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

Il s'agit d'une régression logistique, donc

$$\pi(x_1, x_2) = \mathbb{E}(Y|X_1 = x_1, X_2 = x_2) = \frac{\exp[\beta_0 + \beta_1 \mathbf{1}(x_1 = H) + \beta_2 x_2]}{1 + \exp[\beta_0 + \beta_1 \mathbf{1}(x_1 = H) + \beta_2 x_2]}$$

en utilisant les notations usuelles. Aussi la prévision est ici

$$\widehat{\pi}(x_1, x_2) = \frac{\exp[\widehat{\beta}_0 + \widehat{\beta}_1 \mathbf{1}(x_1 = H) + \widehat{\beta}_2 x_2]}{1 + \exp[\widehat{\beta}_0 + \widehat{\beta}_1 \mathbf{1}(x_1 = H) + \widehat{\beta}_2 x_2]}$$

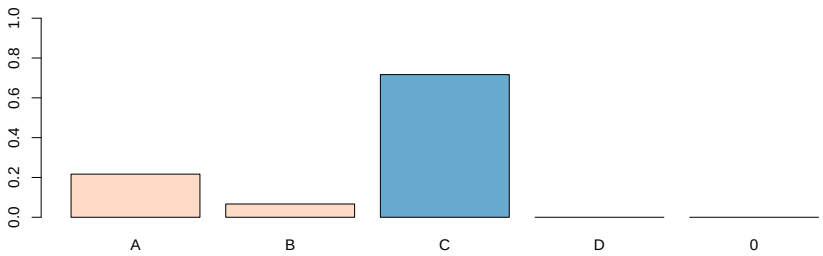
soit

$$\widehat{\pi}(x_1, x_2) = \frac{\exp[0.7044 - 3.0698 \mathbf{1}(x_1 = H) - 0.0207 x_2]}{1 + \exp[0.7044 - 3.0698 \mathbf{1}(x_1 = H) - 0.0207 x_2]}$$

Ici, on souhaite une prévision pour $x_1 = H$ et $x_2 = 25$, i.e.

$$\widehat{\pi}(x_1, x_2) = \frac{\exp[0.7044 - 3.0698 \times 1 - 0.0207 \times 25]}{1 + \exp[0.7044 - 3.0698 \times 1 - 0.0207 \times 25]} \sim \frac{\exp[-2.8829]}{1 + \exp[-2.8829]} = 0.053,$$

ce qui correspond à la réponse C.



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

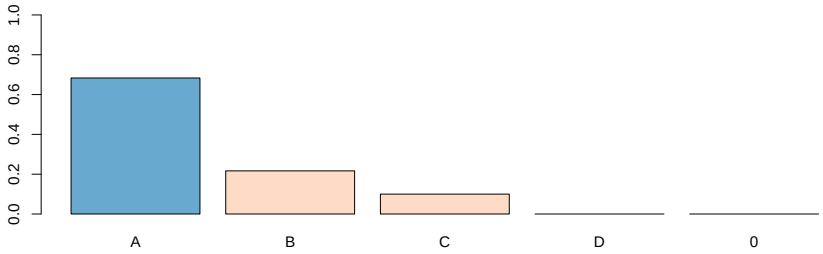
C. 0.0546

D. 0.0180

Il suffit de faire un petit changement dans le calcul précédant,

$$\widehat{\pi}(x_1, x_2) = \frac{\exp[0.7044 - 3.0698 \times 0 - 0.0207 \times 25]}{1 + \exp[0.7044 - 3.0698 \times 0 - 0.0207 \times 25]} \sim \frac{\exp[0.1869]}{1 + \exp[0.1869]} = 0.5463,$$

qui correspond à la réponse A.



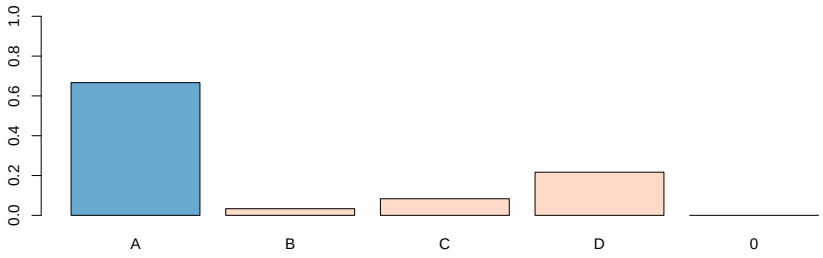
Question 3. Dans la sortie `regH`, qu'a été effacé dans la 3ème colonne (`z value`) ?

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

La z-value est la statistique obtenue pour un test de significativité (équivalent de la t-value dans un modèle linéaire: dans un modèle linéaire, sous hypothèse normalité, les estimateurs sont Gaussien, donc le ratio entre la valeur estimée et l'*estimateur* de son écart-type suit une loi de Student. Ici, on n'a plus cette propriété, mais asymptotiquement, comme on estime à l'aide du maximum de vraisemblance, on récupère la normalité asymptotique. Bref, on continue à utiliser le ratio entre la valeur estimée et l'*estimateur* de son écart-type, qui va suivre asymptotiquement une loi normale centrée réduite). Toute cette histoire pour dire que

$$z = \frac{\widehat{\beta}_0}{\sqrt{\widehat{\mathrm{Var}}(\widehat{\beta}_0)}} = \frac{0.70444}{0.56621} = 1.244$$

qui est la réponse A.



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%

Cette fois, on va utiliser plus en détails ce qui était expliqué dans la réponse précédante. À savoir que - sous H_0 i.e. $\beta_0 = 0$ - Z suit une loi normale centrée réduite. On veut donc calculer, comme l'indique la notation

$$\mathbb{P}(|Z| > 1.244) \text{ où } Z \sim \mathcal{N}(0,1).$$

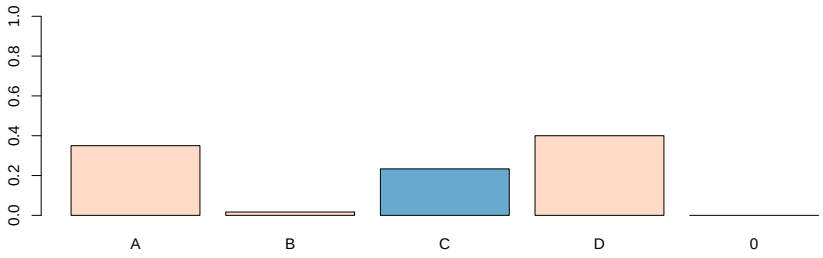
Quelques souvenirs de son cours de probabilité s'imposent ici:)

$$\mathbb{P}(|Z| > z) = \mathbb{P}(\{Z > z\} \text{ ou } \{-Z > z\}) = \mathbb{P}(\{Z > z\} \text{ ou } \{Z < -z\}) = \mathbb{P}(Z > z) + \mathbb{P}(Z < -z)$$

par symmétrie de la densité de la loi normale centrée (autour de 0, on en déduit

$$\mathbb{P}(|Z| > z) = \mathbb{P}(Z > z) + \mathbb{P}(Z > -(-z)) = 2 \times \mathbb{P}(Z > z)$$

On peut alors utiliser le tableau de la première page, i.e. $\mathbb{P}(Z > 1.28) = 10\%$, donc $\mathbb{P}(|Z| > 1.28) = 20\%$. J'en conviens, on n'avait pas 1.28 mais 1.24. Toutefois, on se doute que la réponse ne doit pas être éloignée de 20%. Comme on est plus petit que 1.28, on peut même affirmer que la probabilité doit être légèrement plus grande que 20%. Bref, la bonne réponse était 21.50%, qui est la réponse C. Beaucoup ont répond D: il ne faut pas oublier que l'on fait un test *bilattéral*: l'alternative est que le coefficient soit non-null. On cherche donc une probabilité que $|Z|$ dépasse un seuil, et pas Z. Ce qui explique que la bonne réponse soit le double.



