Compréhension d'une régression logistique

Jean-Paul Fischer

Compréhension d'une régression logistique

(J.-P. Fischer, Document de travail en date du 02/02/2024)

« Pour l'honneur de l'esprit humain »
(Titre du livre de Jean Dieudonné, 1987)

Sommaire

0. Introduction

Dans leur pratique, beaucoup de chercheurs contemporains – et je m'y inclus – utilisent des programmes informatiques sophistiqués qu'ils ne comprennent pas profondément. Souvent, cette utilisation s'accompagne, toujours sans grande compréhension, par des précisions techniques recopiées des manuels de référence ou fournies par le statisticien qui a analysé les données. L'objet principal de ce document est de tenter de comprendre la signification des coefficients produits par une régression logistique, une modélisation très usitée, en psychologie et, surtout, en médecine.

Cette utilisation est loin de se tarir. Dans la dernière version du manuel de référence de Stan (Stan Development Team, non daté) – un logiciel présenté comme une

plateforme de pointe pour la modélisation statistique et le calcul statistique à haute performance – j'ai relevé 105 occurrences de « logit » (voir suite). Cette omniprésence de la régression logistique n'est pas seulement due aux nombreuses variables dépendantes fondamentalement binaires (e.g., guéri / non guéri ; dyscalculique / non dyscalculique), mais sert aussi comme outil, par exemple dans la construction du score de propension défini comme la probabilité d'un participant d'être sélectionné pour recevoir un traitement spécifique conditionnellement à ses caractéristiques observées.

Dans la plupart des utilisations que je connais, les chercheurs (y compris moi-même : voir Fischer & Charron, 2009) se précipitent sur la significativité des coefficients et leur transformation en rapport de cotes (*odds ratio*) par la fonction exponentielle (e^x). Cette transformation a pu me perturber un temps car si e^x est bien approximé par 1+x lorsque x est petit, il ne l'est plus quand x est grand. Par exemple, si un coefficient β vaut 0.09, son exponentielle vaut, avec deux chiffres après la virgule, 1.09. On peut donc écrire qu'un participant du groupe A (dont la cote est au numérateur du rapport) a une cote (de guérison par exemple) supérieure de 9% à celle d'un participant du groupe B (dont la cote est au dénominateur du rapport). Mais dès β = .10 la coïncidence (ici entre .09 et 9%) disparaît puisque, en pourcentage, il faudra écrire que ce participant a une cote de guérison supérieure de 11%. En outre, on peut aussi écrire, dans le premier cas, que le participant du groupe B a une cote de guérison inférieure de 9%, mais, dans le second cas, elle est inférieure de 10% (pas 11% !).

Gelman et al. (2021, p. 221) trouvent que le concept de cote est difficile à comprendre (et on ne peut guère les soupçonner de craindre la complexité !), et que les rapports de cote sont encore plus obscurs[1]. Pour tenter de comprendre les coefficients d'une régression logistique, et avoir leur interprétation directe en termes de probabilités, je m'appuie donc quasi-exclusivement sur leur livre.

La courbe logistique s'obtient avec la fonction *plogis* de R (R Core Team, 2023) ou, je préfère pour raison pédagogique, avec la fonction *inv.logit* du package *boot*. Elle est définie mathématiquement par $\text{logit}^{-1}(x) = \text{inv.logit}(x) = \frac{e^x}{1+e^x}$. C'est l'inverse de $\text{logit}(x) = \log(\frac{x}{1-x})$. Elle est représentée graphiquement sur la figure 0. Elle s'avère intéressante pour modéliser une probabilité car cette fonction varie de 0 (pour x = -∞) à +1 (pour x = +∞), tout en restant toujours dans l'intervalle ouvert]0,1[.

[1] Howell (1998, p. 182-3) souligne les utilité et avantage du rapport de cotes qu'il (ou sa traductrice ?) appelle « chances ». Mais sa formulation de la signification du rapport de chances est simplifiée au point d'être fausse : « une personne incluse dans le groupe témoin est 1.83 fois plus susceptible d'avoir une crise cardiaque (que de ne pas en avoir) qu'une personne incluse dans le groupe expérimental » (en rouge, la partie manquante que j'ai ajoutée et qui « obscurcit » la phrase).

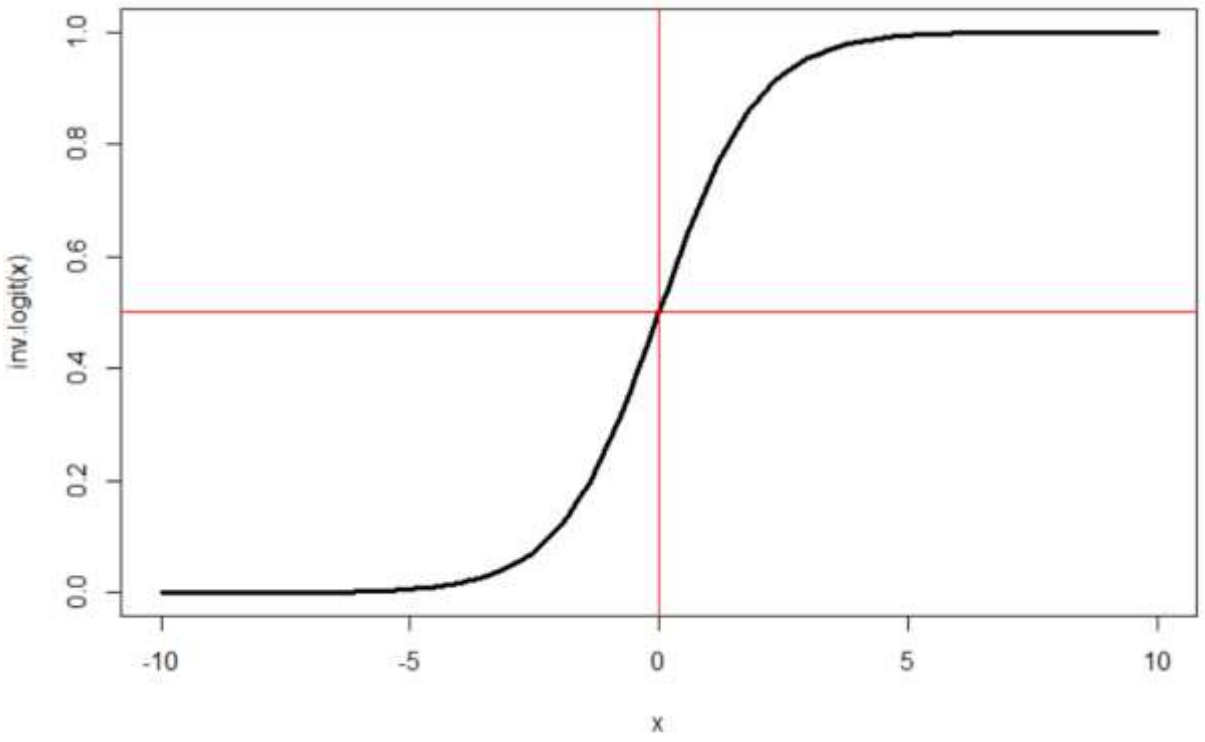

Figure 0 : La courbe logistique avec visualisation de son point d'inflexion

Sur la figure 0, le point (0, 0.5), intersection des deux droites rouges, est le point d'inflexion de la courbe. Il est déterminé mathématiquement par l'annulation de la dérivée seconde de la fonction f(x) = $\frac{e^x}{1+e^x}$. En réveillant mes souvenirs, f'(x) = $\frac{e^x}{(1+e^x)^2}$ et f''(x) = $\frac{e^x(e^x-1)}{(1+e^x)^3}$ qui ne s'annule que pour x = 0. Visuellement, la pente qui part presque de 0, augmente avant le point d'inflexion, et se met alors à diminuer pour revenir presque à 0. Au point d'inflexion la pente vaut f'(0) = $\frac{1}{4}$. Attention, à cause des axes non normés la pente de $\frac{1}{4}$ ne rappelle pas une pente de 25% que l'on pourrait rencontrer sur un sentier.

