez-vous ?

A. 4.5%
B. 18.6%
C. 50.2%
D. 77.7%

Question 34. Dans la **sortie 8**, on suppose que π_i dépend de la variable `SATISFACTION` de l'individu i (au travers d'une régression logistique). Comparer π_i pour une personne malheureuse en amour (`SATISFACTION=1`), noté π_1 et pour une personne heureuse en amour (`SATISFACTION=5`), noté π_5

A. $\pi_1/\pi_5 \sim 40\%$
B. $\pi_1/\pi_5 \sim 70\%$
C. $\pi_1/\pi_5 \sim 90\%$
D. $\pi_1/\pi_5 \sim 120\%$