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

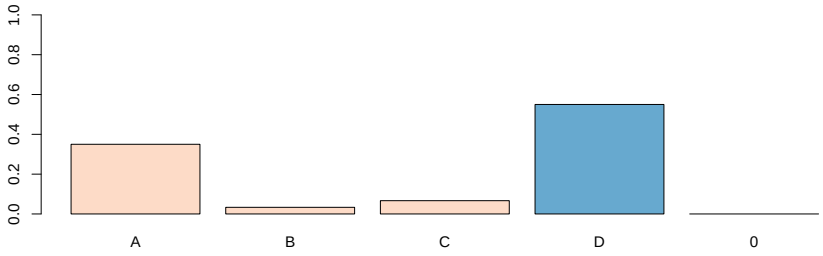
A. `***`

B. `*`

C. `.`

D. rien

Comme indiqué dans la sortie, on a des *Signif. codes*: Si la p-value (i.e. la valeur qu'on vient de calculer) dépasse 10%, on ne met *rien*. C'était la réponse D.



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

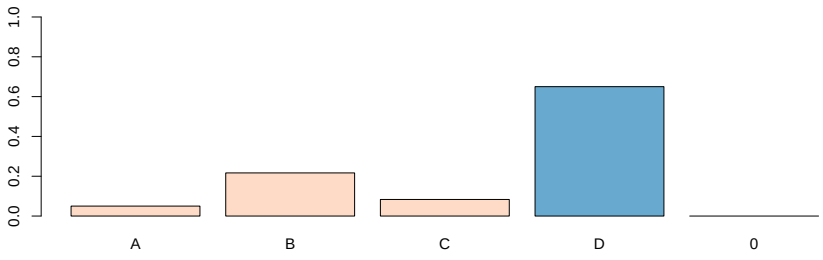
Bon, cette fois, on utilise les deux sorties. Car les deux sorties sont *équivalentes*: c'est le même modèlesqui est estimé, c'est juste qu'il n'est pas présenté pareil. Travaillons avec X_2 nul (les deux modèles doivent donner les mêmes prévisions pour tout X_2), et travaillons juste sur $\widehat{\eta}$ i.e. la combinaison linéaire des variables explicatives. Comme la fonction lien est bijective, les probabilités sont égales si et seulement si les combinaisons linéaires sont égales. On note β le coefficient que l'on cherche

- modèle regH: pour H: $\widehat{\eta} = 0.7044 - 3.0698$ et pour F: $\widehat{\eta} = 0.7044$
- modèle regF: pour H: $\widehat{\eta} = \beta$ et pour F: $\widehat{\eta} = \beta + 3.0698$

On a deux équations, une seule inconnue... mais les deux équations sont identiques:

$$\begin{cases} 0.7044 - 3.0698 = \beta \\ 0.7044 = \beta + 3.0698 \end{cases}$$

Bref, β vaut ici $0.7044 - 3.0698 = -2.3654$ qui était la réponse D. On notera que l'on pouvait rajouter x_2 dans le système, cela ne changerait rien...



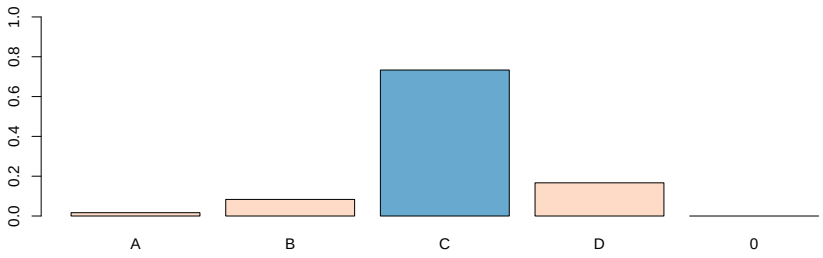
Question 7. Dans la sortie `regF`, qu'a été effacé dans la 2ème colonne (`Std. Error`) ?

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

Utilisons ce que l'on a vu auparavant. Mais dans l'autre sens cette fois, car on nous donne z-value: on sait que

$$\frac{-2.3654}{?} = -3.567$$

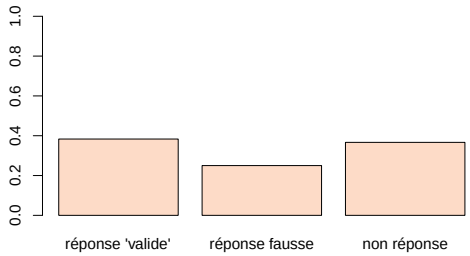
c'est à dire que la valeur que l'on cherche est 0.66313, qui correspond à la réponse C.



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

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

La place impartie était restreinte, on ne va pas raconter sa vie ici, mais cette phrase n'apparaissait pas dans le modèle linéaire *classique* car on avait une solution explicite à la maximisation de la (log)-vraisemblance, qui coïncidait avec l'estimateur obtenu par moindres carrés. Or ici, la log-vraisemblance est plus complexe, et la condition du premier ordre n'admet pas de solution explicite. Bref, on utilise un algorithme *numérique* pour résoudre la condition du premier ordre, et donner des valeurs numériques qui maximisent cette (log)-vraisemblance. La recherche des zéros d'une fonction (on veut les valeurs qui font que le gradient sera nul) se fait par l'algorithme de Newton-Raphson, correspondant ici au score de Fisher (pour rappel, la matrice Hessienne est ici la matrice d'information de Fisher, et le *score* est tout simplement le gradient de la log-vraisemblance, cf cours de stats). Bref, cette phrase nous rappelle quel la recherche du maximum de vraisemblance se fait numériquement, par un algorithme *itératif*. Et en l'occurence, on nous dit que l'algorithme numérique a convergé. Ah oui.... aucun rapport avec le test de Fisher ! Ronald Fisher (1890-1962) a (malheureusement) fait beaucoup beaucoup de choses, et on retrouve son nom un peu partout en statistique: le *score de Fisher*, *l'information de Fisher*, la *loi de Fisher*, le *test de Fisher*, le *noyau de Fisher*, l'*inégalité de Fisher*, le *théorème de Fisher* (et Tippett), sans parler de tous ses travaux en génétique et en études causales.