Les données que j'utilise pour ma modélisation par régression logistique (que je compare systématiquement à la régression linéaire simple) sont issues des écritures de chiffres par 691 élèves de GS maternelle, âgés entre 5 ans et 6 ans ½ (Fischer, en préparation). Chacun a écrit, dans un ordre variable, sous dictée, les dix chiffres de notre système de numération quatre fois. Lorsque, pour le chiffre 2, l'élève i a produit au moins une écriture en miroir (inversion horizontale : 2 écrit Ƨ) au cours des quatre possibilités, on lui attribue le score Y_i = 1, et Y_i = 0 sinon. Lorsque Y_i = 1 l'élève i est qualifié d'inverseur (de 2). Les deux variables-inputs Age et Sexe serviront à modéliser cette variable dépendante binaire Y (0/1). Les données proviennent d'un fichier RL (dérivé de celui de Fischer, en préparation), dans lequel la variable Item2 code les 691 Y_i résultant des inversions de 2.

1. Régressions avec intercept seulement

Bien qu'une régression logistique s'impose avec une variable dépendante binaire comme Y, R ne fait aucune difficulté pour ajuster une régression simple avec la fonction *lm* (0 et 1 sont codés en tant que nombres).

1.1. Régression linéaire simple

```
> RL0 <- lm(RL$Item2 ~ 1, data=RL);summary(RL0)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.40810    0.01871   21.81   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

L'intercept donne la probabilité (= 0.408) d'être inverseur ($Y_i = 1$) dans l'ensemble des élèves.

1.2. Régression logistique

La régression logistique relève du modèle linéaire général et se fait donc avec la fonction *glm* de R. La spécification family=binomial(link="logit") précise qu'il s'agit d'une régression logistique.

```
> RL0l <- glm(RL$Item2 ~ 1, family=binomial(link="logit"),data=RL); summary
(RL0l)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.3718     0.0774  -4.804 1.56e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

`> inv.logit(coef(RL0l[1])) = 0.4081042` redonne la probabilité d'être inverseur dans l'ensemble des élèves. Mais attention : cela peut induire l'idée fausse que les coefficients d'une régression logistique se transforment simplement en probabilités avec la fonction inv.logit.

2. Régressions avec la seule variable-input continue centrée Age

2.1. Régression linéaire simple

```
> RL1 <- lm(RL$Item2 ~ c_Age, data=RL);summary(RL1);

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.40810    0.01872  21.805   <2e-16 ***
c_Age       -0.04671    0.06017  -0.776    0.438
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

L'intercept 0.40810 est la moyenne des Y_i. La régression sur l'âge s'écrit Pr(Y=1) = 0.408 - 0.047x. Mais, par exemple pour x = 10, on a Pr(Y=1) = 0.408-0.470 = - 0.062, c'est-à-dire une probabilité négative. L'âge centré (mais pas standardisé) rend la valeur x = 10 irréaliste avec

mes données, mais dans l'absolu l'âge centré peut théoriquement varier, dans des échantillons extrêmes, entre presque -120 ans et + 120 ans.

Sur les figures 2.1, où une fonction *jitter* décale verticalement les points en maintenant leur ordonnée entre 0 et 1, je visualise la droite de régression sur l'intervalle [-2, +2] (Fig. 2.1a), puis sur l'intervalle [-25, +25] (Fig. 2.1b). Les traits verts délimitent l'intervalle des âges observés (j'ai arrondi les bornes à l'entier supérieur en valeur absolue).

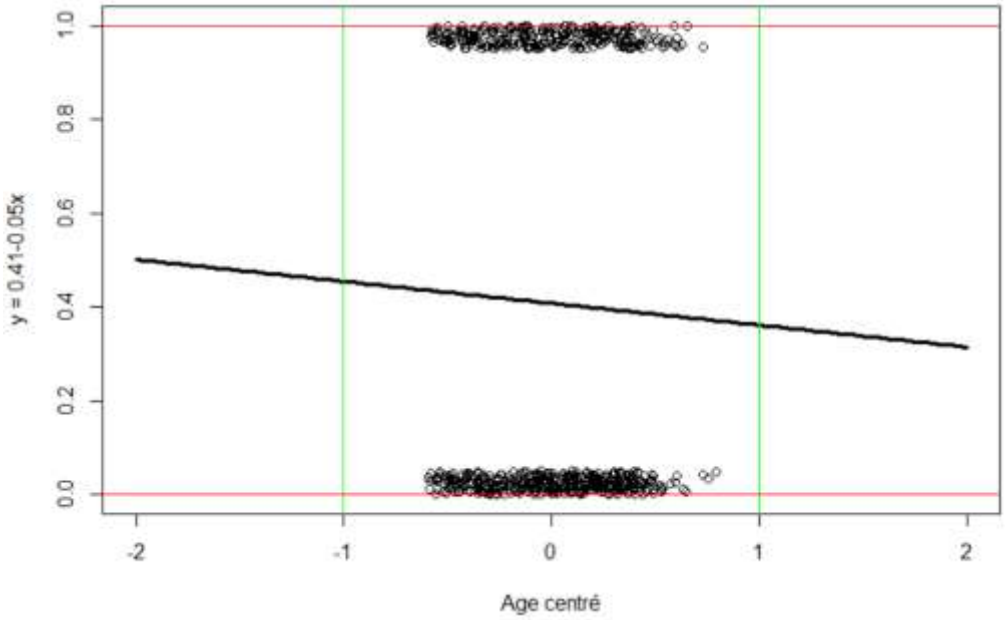

Figure 2.1a : Droite de régression du pourcentage d'inverseurs sur l'âge des élèves, sur l'intervalle [-2, +2]

En élargissant l'intervalle de variation de x à [-25, +25] on obtient :

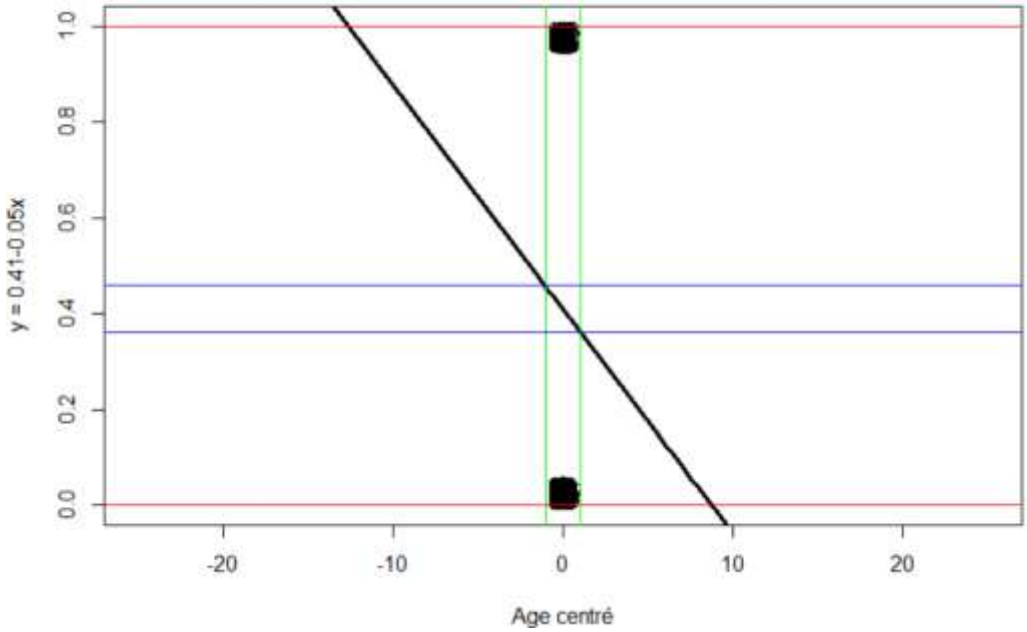

Figure 2.1b : Droite de régression du pourcentage d'inverseurs sur l'âge des élèves, prolongée sur l'intervalle [-25, +25]

Le graphique 2.1b confirme que, pour des valeurs extrêmes de l'âge centré (< -15 ou > 10), la probabilité Pr(Y=1), c'est-à-dire la probabilité d'avoir inversé 2 au moins une fois, est supérieure à 1 ou inférieure à 0. Cela est rédhibitoire pour une modélisation et peut suffire à justifier la régression logistique.

2.2. Régression logistique

```
> RL1l <- glm(RL$Item2 ~ c_Age, family=binomial(link="logit"),data=RL); sum
mary(RL1l)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.37214    0.07744  -4.805 1.54e-06 ***
c_Age       -0.19349    0.24906  -0.777    0.437
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Pr(Y=1) = logit^{-1}(-0.37214 -0.19349*c_Age)

inv.logit(-0.37214) = 0.408024 redonne bien la moyenne des 691 élèves.

