Université de Strasbourg
M2 - Statistique
Modèles linéaires généralisés

Année 2020/2021
Ségolen Geffray
geffray@math.unistra.fr

TD 1 : modèles à effets fixes

Exercice 1.
Soit $(\varepsilon_i)_{i=1,..,n}$ une suite de v.a.i.i.d. de loi $\mathcal{N}(0, \sigma^2)$. Parmi les modèles suivants, quels sont ceux qui correspondent (éventuellement après transformation) à un modèle de régression linéaire ?

1. $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$
2. $Y_i = \beta_0 + \beta_1 (X_i - \alpha)^2 + \varepsilon_i$
3. $Y_i = \beta_0 + \exp(\beta_1) X_i + \varepsilon_i$
4. $Y_i = \beta_0 + \beta_1 \exp(X_i) + \varepsilon_i$
5. $Y_i = \beta_0 \exp(\beta_1 X_i)\, |\varepsilon_i|,\ \beta_0 > 0,$
6. $Y_i = \beta_0 \exp(\beta_1 X_i) + \varepsilon_i$
7. $Y_i = \dfrac{\beta_0}{1 + \beta_1 X_i} + \varepsilon_i$
8. $Y_i = \alpha \dfrac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)} + \varepsilon_i$
9. $\dfrac{Y_i}{X_i^{(1)}} = \beta_0 + \beta_1 X_i^{(2)} + \varepsilon_i$

Exercice 2.

1. Ecrire l'équation et les hypothèses définissant un modèle de régression pour la variable réponse Y représentant le prix d'une bouteille d'un certain vin AOC en fonction de variables climatologiques : quantité cumulée de pluie durant l'hiver, température diurne journalière moyenne pendant l'été, quantité cumulée de pluie pendant le mois précédant les vendanges et l'année de récolte.
2. Ecrire l'équation et les hypothèses définissant un modèle de régression pour la variable réponse Y représentant l'occurrence d'un cancer de l'œsophage incluant différents prédicteurs : le taux sanguin de sélénium (en mg/L), la prise chronique d'aspirine, l'âge, le sexe, la consommation hebdomadaire moyenne d'alcool (exprimée en verres de vin).

Exercice 3.
Dans le contexte de la recherche agronomique, on mesure le rendement à l'hectare de champs plantés de pommes de terre obtenu dans trois conditions expérimentales différentes. Dix parcelles de 1 ha servent de contrôle et ne reçoivent aucun fertilisant. Dix autres parcelles de 1 ha reçoivent un fertilisant A tandis que dix autres parcelles de 1 ha reçoivent un fertilisant B. On recueille les données suivantes : en dessous des rendements figure la pluviométrie mensuelle correspondante exprimée en $mm \times (30j)$

rendement : contrôle	4.17	5.58	5.18	6.11	4.50	4.61	5.17	4.53	5.33	5.14
pluviométrie	52.3	58.1	62.1	59.0	55.3	48.9	62.5	50.1	57.6	54.2
rendement : fertilisant A	4.81	4.17	4.41	3.59	5.87	3.83	6.03	4.89	4.32	4.69
pluviométrie	52.4	57.1	59.1	39.0	55.0	48.2	62.6	51.1	55.6	52.2
rendement : fertilisant B	6.31	5.12	5.54	5.50	5.37	5.29	4.92	6.15	5.80	5.26
pluviométrie	51.3	57.8	60.1	49.2	52.3	47.9	60.5	58.6	57.7	56.2

On se demande si le rendement est influencé par le traitement du sol. Quelle est votre démarche ? Que se passe-t-il si cette expérience est effectuée avec une plante fourragère à croissance rapide de sorte qu'il y a trois récoltes sur chaque parcelle au cours de la belle saison ?

Exercice 4.
Dans le contexte de la recherche agronomique, on évalue l'effet des fertilisations potassique et magnésienne sur le rendement du blé. Le dispositif expérimental comporte quatre doses de potassium et quatre doses de magnésium (en kg/ha). L'allocation des doses sur les différentes parcelles de plantation a été effectuée par tirage au sort. Les résultats de rendement (en kg/ha) sont les suivants. Quelle est votre analyse ?