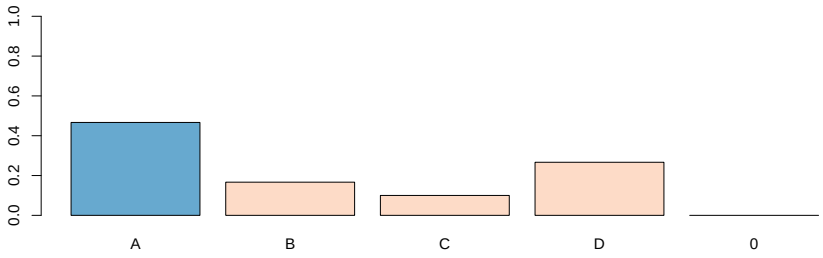
Question 9. Lors de la modélisation à l'aide d'une régression logistique, on a prédit 0.314 pour $\mathbb{E}(Y|X_1, X_2)$. Que peut-on prédire pour $\text{Var}(Y|X_1, X_2)$

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

Le modèle logit (ou probit d'ailleurs, on s'en moque ici) nous permet d'estimer un modèle de la forme $Y|\boldsymbol{X} \sim \mathcal{B}(\pi_{\boldsymbol{X}})$. Bref, le point important est que conditionnellement aux variables explicatives, on a une loi de Bernoulli. Or, pour une loi de Bernoulli, la variance est $\pi(1-\pi)$, donc ici

$$\text{Var}(Y|X_1, X_2) = \pi_{\boldsymbol{X}} \cdot (1 - \pi_{\boldsymbol{X}}) = \mathbb{E}(Y|X_1, X_2) \cdot (1 - \mathbb{E}(Y|X_1, X_2))$$

car pour rappel, $\mathbb{E}(Y|X_1, X_2) = \pi_{\boldsymbol{X}}$. Donc ici, $\text{Var}(Y|X_1, X_2) = 0.314 \cdot (1 - 0.314) = 0.215$. C'est la réponse A. Notons que la loi de Bernoulli (comme plus généralement la loi binomiale) est une loi à sous-dispersion: la variance est *toujours* plus faible que l'espérance. Seule la réponse A pouvait correspondre.