Si je reprends mon exemple d'âge centré = 10, j'ai cette fois-ci

Pr(inverseur2) = logit^{-1}(-0.37214 -1.9349) = 0.091

qui est bien positive ; même avec une valeur totalement irréaliste (avec mes données) de l'âge centré = 50, j'obtiens une valeur certes très petite ($4*10^{-5}$) mais toujours positive conformément à ce qui est attendu pour une probabilité.

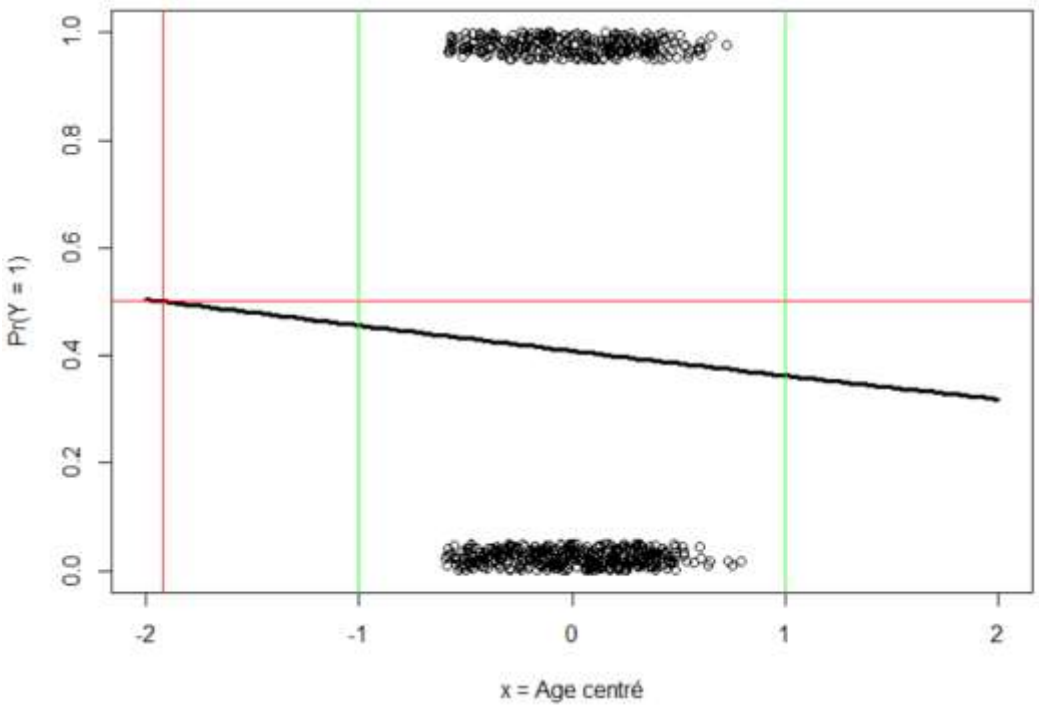

Figure 2.2a : Courbe de régression logistique Pr(Y=1) = logit^{-1}(-0.37 – 0.19*c_Age) sur l'intervalle [-2, +2]

Sur la figure 2.2a, la courbure de la représentation de la régression logistique n'apparait guère. Le calcul permet cependant d'observer que la pente p = -0.047 entre -1 et +1 est presque égale, très légèrement supérieure à celle du coefficient de régression logistique **divisé par 4** (= -0.048).

Explications. La pente entre -1 et +1 est presque identique à la pente de la courbe logistique au point d'inflexion (intersection des droites rouge) car ce dernier est proche. Elle est légèrement supérieure car – cela se verra un peu sur la figure 2.2b – la pente de la courbe commence légèrement à remonter après le point d'inflexion. Mais la courbe logistique est la plus pentue au point d'inflexion, où $y = \alpha + \beta x = 0$ de telle sorte que $\text{logit}^{-1}(\alpha + \beta x) = 0.5$. Comme cette pente maximale (en valeur absolue) est la valeur de la dérivée de la fonction $\text{logit}^{-1}(y) = e^y/(1+e^y)$ elle vaut, en ce point, $\beta e^0/(1+e^0)^2 = \beta/4$. Si on veut faire le calcul de la dérivée à la main, il est bon de se souvenir (aussi !) des dérivées $(u/v)' = (u'v - v'u)/v^2$ et $(e^u)' = u'e^u$.

Sur la figure 2.2b, on voit mieux, que sur la figure 2.2a, pourquoi la pente calculée entre -1 et +1 est légèrement supérieure au coefficient de régression divisé par 4. En effet, le point d'intersection des droites rouges, centre de la courbe logistique où est calculé la pente (que l'on divise ensuite par 4) est à gauche de l'intervalle (entre les traits verts) où j'ai calculé cette pente. Or, visuellement, on voit que la pente de la courbe a déjà très légèrement entamé sa remontée vers 0.

Attention : du fait que le pourcentage d'inverseurs décroît avec l'âge, la courbe logistique est inversée (symétrique par rapport à un axe horizontal) par rapport à sa représentation classique sur la figure 0 ; en outre, cela conduit à des comparaisons de pentes négatives qui diffèrent de leurs comparaisons en valeurs absolues (-0.47 est supérieur à -0.48).

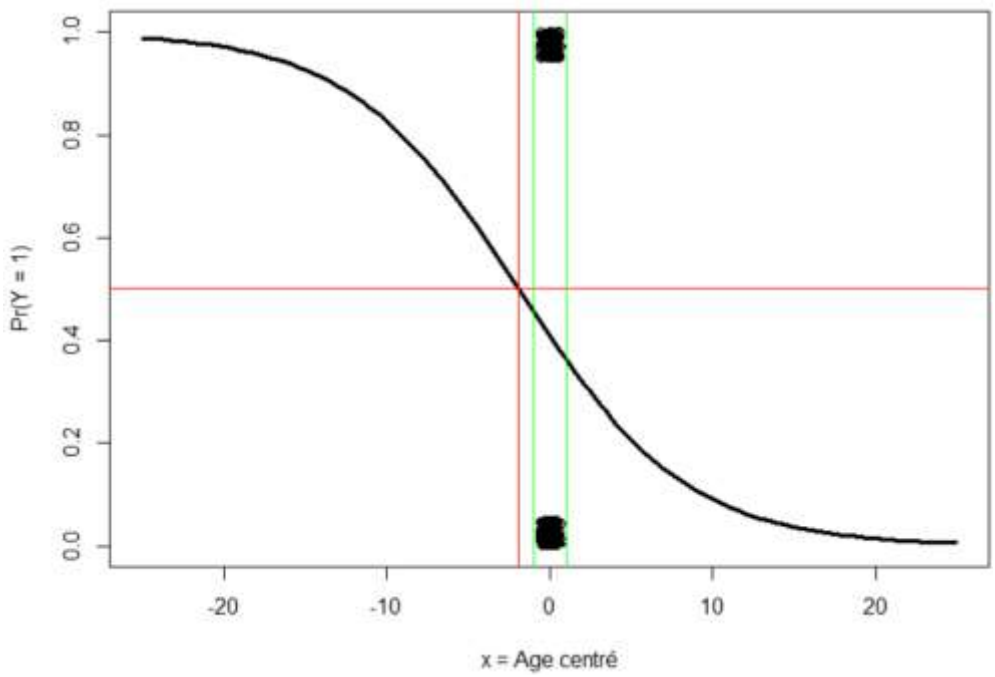

Figure 2.2b : Courbe de régression logistique $\Pr(Y=1) = \text{logit}^{-1}(-0.37 - 0.19*c\_Age)$ prolongée sur l'intervalle [-25, +25]

3. Régressions avec la variable-input binaire Sexe

3.1. Régression linéaire simple

Rappelons qu'estimer une différence est la même chose que régresser une variable indicatrice (Gelman et al., fin du chapitre 7).

```
> RL2 <- lm(RL$Item2 ~ Sex, data=RL); summary(RL2)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.44872    0.02779  16.149   <2e-16 ***
Sex1        -0.07405    0.03752  -1.974   0.0488 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

L'intercept 0.44872 est bien la moyenne parfaite des filles inverseuses. Celle des garçons inverseurs est coef(RL2)[1] + coef(RL2)[2] : 0.3746702 avec 7 décimales ! (Dans R, les coefficients sont numérotés d'après leur ordre dans le tableau des coefficients).

Le coefficient -0.074 négatif montre que le pourcentage des inverseurs décroît lorsqu'on passe des filles (code 0) aux garçons (code 1), et a permis d'obtenir le pourcentage d'inverseurs des garçons.

En accord avec le rappel initial, un t-test conduit exactement à la même p :

```
t.test(RL$Item2[RL$Sex==0],RL$Item2[RL$Sex==1], var.equal = TRUE,alternativ
e = "two.sided"): t = 1.9736, df = 689, p-value = 0.04883
```

À noter que j'ai dû forcer un test t avec variances égales car R utilise par défaut la procédure pour variances inégales de Welch qui donnait t = 1.9683, df = 656.54, p-value = 0.04945. La différence entre les deux tests suggère, au passage, que les variances ne sont pas parfaitement égales (car si les variances étaient égales, la procédure de Welch donnerait le même résultat).

On peut aussi – en particulier du fait que le score 0 ou 1 n'est pas une bonne variable quantitative – considérer que la proportion d'inverseuses dans les 312 filles est 140/312 = 0.4487, avec une SE = racine(0.4487*(1-0.4487)/312) = 0.0282, et la proportion d'inverseurs dans les 379 garçons de 142/379 = 0.3747, avec une SE = racine(0.3747*(1-0.3747)/379) = 0.0249, et alors estimer la différence à 0.4487 - 0.3747 = 0.0740, avec une SE = racine(0.0282^2+0.0249^2) = 0.0376. La différence est identique à celle obtenue avec la régression, et la SE associée, à peu près la moitié de cette différence[2], s'accorde avec p, proche de .05, obtenue aussi bien avec la régression qu'avec le test de Student.

[2] Gelman et al. comparent systématiquement l'estimation du coefficient (en valeur absolue) avec deux fois sa SE pour le qualifier de substantiel (ou analogue : ils évitent de parler de significativité).

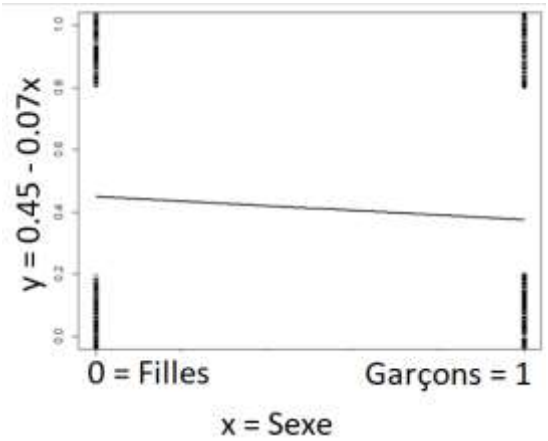

Figure 3.1 : Droite de régression du pourcentage d'inverseurs sur le sexe des élèves

Pour la figure 3.1, j'ai décalé (*jitter*) verticalement les points associés aux codes 0 et 1 (numériques) de la variable dépendante, mais pas ceux de la variable catégorielle Sexe (*jitter* ne s'applique pas aux codes non numériques). La droite de régression montre la baisse du pourcentage d'inverseurs lorsqu'on passe des filles aux garçons. Ici l'argument que l'on obtiendrait des probabilités inférieures à 0 ou supérieures à 1 ne s'applique guère car x ne varie pas en dehors, ni même dans l'intervalle [0, 1]. La régression logistique s'impose donc moins que dans le cas de la variable continue Age. Je l'ai néanmoins pratiquée.

3.2. Régression logistique

La régression logistique sur un simple prédicteur binaire équivaut à une comparaison de proportions (Gelman et al., p. 225). Cela ne doit pas suggérer de mettre en œuvre des régressions logistiques pour de simples comparaisons de proportions, mais démontre l'unité sous-jacente des comparaisons et de la régression logistique pour les données binaires (ibidem, p. 226).

```
> RL2l <- glm(RL$Item2 ~ Sex, family=binomial(link="logit"),data=RL); summa
ry(RL2l)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.2059     0.1138  -1.808   0.0705 .
Sex1         -0.3064     0.1556  -1.969   0.0490 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Pr(Y = 1) = logit^{-1}(-0.2059 - 0.3064*Sex)
L'intercept et le coefficient sur l'échelle logarithmique ne sont pas intuitifs. Pour l'intercept, on obtient à nouveau la moyenne des filles inverseuses avec inv.logit(-0.2059) = 0.4487061. Comme les quatre chiffres manquent de précision, j'en ai mis 6 : round(coef(RL2l)[1], 6) = -0.205852 et inv.logit(-0.205852) = 0.448718 ce qui est bien la valeur de l'intercept avec lm simple.

Pour le coefficient de Sexe, cela se complique : il ne suffit pas d'inverser le coef de la régression logistique pour retrouver celui de la régression simple car $\text{logit}^{-1}(a+b) \neq \text{logit}^{-1}(a) + \text{logit}^{-1}(b)$. Il faut (ou on peut ?) calculer la probabilité d'être inverseur d'un garçon en utilisant le coefficient de sexe sur l'échelle logarithmique:

> inv.logit(coef(RL2l)[1] + coef(RL2l)[2]) = 0.3746702 et donc :

inv.logit(coef(RL2l)[1]+coef(RL2l)[2]) - inv.logit(coef(RL2l)[1]) =-0.07404776 qui, avec 5 décimales, est aussi le coefficient de Sex dans la régression lm (=-0.07405).

Pour arriver simplement à une valeur approximative interprétable de ce coefficient sur l'échelle logarithmique, on peut le diviser par 4 : -0.3064/4 = -0.077. Cela est un peu moins que -0.074 mais devrait être un peu plus. La règle approximative de division par 4 ne s'applique probablement pas et la représentation graphique confirme l'inintérêt de la régression logistique pratiquement identique à la régression simple et, comme déjà dit, à une comparaison de proportions. Le point central de la courbe logistique (-0.67, 0.50) est d'ailleurs en dehors du graphique.

Figure 3.2 : Courbe de régression logistique Pr(Y=1) = logit^{-1}(-0.2059 – 0.3064*Sexe)