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
```

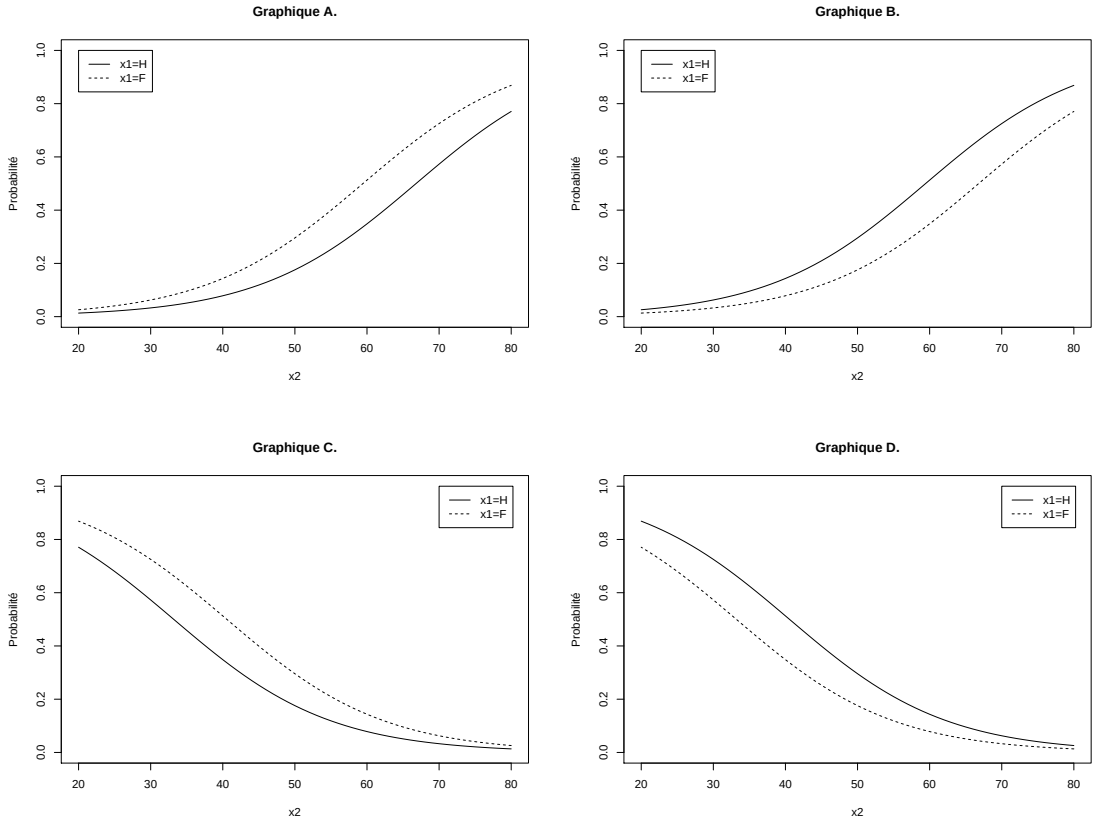
Question 10. Quel graphique ci-dessous correspond à ce qui a été demandé ?

A. le graphique A.

B. le graphique B.

C. le graphique C.

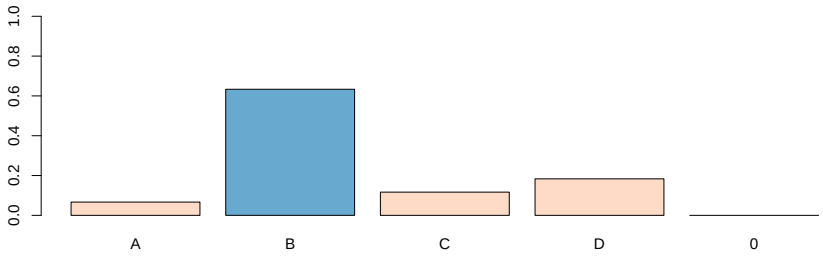
D. le graphique D.



Pour rappel, la fonction de lien est une fonction *croissante*, au sens où

$$x \mapsto \frac{\exp[\beta x + \star]}{1 + \exp[\beta x + \star]}$$

est une fonction *croissante* en x si $\beta > 0$ et *décroissante* en x si $\beta < 0$. Bref, comme β_2 est (strictement) positif, la prédiction doit être croissante en x_2. Pour les deux valeurs de x_1. On a donc le choix entre A et B. Pour trancher, il faut juste voir - en fonction des deux valeurs possibles de x_1 - laquelle sera *toujours* en dessous de l'autre. Comme $\widehat{\beta}_F$ est (strictement) inférieure à $\widehat{\beta}_H$, la prédiction sera aussi *toujours* inférieure, par croissance de la fonction de lien. Donc pour x_2 prenant la valeur F la prédiction (et donc la courbe en pointillé) sera *toujours* en dessous de l'autre. C'est ce que l'on a pour le graphique B.



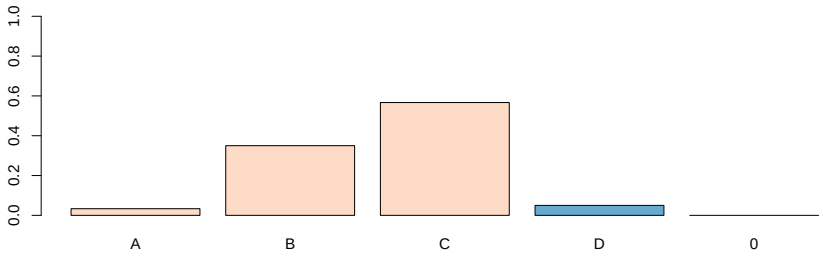
Question 11. Différentes écritures pour la régression de Poisson sont proposées

(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

Aucune de ces écriture n'a été évoqué en cours, aucune n'est va une valide. La raison fondamentale - je vais tenter de résumer en deux phrases ce qui a été pendant plusieurs en cours - avec les GLM, on ne modélise plus la variable d'intérêt, mais son espérance. On va supposer que c'est $\mathbb{E}(Y|X_1)$ qui va être une fonction d'une combinaison linéaire des variables explicatives. En l'occurence, $\mathbb{E}(Y|X_1) = \exp[\beta_0+\beta_1 X_1]$. Autant, dans un modèle Gaussien, si $Y \sim \mathcal{N}(\mu, \sigma^2)$, alors $Y = \mu+\varepsilon$ avec $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, autant pour les autres lois de la famille exponentielle on n'a pas d'écriture de ce genre. On ne peut pas dire que si $N \sim \mathcal{P}(\lambda)$ alors $N = \lambda+\varepsilon$, avec une forme simple et standard pour ε, et surtout identiquement distribué, car la régression de Poisson est fondamentalement hétéroscédastique: la régression de Poisson était hétéroscédastrique, aucune réponse nétait possible (si on suppose les bruits i.i.d.). Bref, aucune réponse ne convenait. Il fallait répondre D. J'attends d'ailleurs que l'on m'explique ce qu'est cette loi 'log-Poisson' (?).



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)     [redacted]    [redacted]  [redacted]  3.45e-11  [redacted]
X1 == "H"TRUE   [redacted]    [redacted]  [redacted]    0.0449  [redacted]
---
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

Une première question facile, qui est la version empirique de la relation de projection de l'espérance conditionnellement

$$\mathbb{E}(N) = \mathbb{E}(\mathbb{E}(N|X_1))$$

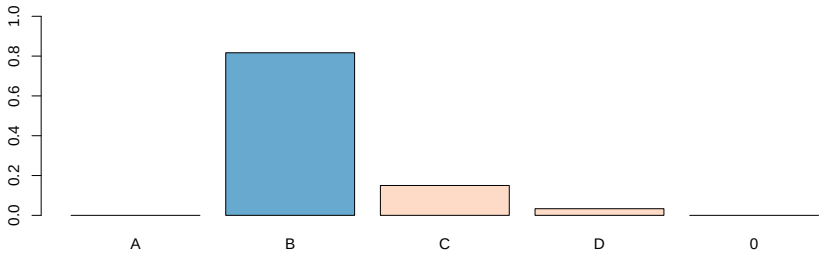
i.e.

$$\mathbb{E}(N) = \mathbb{E}(N|X_1 = H) \cdot \mathbb{P}(X_1 = H) + \mathbb{E}(N|X_1 = F) \cdot \mathbb{P}(X_1 = F)$$

On remplace ici les espérance par les moyennes empiriques, et les probabilités par les fréquences empiriques,

$$\overline{N} = 0.23611 \cdot \frac{144}{144+96} + 0.1145 \cdot \frac{96}{144+96} = 0.1875$$

qui était la réponse B. C'était la question du niveau du cours de Probas 1, histoire d'offrir des points.



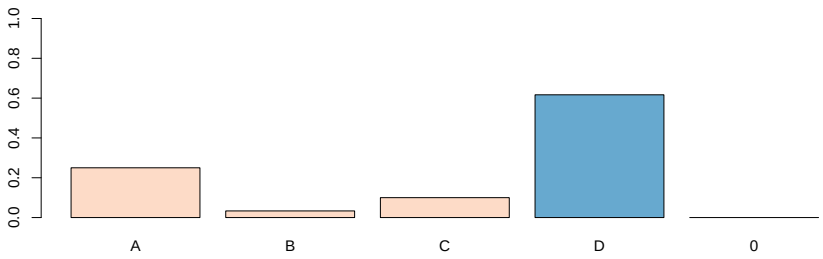
Question 13. Dans la sortie de régression `reg`, quelle est la valeur de `estimate` associée au coefficient `(Intercept)`

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

Comme on régresse sur un facteur, la constante est ici la valeur obtenue pour la modalité de référence, i.e. F. Or on a vu, qu'avec une r'egression de Poisson, la somme prédiction coïndicait avec la somme des observations. Autrement dit, la prédiction pour une personne F *doit* coïncideer avec la moyenne pour les personnes de type F. Cela se traduit par

$$\widehat{\lambda}_F = 0.1145833 = \exp[\widehat{\beta}_0]$$