4. Régressions avec les variables-inputs Sexe et Age

4.1. Régression linéaire simple

```
> RL3 <- lm(RL$Item2 ~ Sex + c_Age, data=RL); summary(RL3)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.44936    0.02780  16.163   <2e-16 ***
Sex1        -0.07523    0.03755  -2.003   0.0455 *
c_Age       -0.05115    0.06008  -0.851   0.3948
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Cette fois-ci, l'intercept est la probabilité estimée d'être une fille inverseuse si l'âge moyen centré des filles est 0, ce qu'il n'est pas tout à fait: cela peut expliquer pourquoi cette estimation diffère très légèrement de la proportion réelle des filles inverseuses (=0.44872).

Hormis illustrer encore une fois la probabilité supérieure des filles, les figures 4.1 ne nous apprennent rien de nouveau.

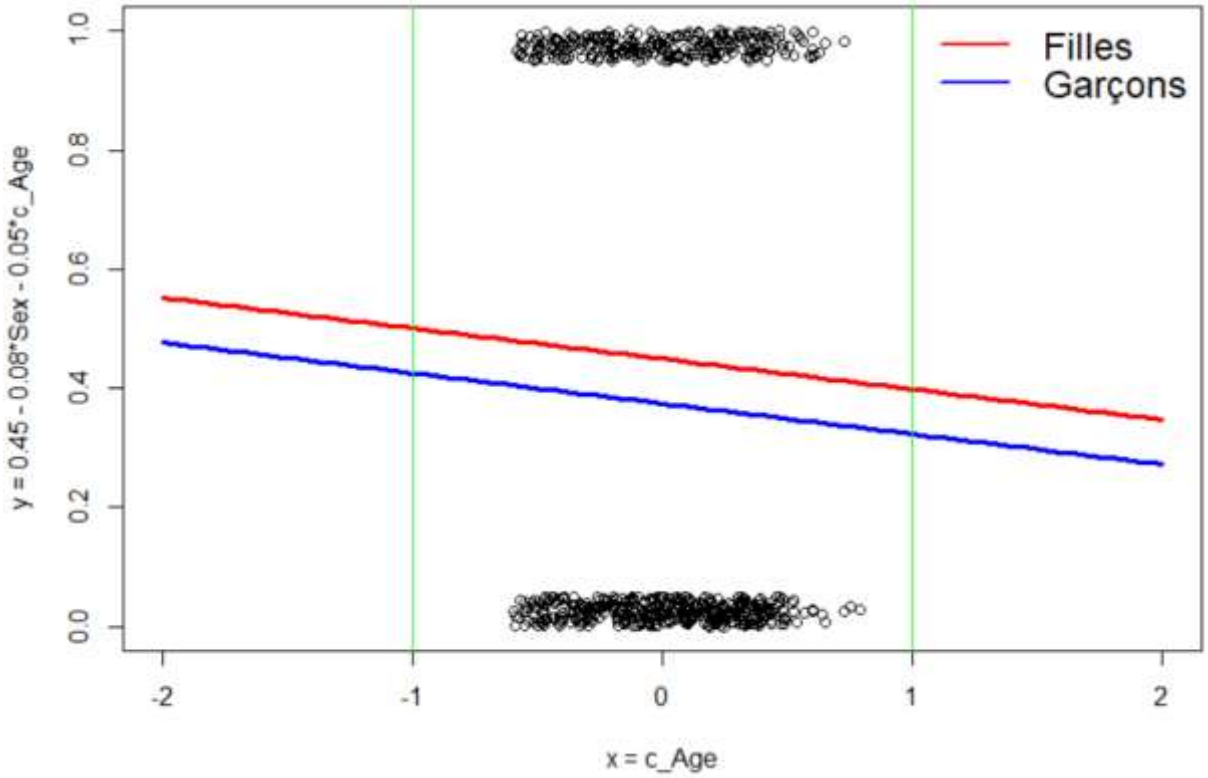

Figure 4.1a : Droite de régression y = 0.449 - 0.075*Sex – 0.051*c_Age du pourcentage d'inverseurs sur l'âge des élèves, en fonction de leur Sexe, sur l'intervalle [-2, +2]

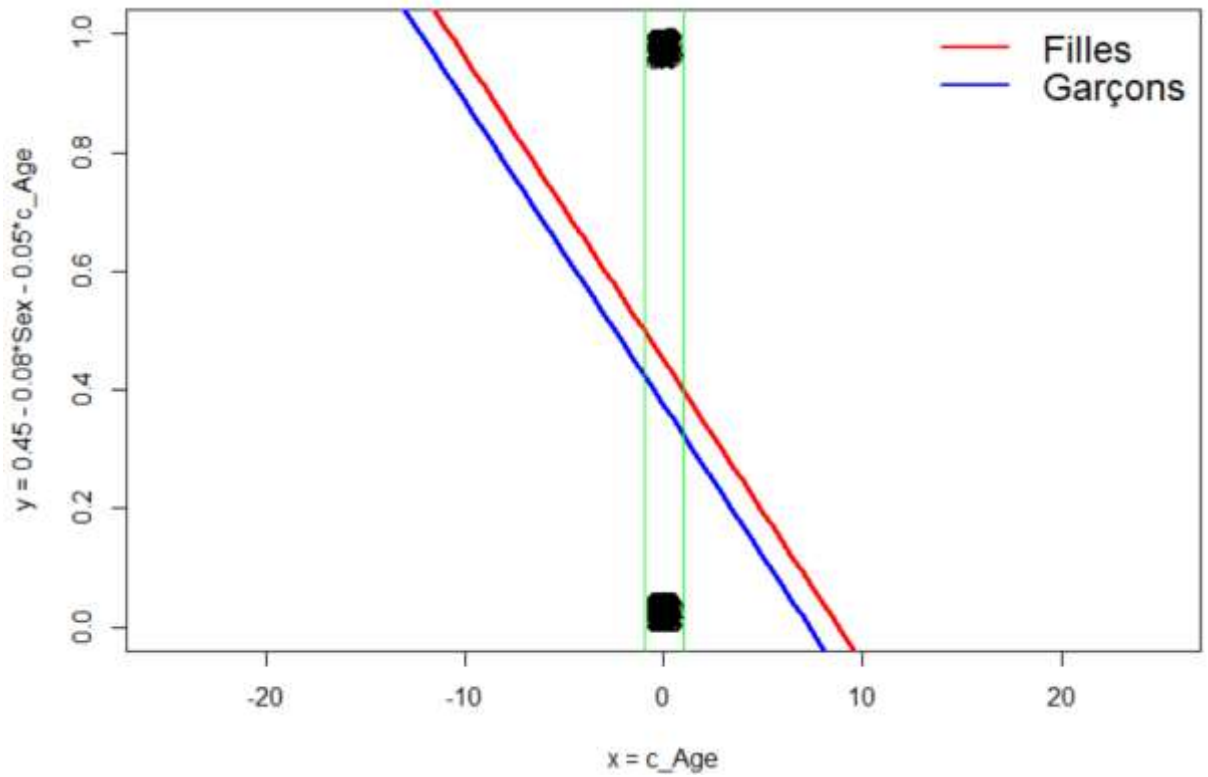

Figure 4.1b : Droite de régression du pourcentage d'inverseurs sur l'âge des élèves, en fonction de leur Sexe, prolongée sur l'intervalle [-25, +25]

4.2. Régression logistique

```
> RL3l <- glm(RL$Item2 ~ Sex + c_Age, family=binomial(link="logit"),data=RL
); summary(RL3l)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.2034     0.1139  -1.785   0.0742 .
Sex1         -0.3116     0.1558  -2.000   0.0455 *
c_Age        -0.2131     0.2500  -0.852   0.3940
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

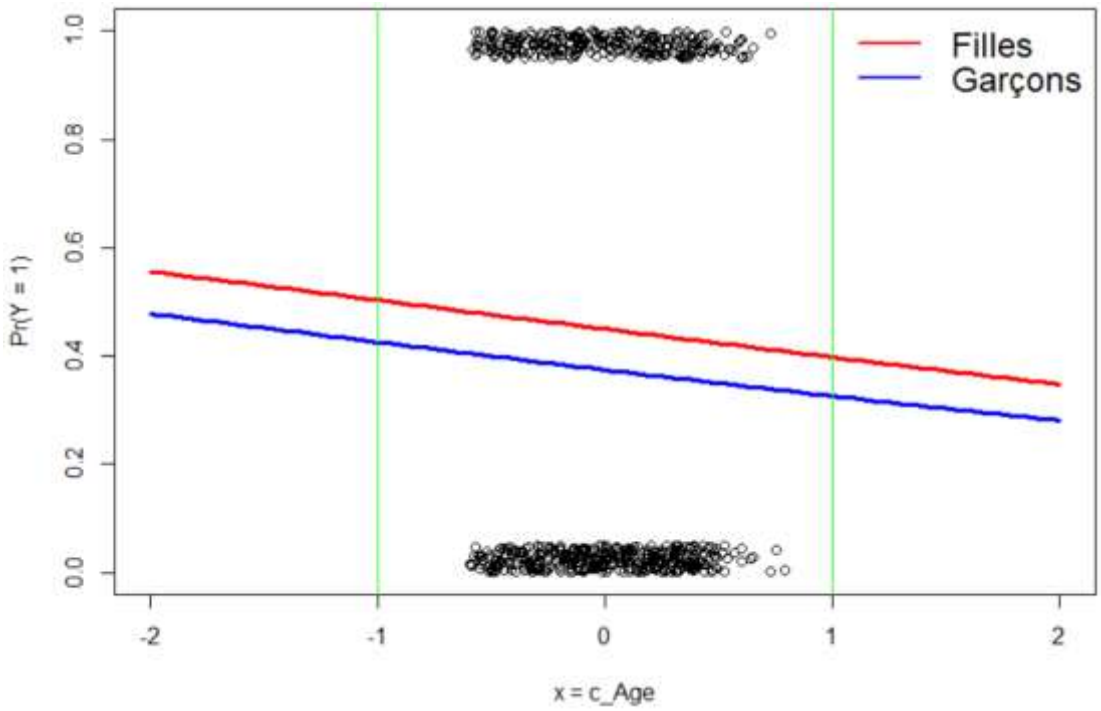

Figure 4.2a : Courbe de régression logistique de Pr(Y=1) = logit^{-1}(-0.2034 – 0.3116*Sex – 0.2131*c_Age) sur l’âge des élèves, en fonction de leur Sexe, sur l’intervalle [-2, +2]

Pour une meilleure visualisation de la courbe j'ai agrandi l'intervalle de variation de x :

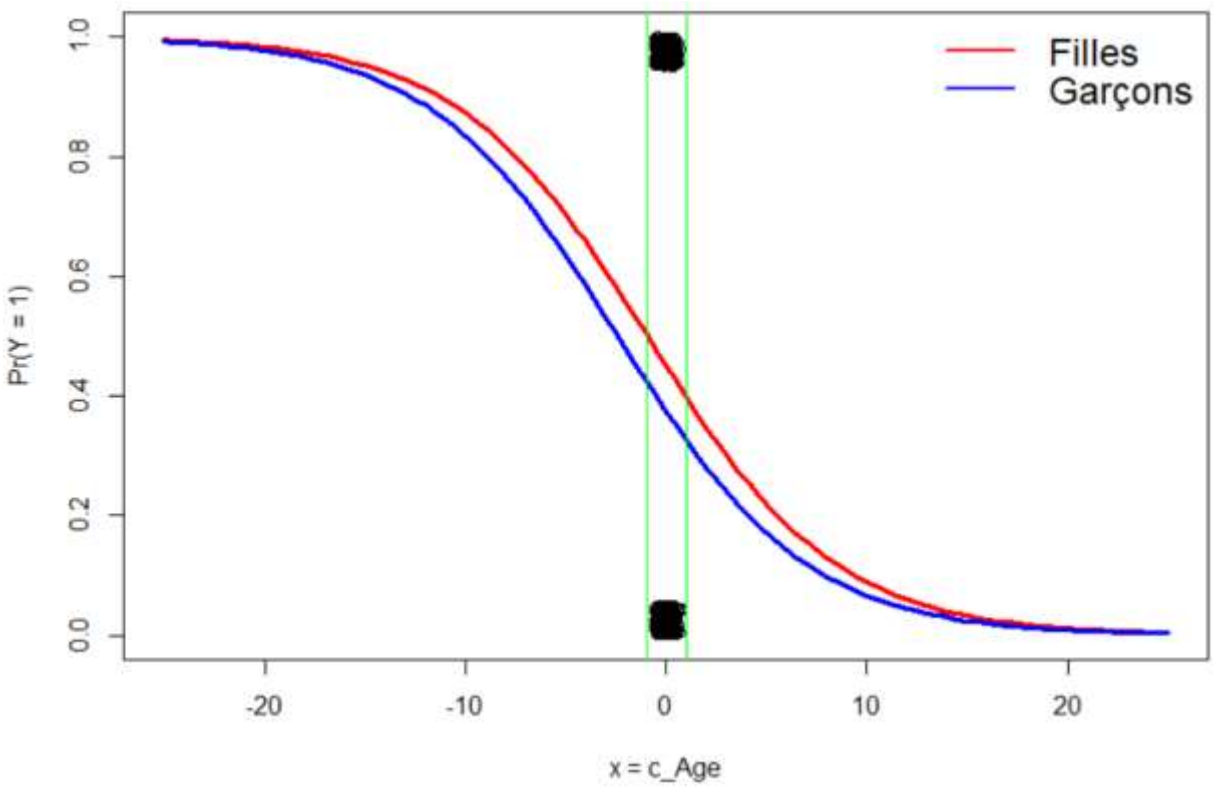

Figure 4.2b : Courbe de régression logistique de Pr(Y=1) = logit^{-1}(-0.2034 – 0.3116*Sex – 0.2131*c_Age) sur l'âge des élèves, en fonction de leur Sexe, prolongée sur l'intervalle [-25, +25]

Le calcul des pentes entre x = -1 et x = +1 confirme que la pente de la courbe des garçons est très légèrement supérieure à celle des filles : p = -0.050 vs. p = -0.053. Ces deux pentes sont proches, légèrement supérieures en moyenne au coefficient de c-Age divisé par 4 (= 0.053).

5. Régressions avec l'interaction Sexe*Age

5.1. Régression linéaire simple

```
> RL4 <- lm(RL$Item2 ~ Sex*c_Age, data=RL); summary(RL4)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.44925    0.02783  16.140   <2e-16 ***
Sex1        -0.07519    0.03758  -2.001   0.0458 *
c_Age       -0.04175    0.08887  -0.470   0.6387
Sex1:c_Age  -0.01734    0.12067  -0.144   0.8858
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Le coefficient de l'interaction, bien que très inférieur (en valeur absolue) à l'erreur-type associée, s'interprète dans deux directions: (1) En passant des filles aux garçons (i.e., de 0 à 1 : changement d'une unité), il faut ajouter -0.017 au coefficient -0.042 de l'âge ; (2) pour chaque année supplémentaire d'âge, la valeur -0.017 est ajoutée au coefficient du Sexe.

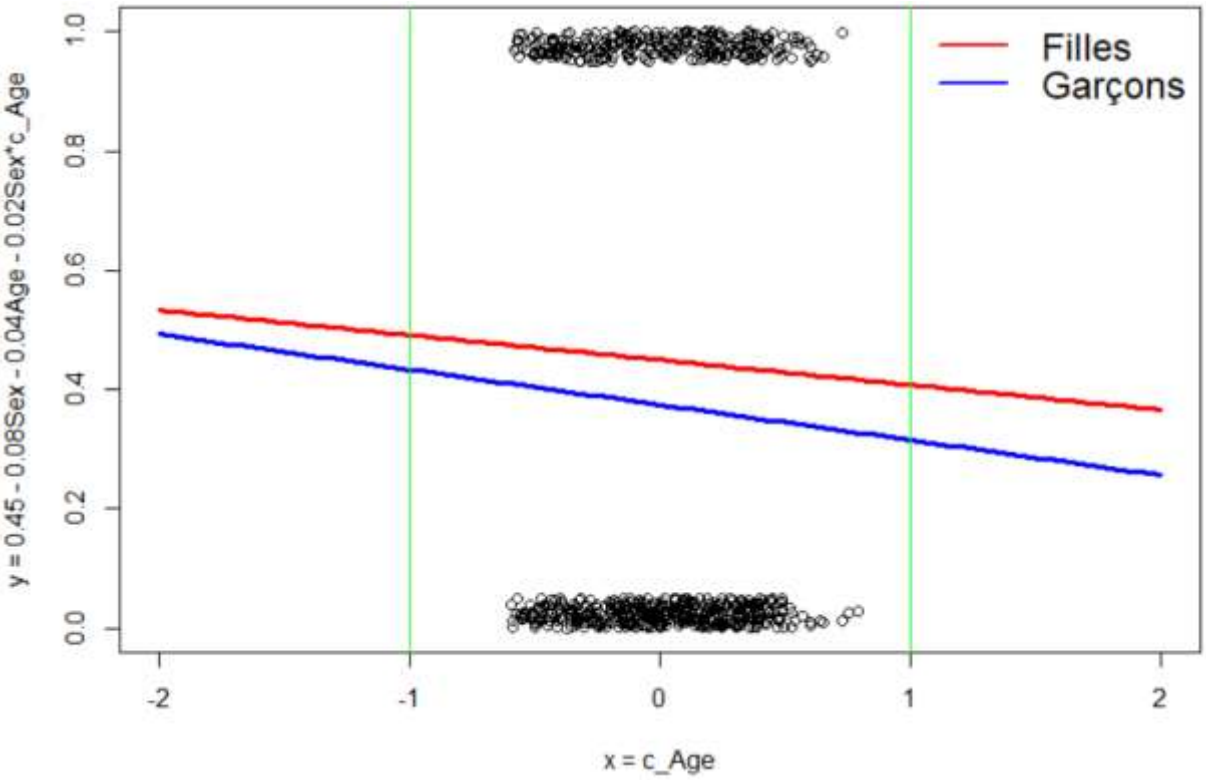

Figure 5.1a : Droite de régression du pourcentage d'inverseurs sur l'âge des élèves, en fonction de leur Sexe, sur l'intervalle [-2, +2]

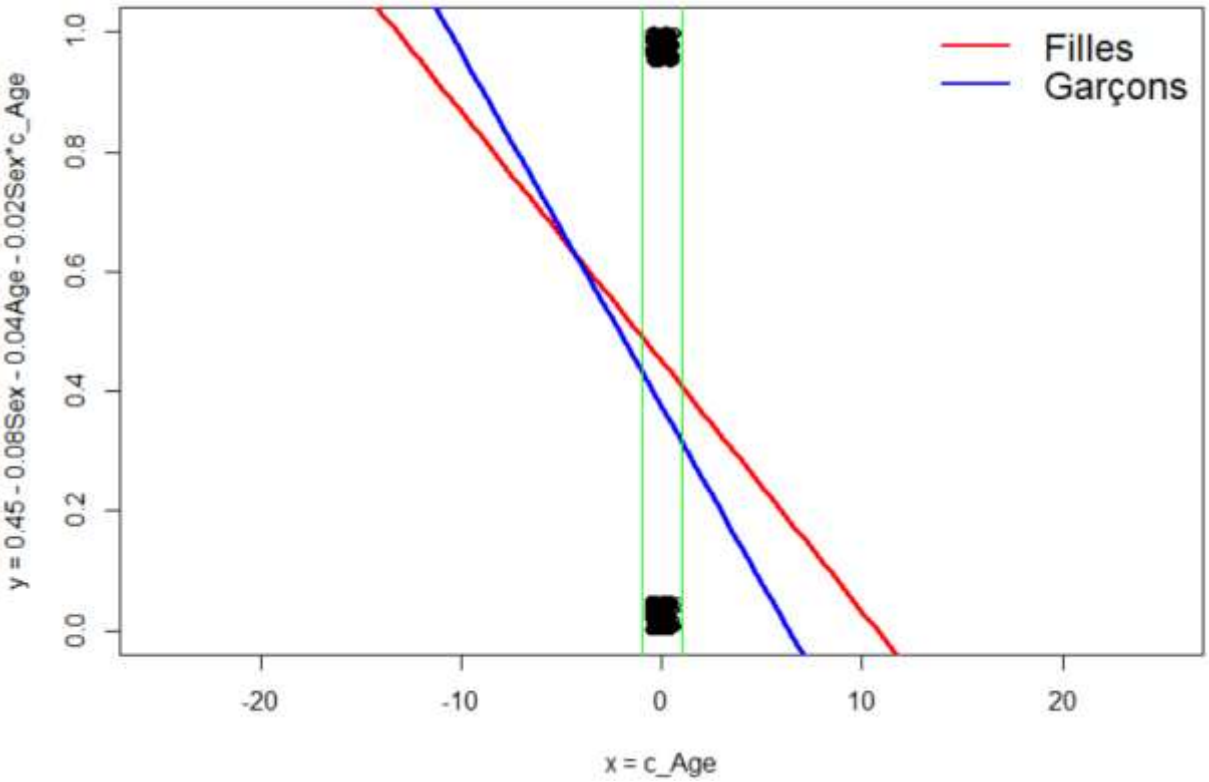

Figure 5.1b : Droite de régression du pourcentage d'inverseurs sur l'âge des élèves, en fonction de leur Sexe, prolongée sur l'intervalle sur [-25, +25]

L'interaction est graphiquement sensible car même si l'erreur-type associée est trop grande pour croire l'interaction fiable, il n'en demeure pas moins que son estimation, au vu des autres coefficients, n'est pas négligeable.

5.2. Régression logistique