autrement dit $\widehat{\beta}_0 = \log(0.1145833) \sim -2.166$, qui était la réponse D.



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

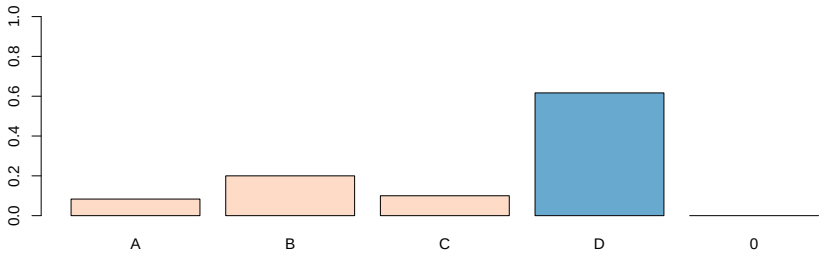
On utilise un raisonnement similaire, on notant ici que la prédiction s'écrit de manière légèrement différente,

$$\widehat{\lambda}_H = 0.23611 = \exp[\widehat{\beta}_0 + \widehat{\beta}_1] = \underbrace{\exp[\widehat{\beta}_0]}_{\widehat{\lambda}_F} \cdot \exp[\widehat{\beta}_1]$$

ce qui permet décrire

$$\exp[\widehat{\beta}_1] = \frac{\widehat{\lambda}_H}{\widehat{\lambda}_F} = \frac{0.2361111}{0.114583} \sim 2.060612$$

ou encore $\widehat{\beta}_1 = \log[2.060612] \sim 0.723$ qui était la réponse D.



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`

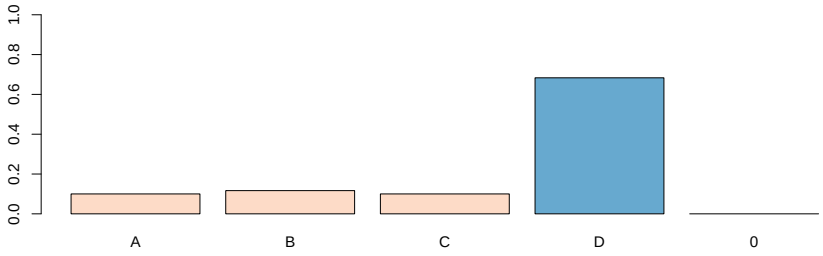
On se lance dans les calculs. La significativité se teste à partir du ratio de la valeur estimée pour un coefficient et de son écart-type (estimé). Pour la constante, on a

$$|z| = \frac{2.166}{0.3015} = 7.184$$

qui est (bien) supérieur à 1.96 (valeur critique pour un test à 95%). Donc la constante est significativement non nulle. Pour le second

$$|z| = \frac{0.723}{0.3469} = 2.0841$$

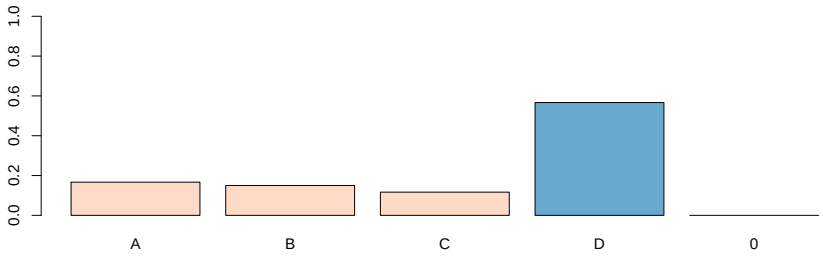
qui est (tout juste) supérieur à 1.96 (valeur critique pour un test à 95%). On accepte donc l'hypothèse de significativité du coefficient. Bref, les deux coefficients sont significativement non-nuls, on retient donc la réponse D.



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

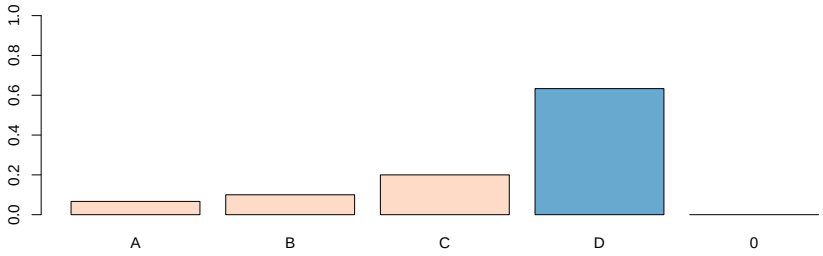
On attaque ici la régression *quasi*-Poisson. Comme on l'a vu en cours, la prédiction est ici la même (exactement la même) qu'avec une régression de Poisson, le seul changement se faisant sur la variance de Y, et donc sur la variance des estimateurs, et donc sur leur significativité (cf le cours). On a donc ici la même valeur qu'avec la régression de Poisson, i.e. -2.166 (cf question 13). On retient donc la réponse D.



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

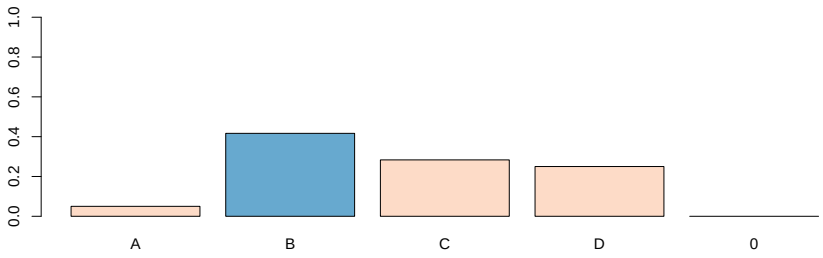
Je ne vais pas répéter ce que je disais dans la question 16, donc on conviendra que la réponse D est la bonne.



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

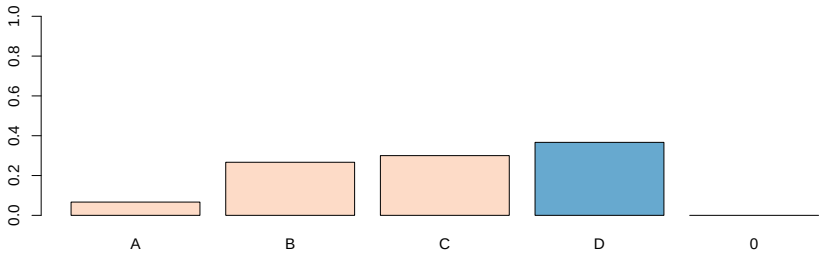
Cette fois, les choses vont changer un peu: rappelons qu'avec une loi de Poisson, on avait $\text{Var}(Y|X) = \mathbb{E}(Y|X)$. Avec une loi *quasi*-Poisson, $\text{Var}(Y|X) = \varphi \cdot \mathbb{E}(Y|X)$ On retrouve une propriété similaire sur les estimateurs $\widehat{\beta}$ et leur variance. Je n'étais pas rentré dans les détails, mais on avait vu en classe qu'avec un modèle *quasi*-Poisson surdispersé, on a avait plus de variance sur notre estimateur. Entre les variances des estimateurs de la régression dans le modèle de Poisson, et quasi-Poisson, il y a un rapport ϕ, et donc entre les écart-type, on peut s'attendre à avoir un rapport de l'ordre de $\sqrt{\phi}$. Or ici $\sqrt{\phi}$ est de l'ordre de 1.03. Il fallait donc rajouter environ 3% à l'estimation de l'écart-type.



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

L'explication est la même que pour la question précédante. La bonne réponse est la réponse D.

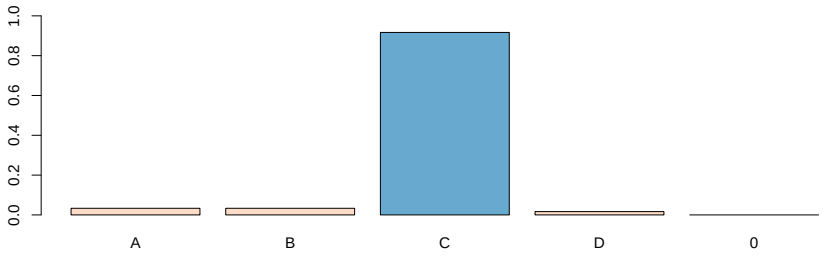


2. Analyse des données

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

On sait - ou on devrait savoir, ce n'est jamais que le 8ème cours qui utilise la loi de Poisson, que si $Y \sim \mathcal{P}(\lambda)$, alors $\mathbb{E}(Y) = \lambda$, donc l'estimateur par la méthode des moments de λ est la moyenne empirique. Mais c'est aussi l'estimateur par la méthode du maximum de vraisemblance, on l'a revu en cours (d'où l'intérêt de venir en cours ou de lire les transparents). Bref, $\widehat{\lambda} = \overline{Y}$, la valeur numérique était donnée dans la seconde ligne, 0.63055. La bonne réponse était la réponse C.



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%

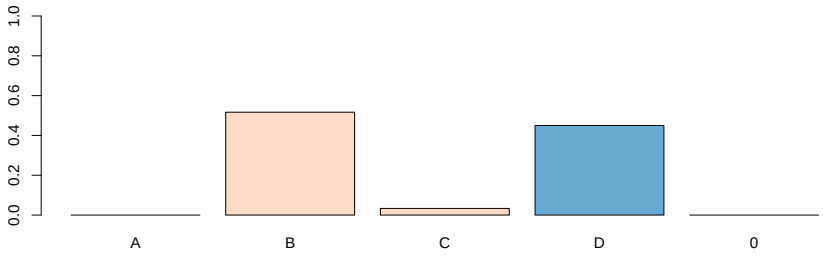
La phrase "*à l'aide du modèle prédédant*" signifie qu'on nous demande de calculer, pour $Y \sim \mathcal{P}(\lambda)$, $\mathbb{P}(Y > 0)$. Mais ça on connait - depuis le premier cours de proba - c'est