```
> RL4l <- glm(RL$Item2 ~ Sex*c_Age, family=binomial(link="logit"), data=RL)
; summary(RL4l)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.20386    0.11394  -1.789   0.0736 .
Sex1        -0.31177    0.15586  -2.000   0.0455 *
c_Age       -0.16881    0.36391  -0.464   0.6427
Sex1:c_Age  -0.08376    0.50076  -0.167   0.8672
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

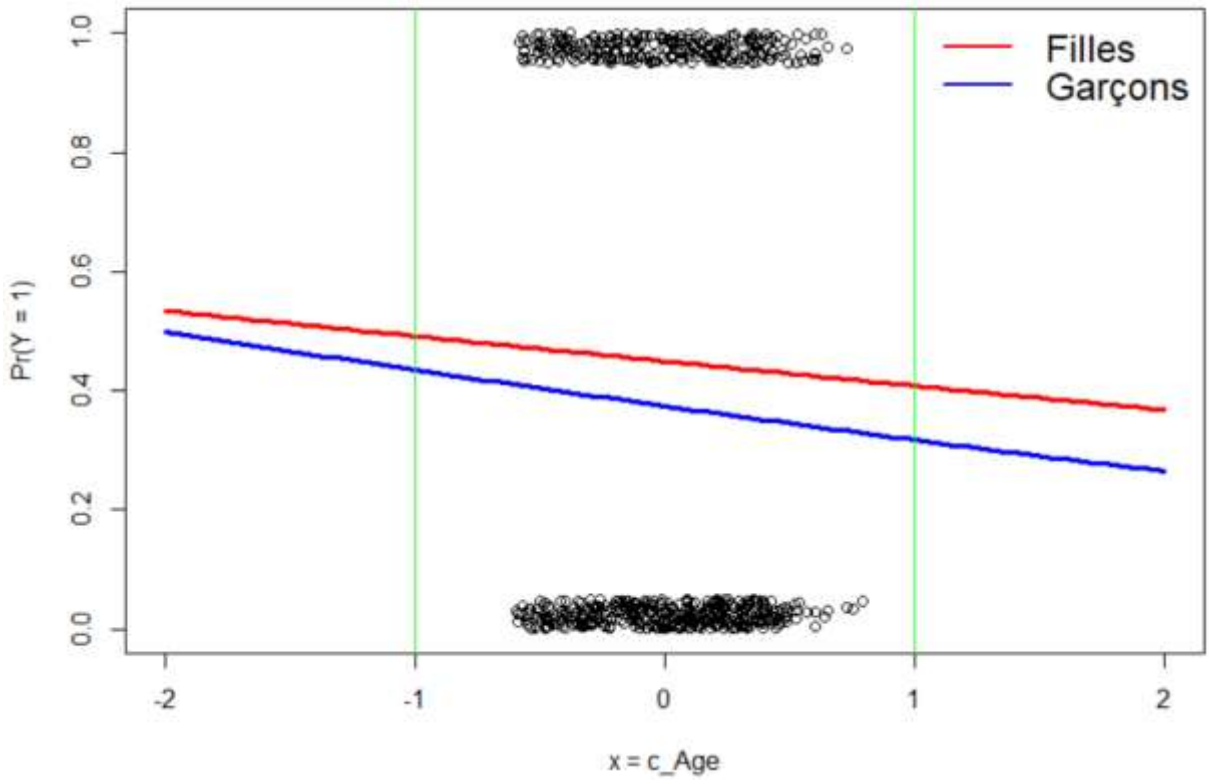

Figure 5.2a : Courbe de régression logistique Pr(Y=1) = logit^{-1}(-0.20386 - 0.31177*Sex - 0.16881*c_Age – 0.08376*(Sex:c_Age)) sur l'âge des élèves, en fonction de leur Sexe, sur l'intervalle [-2, +2]

Pour mieux visualiser la courbe logistique, je la prolonge sur l'intervalle [-25, +25] :