$$\mathbb{P}(Y > 0) = 1 - \mathbb{P}(Y = 0) = 1 - \frac{\exp[-\lambda]\lambda^0}{0!} = 1 - \exp[-\lambda]$$

Bon, maintenant, si on veut un estimateur de cette quantité, on remplace dans le calcul λ par son estimateur. Donc la probabilité qu'on nous demande est

$$1 - \exp\left[-\widehat{\lambda}\right] = 1 - \exp[-\overline{Y}] = 1 - \exp[-0.631] = 0.467$$

qui est la réponse D.



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 ?

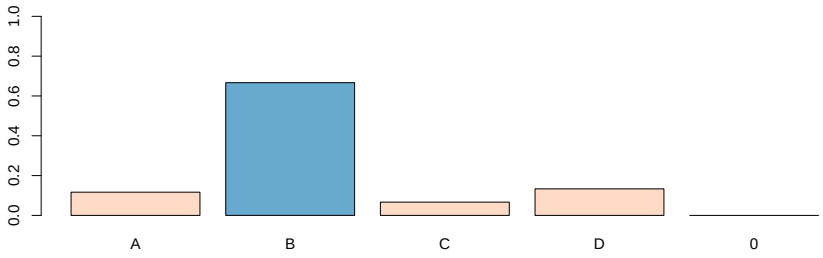
A. 80.1%
B. 19.9%
C. 24.9%

D. 46.7%

Cette fois, on utilise une loi de Bernoulli, $Y_0 \sim \mathcal{B}(\pi)$. Là aussi, on a vu en cours que pour une loi de Bernoulli, l'estimateur du maximum de vraisemblance pour π est $\overline{Y}_0$, i.e. la fréquence empirique des non-"0" dans la base. Or on nous donne la répartition. En l'occuremce, sur 563 observations, on a eu 451 fois "0". Donc

$$\overline{Y}_0 = \frac{563 - 451}{563} = \frac{}{563} = 0.199$$

aussi, la bonne réponse était la réponse B (qui, soit dit en passant, est très différente de celle obtenue avec le modèle de Poisson).



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%

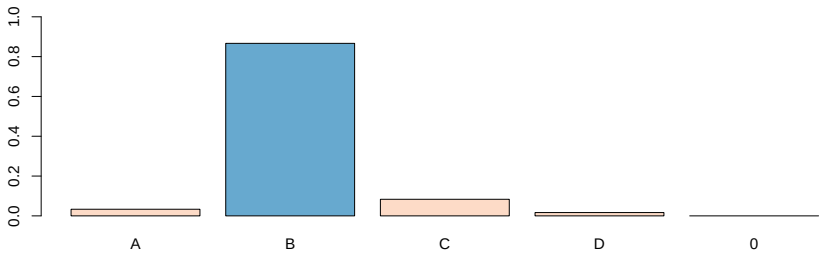
On retrouve la régression logistique (des premières questions de l'examen, on va donc pouvoir aller vite). Un homme marié depuis 10 ans, c'est `YEARMARRIAGE` prenant la valeur 10, et pour les facteurs, `SEXEF` vaut 0 alors que `SEXEH` vaut 1. Bon, maintenant on remplace...

$$\widehat{\pi} = \frac{\exp[\widehat{\beta} \times 10 + \widehat{\beta}_F \times 0 + \widehat{\beta}_H \times 1]}{1 + \exp[\exp[\widehat{\beta} \times 10 + \widehat{\beta}_F \times 0 + \widehat{\beta}_H \times 1]]}$$

et si on met les valeurs numériques, on obtient

$$\widehat{\pi} = \frac{\exp[0.03738 \times 10 - 1.55898]}{1 + \exp[0.03738 \times 10 - 1.55898]} = \frac{\exp[-1.185]}{1 + \exp[-1.185]} \sim 0.234,$$

qui correspond à la réponse B.



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%

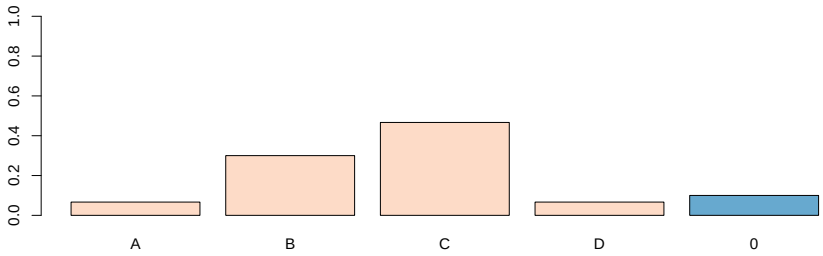
C'est la même question que celle d'avant, mais avec un autre modèle, tout comme la question 22 était distincte de la question 21. On va donc faire pareil: estimer λ et en déduire la probabilité de valoir 0. C'est parti (et on va vite dans les calculs, c'est toujours pareil..)

$$\widehat{\lambda} = \exp[0.044486 \times 10 - 0.746552] = \exp[-0.30169] = 0.7396$$

Or on avait vu dans la question 21 que dans ce cas, un estimateur de $\mathbb{P}(Y > 0|\boldsymbol{X})$ était obtenu en considérant

$$1 - \exp[\widehat{\lambda}] \text{ soit ici } 1 - \exp[-\exp[0.044486 \times 10 - 0.746552]] \sim 0.52267$$

houps. La réponse n'était pas proposée... Désolé (et bravo aux étudiants qui ont sur sa copie qu'aucune réponse proposée n'était valide, presque toujours en donnant comme *bonne* réponse 52.3%, un bonus leur a été accordé). La question a été supprimée. Désolé.



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

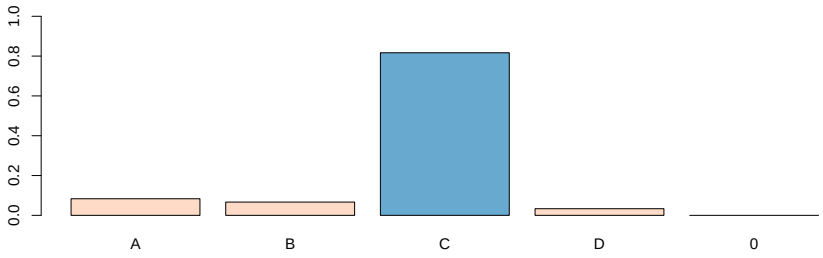
A. non, car seuls les hommes ont été pris en compte (variable `SEXEH`), il faudrait faire la même régression en prenant en compte les femmes (variable `SEXEF`)

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.

C'est la réponse C, je renvoie à tout ce que j'ai pu dire en cours, car ce point à été vu à maintes reprises... Maintenant, la réponse D pourrait - en théorie - être valide, si les personnes se trompaient au sein de la population interrogée. Dans ce cas, effectivement, le nombre d'adultères commis par les femmes devrait être égal au nombre d'adultères commis par les hommes. Mais ce n'est pas le cas... Donc cette réponse ne pouvait être acceptée (surtout que la C) correspondait à *la* réponse que j'ai toujours donné au tableau en cours..



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]

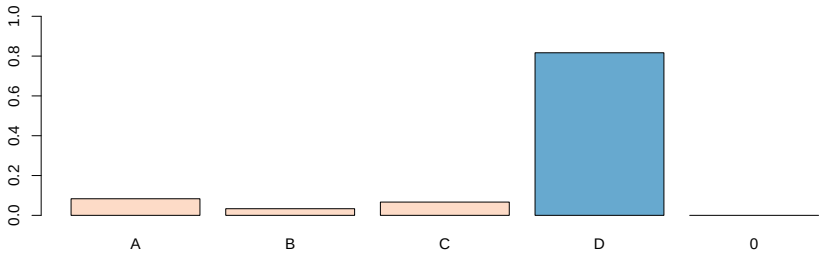
Pour les intervalles de confiance, on utilise la propriété de normalité asymptotique des estimateurs du maximum de vraisemblance. Pour faire simple (je renvoie au cours pour les subtilités), un intervalle de confiance à 95% pour β est de la forme