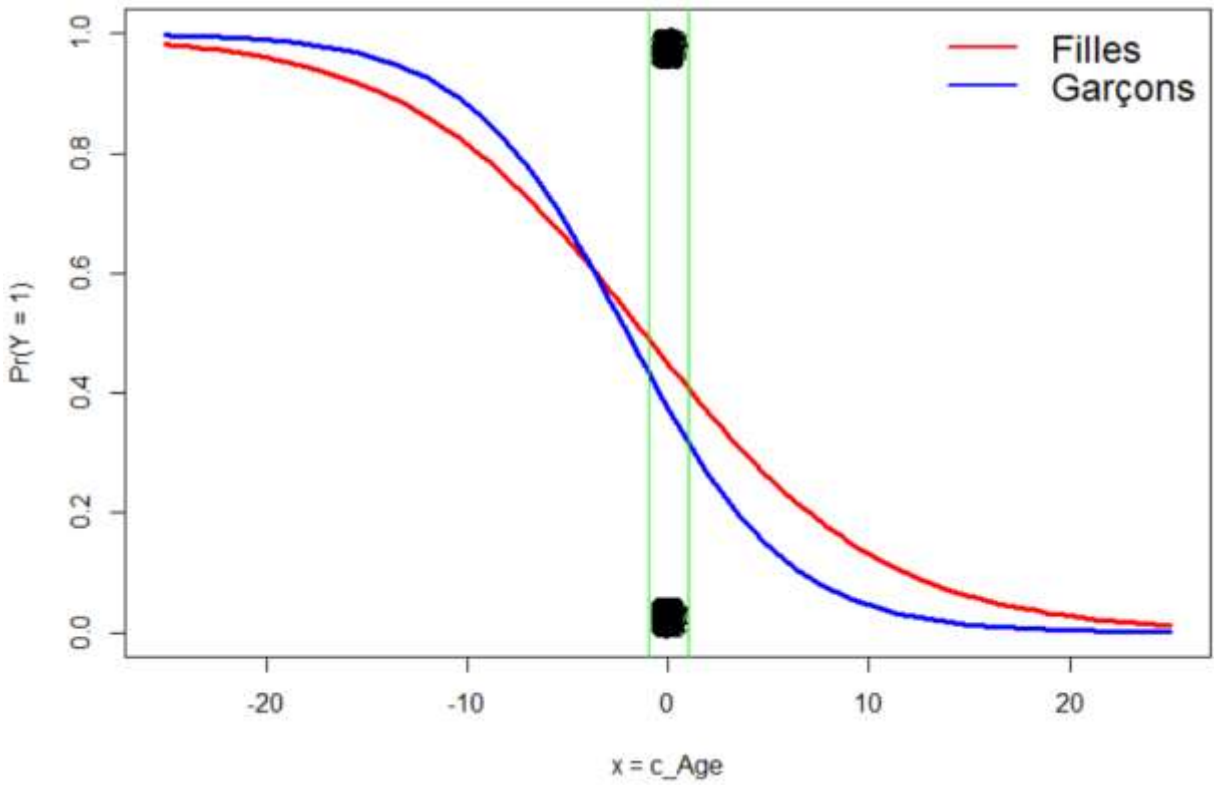

Figure 5.2b : Courbe de régression logistique Pr(Y=1) = logit^{-1}(-0.20386 - 0.31177*Sex - 0.16881*c_Age – 0.08376*(Sex:c_Age)) sur l'âge des élèves, en fonction de leur Sexe, prolongée sur l'intervalle [-25, +25]

Le calcul des pentes filles (= -0.042) et garçons (= -0.059) confirme l'impression visuelle et colle assez bien avec les coefficients respectifs divisés par 4 (= -0.042 et -0.063).

6. Exercice

Il s'agit de l'exercice 13.5 (p. 238) du livre de Gelman et al. (2021). Cet exercice réfère à une modélisation des décisions des ménages au Bangladesh concernant le changement de leur source d'eau potable, un exemple longuement analysé dans le livre. De nombreux puits utilisés pour l'eau potable au Bangladesh sont contaminés par de l'arsenic naturel. Une équipe de recherche a mesuré tous les puits et les a étiquetés comme "sûr" ou "insalubre". Les personnes dont les puits n'étaient pas sûrs ont été encouragées à se tourner vers des puits privés ou communautaires situés à proximité ou vers de nouveaux puits qu'elles ont construits elles-mêmes. Quelques années plus tard, les chercheurs sont retournés sur place pour savoir qui avait changé de puits. La régression logistique essaie de comprendre les facteurs prédictifs du changement (*switch*) de puits parmi les utilisateurs de puits insalubres: *switch* = 1 si le ménage a changé de puits et *switch* = 0 si le ménage a continué à utiliser son propre puits.

La régression logistique considère deux variables-inputs :
- La distance (en centaines de mètres) par rapport au puits sûr connu le plus proche, notée dist100;
- Le niveau d'arsenic du puits de l'enquêté.

```
glm(formula = switch ~ dist100 + arsenic, family=binomial(link="logit"))
            coef.est    coef.se
(Intercept)  0.00       0.08
dist100     -0.90       0.10
arsenic      0.46       0.04
n = 3020, k = 3
```

L'exercice demande de comparer deux personnes qui vivent à la même distance du puits le plus proche mais dont les niveaux d'arsenic diffèrent, l'une ayant un niveau d'arsenic de 0.5 et l'autre un niveau de 1.0. Quelle est approximativement (sachant que la probabilité de *switch* = 0.575) la probabilité additionnelle pour que cette deuxième personne change de puits ? Donnez une estimation approximative, une erreur-type et un intervalle à 95%.

Solution : La règle de la division par 4 fonctionne car les probabilités prédites sont proches de 50/50. La différence attendue dans Pr(switch), par changement unitaire dans le niveau d'arsenic, est approximativement de 0.46/4 = 0.11 (rappelons qu'avec la règle de la division par 4, on arrondit à l'unité inférieure) avec une erreur-type de 0.01. Mais, comme il s'agit d'une différence de 0.5, nous devons multiplier ces coefficients par 0.5, soit 0.055 avec une erreur type de 0.005 et un intervalle à 95 % de [0.055 +/- 2*0.005] = [0.045, 0.065].

Remarque : Cette solution a été fournie par Gelman (2019). Un commentateur, sur le blog de Gelman, conteste – ou en tout cas soutient qu'un calcul exact donne un résultat tout à fait différent – de celui de l'utilisation de la règle approximative de division par 4. Le calcul exact pourrait être:

y = inv.logit(0.00 -0.90*dist100 + 0.46*1) - inv.logit(0.00 -0.90*dist100 + 0.46*0.5)

Il a fixé dist100 à 3 et trouve alors une probabilité de *switch* de 0.018 qui n'est pas du tout de l'ordre de grandeur suggéré par le calcul approximatif avec la règle de division par 4.

Gelman le critique d'abord (à tort) parce qu'il n'a pas utilisé la fonction inv.logit (en fait, Gelman utilise invlogit après avoir renommé *plogis* : invlogit ← plogis), ce qui n'est pas le problème ; puis, Gelman argue que la valeur dist100 = 3 est extrême et que la plupart des valeurs dist100 sont inférieures à 0.50 (ce qui est vrai, mais il y en a quand même quelques-unes qui dépassent 3). Gelman reste sur sa position à propos de la règle de division par 4 et conclut que « dans cet exemple particulier, cela fonctionne parfaitement ». Comme je soutiens qu'un modèle n'est pas la réalité et, parfois pire, que tous les modèles sont faux (Box, 1976), j'adhère à cette position de Gelman.

7. Références

Box G.E.P., 1976. Science and statistics. *Journal of the American Statistical Association, 71*(356), 791-799. doi: 10.1080/01621459.1976.10480949

Dieudonné J., 1987. *Pour l'honneur de l'esprit humain : Les mathématiques aujourd'hui.* Paris : Hachette.

Fischer J.-P., en préparation. Mirror written digits: Is there a difference between boys and girls?

Fischer J.-P. & Charron C., 2009. Une étude de la dyscalculie à l'âge adulte. *Economie et Statistique*, N° 424-425, 87-101. Disponible à l'adresse : http://www.insee.fr/fr/ffc/docs_ffc/ES424-425E.pdf

Gelman A., 2019. https://statmodeling.stat.columbia.edu/2019/06/04/question-4-of-our-applied-regression-final-exam-and-solution-to-question-3/

Gelman A., Hill J. & Vehtari A., 2021. *Regression and Other Stories.* Cambridge: University Press. doi: 10.1017/9781139161879

Howell D.C., 1998. *Méthodes statistiques en sciences humaines.* Paris: De Boeck Université.

R Core Team, 2023. *R: A language and environment for statistical computing* (version 4.2.3). Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Stan Development Team, non daté. Stan user's guide (Version 2.34). Téléchargé le 29/01/2024 à l'adresse https://mc-stan.org/users/documentation/

(Note: Une version pdf du livre de Gelman et al. (2021), disponible gratuitement, a été publiée en 2022, puis 2023, avec, en plus, des corrections. Elle est récupérable sur le site de Vehtari, https://avehtari.github.io/ROS-Examples/. Mes indications des pages réfèrent à ce document car, en outre, le livre d'origine est mal paginé).