$$\left[\widehat{\beta} \pm 1.96\sqrt{\widehat{\text{Var}(\widehat{\beta})}}\right]$$

(c'est le cours de statistique). Bref, si on remplace par les valeurs numériques, l'intervalle de confiance s 'écrit

$$[0.12479 \pm 1.96 \times 0.02627] \text{ soit } [0.072; 0.177]$$

On retiendra donc la réponse D.



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]

On avait rappelé dans une question précédante que la fonction de lien était monotone. Aussi, si x est compris entre deux valeurs (i.e. $x \in [x_-; x_+]$), alors λ_x sera soit dans $[\lambda_{x_-}; \lambda_{x_+}]$ (si le coefficient associé est positif, soit dans $[\lambda_{x_+}; \lambda_{x_-}]$ (si le coefficient associé est négatif). Bref, toute cette histoire pour dire qu'il suffit de faire des prévisions pour les cas extrêmes (ici 1 et 5) pour avoir un intervalle pour les valeurs prédites. Bref, ici, on avait

$$\widehat{\lambda}_x = \exp[-0.0402 \times 35 + 0.09332 \times 10 - 0.25 \times 1 + 0.134 \times 20 - 0.37659 \times x]$$

où ici x est le niveau de la variable `SATISFACTION`. Notons que

$$\widehat{\lambda}_x = \underbrace{\exp[-0.0402 \times 35 + 0.09332 \times 10 - 0.25 \times 1 + 0.134 \times 20]}_{\exp[1.9562]} \exp[-0.37659 \times x]$$

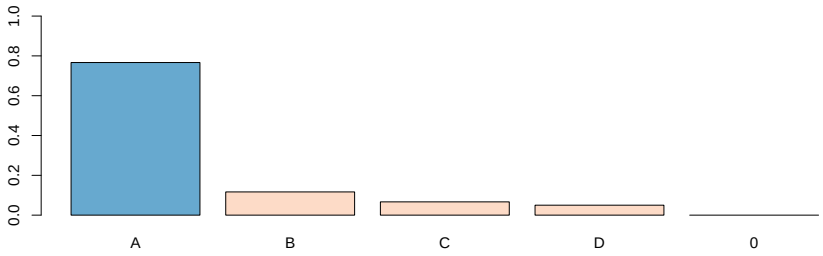
En replaçant x par 1 et 5, on obtient les valeurs

$$\widehat{\lambda}_{x=1} = 7.072 \exp[-0.37659 \times 1] \sim 4.84$$

et

$$\widehat{\lambda}_{x=5} = 7.072 \exp[-0.37659 \times 5] \sim 1.07$$

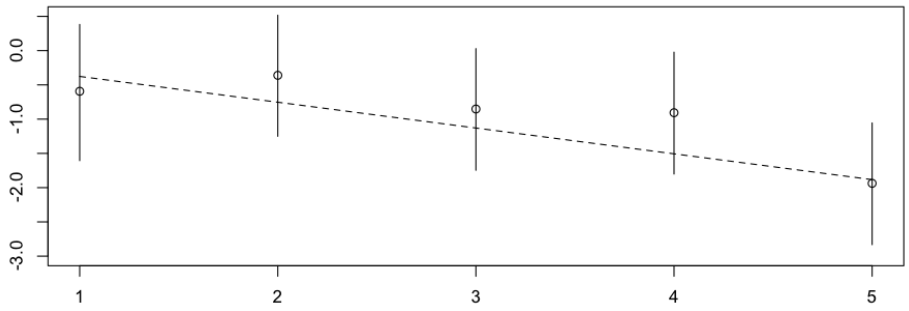
On retrouve les bornes de la réponse A. Qui semble donc être la bonne réponse.



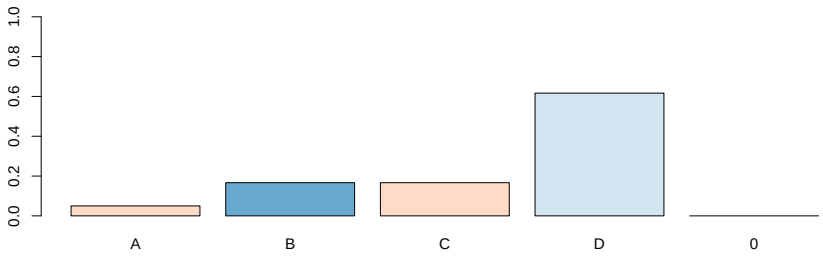
Question 28. Dans la **sortie 5**, on se demande si la variable `SATISFACTION` peut être considérée comme une variable continue, ou s'il faut la prendre en compte comme facteur. Qu'en pensez vous (une seule réponse est autorisée) ?

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)

C'est la réponse B. Pour ceux qui pensaient répondre D, il faut aussi regarder les intervalles de confiance... effectivement, on notait que les valeurs n'étaient pas *stricto sensu* monotones, mais avec l'intervalle de confiance... on ne voit pas d'effet non-linéaire ressortir clairement. La preuve en image...



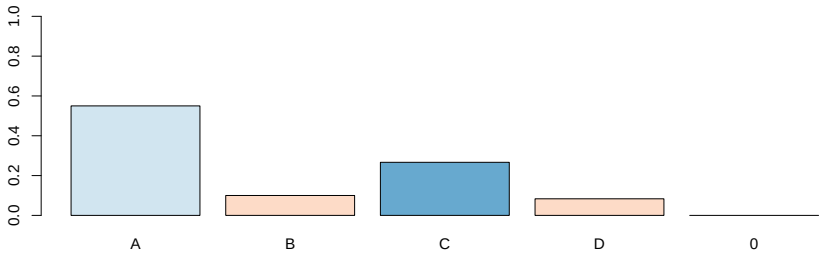
J'ai effectivement dit en cours que si un effet non-linéaire apparaît clairement, il vaut mieux utiliser des facteurs. Cela dit, on aurait tendant à les regrouper car il ne sont pas significativement différents. On le voit sur le graphique, mais un test de Fisher serait plus propre. Je donne 1/2 point pour la réponse D.



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

La réponse A était tentante, et effectivement , de la surdispersion apparaît si $\phi > 1$. Cela dit, pour tester une hypothèse de la forme $H_0 : \phi > 1$, on ne se contente pas de comparer $\widehat{\phi}$ et 1: 3.5285 est autant éloigné de 1 que peut l'être 1.02 ! Il faut prendre en compte la dispersion possible de notre estimateur. Mais on ne l'a pas. En fait, si on veut faire un test propre (cf le cours) on revient à une écriture basée sur un modèle basé sur la loi binomiale négative. Or ici, on a de l'information relative à la variabilité de l'estimateur de nuisance de la loi binomiale négative. Qui permet de tester si la loi binomiale négative est une *vraie* loi binomiale négative, ou une simple loi de Poisson. Bref, le test construit sur la loi binomiale négative nous permet de conclure qu'il *doit* y avoir de la surdispersion, et donc que ϕ est probablement strictement plus grand que 1. La bonne réponse est C. Mais je donnerai 1/2 point pour la réponse A.



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

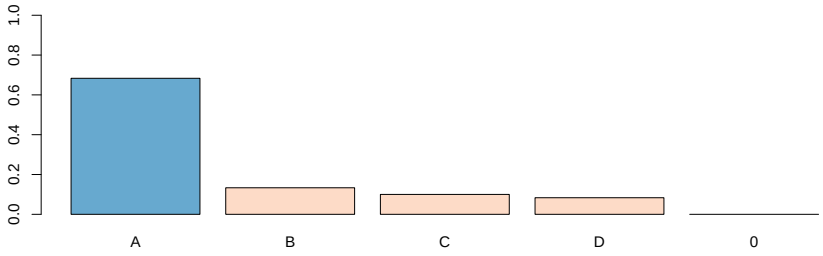
On avait déjà évoqué le modèle *quasi*-Poisson il y a peu. L'idée est que

$$\mathrm{Var}(N|\boldsymbol{X}) = \phi \cdot \mathbb{E}(N|\boldsymbol{X}),$$

soit ici

$$\mathrm{Var}(N|\boldsymbol{X}) = \phi \cdot \exp[\boldsymbol{X}'\boldsymbol{\beta}].$$

Je laisse faire les calculs.... La bonne réponse est la réponse A.



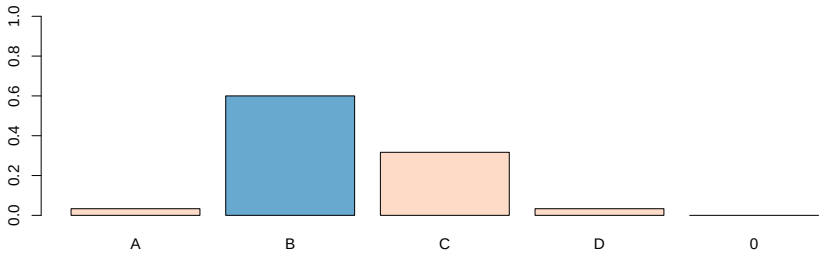
Question 31. [*suite de la question 30*]. Pour cette même personne, à l'aide du modèle `regnegbin`, donnez un estimateur pour $\mathbb{E}(N|\boldsymbol{X} = \boldsymbol{x})$,

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

On a un modèle binomial-négative (on s'en moque un peu pour cette question), avec un lien logarithmique (ça, en revanche, c'est important). Bref, ici, il suffit de prendre l'exponentielle de la combinaison linéaire obtenue avec les valeurs indiquées. Aussi,

$$\widehat{\mu} = \exp[\boldsymbol{X}'\widehat{\boldsymbol{\beta}}] \sim 1.04$$

Il fallait donc répondre B.



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.

Je renvoie au cours... En fait, je l'avais évoqué au tableau. Je demandais *juste* d'écrire la vraisemblance (je n'ai jamais demandé de résoudre les conditions du premier ordre, tout simplement parce qu'il n'y pas de solutions explicites - bravo à ceux qui ont cru les résoudre - c'est pour cela qu'on a évoqué à trois reprises dans le cours les méthodes numériques de descente de gradient),

$$\mathcal{L} = \prod_{i=1}^{n} \mathbb{P}(Y_i = y_i) = \prod_{i=1}^{n} ([\pi_i + [1-\pi_i] \cdot p_i(0)]\mathbf{1}(y_i = 0) + [1-\pi_i] \cdot p_i(k)\mathbf{1}(y_i \neq 0))$$

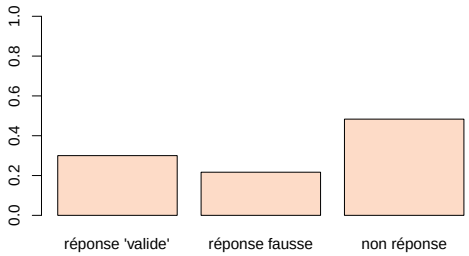
Ensuite, on utilise le fait que

$$p_i(k) = \exp[-\lambda_i]\frac{\lambda_i^k}{k!} = \exp[-\exp[\beta_0 + \beta_1 x_i]]\frac{\exp[\beta_0 + \beta_1 x_i]^k}{k!}$$

et que π_i était le même pour tous, soit

$$\mathcal{L} = \prod_{i=1}^{n} \Big((\pi + [1-\pi] \cdot \exp[-\exp[\beta_0 + \beta_1 x_i]])\mathbf{1}(y_i = 0) + ([1-\pi] \cdot \exp[-\exp[\beta_0 + \beta_1 x_i]]\frac{\exp[\beta_0 + \beta_1 x_i]^k}{k!})\mathbf{1}(y_i \neq$$

Je ne demandais pas d'en mettre plus... Bien sûr que ça peut se simplifier, mais peu importe...



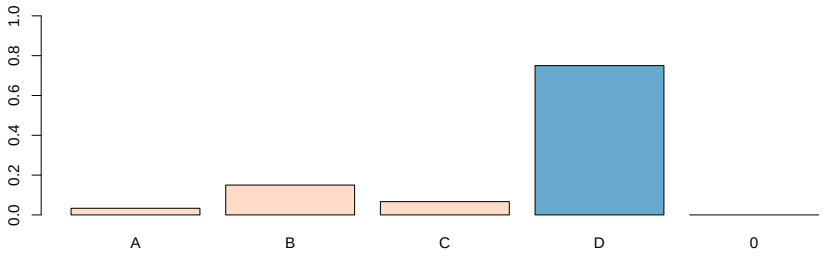
Question 33. Dans la **sortie 7**, on se demande si la surdispersion ne pourrait pas venir d'une sur-représentation des 0 dans notre base de données, qui pourrait être interprété comme un mensonge. Dans un modèle à inflation de zéro (équation (2.1)), où π_i serait considéré comme constant (noté π), quel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

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

Bon, le point essentiel est que pour la sur-représentation de zéros, on a un modèle logistique. Sans variable explicative, certes, mais le lien se fait avec la fonction de répartition de la loi logistique. Ici

$$\widehat{\pi} = \frac{\exp[\widehat{\beta}_0]}{1+\exp[\widehat{\beta}_0]} = \frac{\exp[1.2531]}{1+\exp[1.2531]} \sim \frac{3.50}{1+3.50}$$

c'est à dire environ 0.777. Il fallait répondre D. Comme $\widehat{\beta}_0$ était stricement positive, la probabilité était forcément plus grande que 1/2.



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. $\pi_1/\pi_5 \sim 120\%$

On a presque fini... Comme inscrit au tableau pendant l'intra, la **sortie 8** correspondait à la régression de la page 8/8 (il n'y en avait qu'une). Bref, on utilise le fait que l'on a une régression logistique (cf question précédante) pourraitla sur- représentation de zéros, en rajoutant cette fois la seconde variable,

$$\widehat{\pi}_x = \frac{\exp[\widehat{\beta}_0 + \widehat{\beta}_1 x]}{1+\exp[\widehat{\beta}_0 + \widehat{\beta}_1 x]}$$

On peut noter que

$$\frac{\widehat{\pi}_{x_1}}{\widehat{\pi}_{x_2}} = \frac{\exp[\widehat{\beta}_0 + \widehat{\beta}_1 x_1]}{1+\exp[\widehat{\beta}_0 + \widehat{\beta}_1 x_1]} \cdot \frac{1+\exp[\widehat{\beta}_0 + \widehat{\beta}_1 x_1]}{\exp[\widehat{\beta}_0 + \widehat{\beta}_1 x_1]}$$

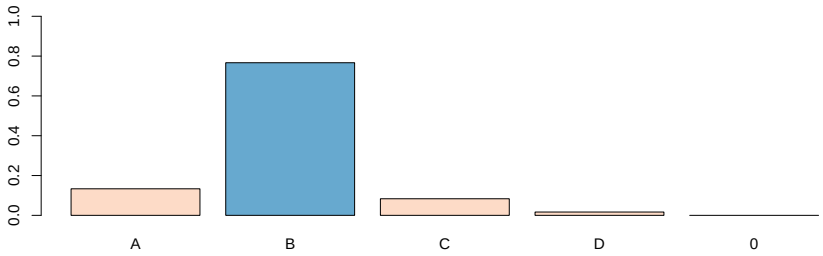
soit

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

Bon, on ne va pas y passer des heures à essayer de faire des simplifications, on peut faire un rapide calcul (surtout que β_0 ici est supposé nul),

$$\frac{\widehat{\pi}_{x=1}}{\widehat{\pi}_{x=5}} = \exp[-0.33694] \cdot \frac{1+\exp[0.3369 \times 5]}{1+\exp[0.3369 \times 1]} \sim 0.6917$$

qui est de l'ordre de 70%. La bonne réponse était la réponse B.


0.0
0.2
0.4
0.6
0.8
1.0
A
B
C
D
0