		Dose K			
		0	80	160	240
Dose Mg	0	203	253	257	261
	20	206	247	270	267
	40	204	244	266	276
	80	179	252	265	259

Exercice 5.
Un agronome étudie un fertilisant d'origine naturelle pour plantes décoratives ou potagères. Il souhaite en évaluer l'efficacité et l'universalité afin d'émettre des recommandations à destination des futurs utilisateurs de ce fertilisant. Pour cela, il mesure le rendement (en kg) obtenu en plantant différentes parcelles d'une surface d'1 m^2 de plantes décoratives ou potagères, avec ou sans fertilisant. Les plantes décoratives sont de trois types possibles : violacées (catégorie A, par exemple pensées), géraniacées (catégorie B, par exemple géranium) ou astéracées (catégorie C, par exemple soucis). Les légumes sont de trois types possibles : apiacées (catégorie A, par exemple carottes), chénopodiacées (catégorie B, par exemple épinards) ou cucurbitacées (catégorie C, par exemple concombres). Il recueille les données suivantes :

plantes décoratives			plantes potagères		
catégorie A	catégorie B	catégorie C	catégorie A	catégorie B	catégorie C
sans/avec	sans/avec	sans/avec	sans/avec	sans/avec	sans/avec
1.0 / 1.2	1.2 / 1.5	4.1 / 4.3	8.4 / 9.1	6.2 / 6.9	3.0 / 4.1
0.9 / 0.9	1.1 / 1.4	5.1 / 5.3	9.3 / 9.6	5.3 / 5.8	2.0 / 3.8
0.9 / 1.3	1.5 / 1.8	5.2 / 5.6	7.2 / 8.2	7.1 / 7.3	2.3 / 3.8
0.9 / 1.4	2.1 / 2.4	4.8 / 4.8	9.0 / 9.1	7.0 / 7.4	3.2 / 3.3
1.1 / 1.0	2.0 /	3.2 / 3.1	8.2 / 8.2	6.8 / 7.1	4.5 / 5.6
1.2 / 1.8		3.6 / 3.9	5.1 /	5.9 / 6.8	/ 6.2
0.9 / 1.8		/ 5.2	7.0 /	7.1 / 7.4	/ 5.6

Comment tester l'efficacité du fertilisant et l'universalité du terreau ?

Exercice 6.
Dans le contexte de la recherche en biologie cellulaire, on évalue l'effet de deux nutriments A et B sur le nombre de colonies de cellules se développant dans une boîte de Pétri. Le dispositif expérimental comporte quatre doses de nutriment A et quatre doses de nutriment B (en μg). Les nombres associés de colonies de cellules sont les suivants.

		dose A			
		0	20	40	80
Dose B	0	2	2	3	5
	20	0	2	0	1
	40	1	0	2	2
	80	1	2	2	4

Quelle est votre analyse ? Ecrire l'équation et les hypothèses définissant un modèle que l'on pourrait ajuster sur ces données. Comment tester H_0 : les nutriments ont des effets similaires ? Comment tester H_0 : les nutriments A et B ont un effet sur le nombre de colonies apparues ?

Exercice 7.
La sclérose en plaques (SEP) est une affection inflammatoire chronique du système nerveux central (SNC) survenant chez l'adulte jeune. Le comptage des zones cérébrales apparaissant lésées à l'IRM peut être utilisé comme biomarqueur dans cettte affection. On réalise une étude chez 188 patients nouvellement diagnostiqués et sans lien entre eux afin d'étudier l'influence potentielle des covariables suivantes sur les comptages obtenus :

— prédisposition génétique (oui/non),
— taux de vitamine D (en ng/mL),
— sexe (H/F),
— âge,
— exposition au tabac (nulle, modérée, importante, très importante).

On admettra qu'il n'y pas lieu de prendre en compte de termes d'interaction entre covariables, ni d'effectuer de transformation des covariables.

1. Ecrire les équations et hypothèses d'un modèle adapté à ce travail.
2. Comment évaluer la qualité de l'ajustement du modèle proposé aux données ?
3. De quelle(s) alternative(s) au modèle initial de la question 1. dispose-t-on ? Préciser le problème que tente de résoudre la(les) alternative(s) proposée(s).
4. Dans le cadre du modèle initial de la question 1., de quel(s) test(s) dispose-t-on pour évaluer l'impact du tabagisme sur la réponse ? Mettre en oeuvre.
5. Supposons maintenant que, dans le cadre du modèle initial de la question 1., on dispose de plusieurs observations distantes de quelques mois pour chacun des patients. Comment ré-écrire le modèle ?

Exercice 8.
Considérons la loi gaussienne inverse dont la loi admet la densité suivante par rapport à la

mesure de Lebesgue :

$$f_{\mu,\sigma^2}(y) = \frac{1}{\sqrt{2\pi\sigma^2 y^3}} \exp\left(-\frac{(y-\mu)^2}{2\mu^2\sigma^2 y}\right), \qquad y > 0, \mu > 0, \sigma > 0\,.$$

1. Montrer que la loi gaussienne inverse appartient à la famille exponentielle.
2. Ecrire les équations et les hypothèses définissant un modèle de régression linéaire généralisé utilisant la loi gaussienne inverse et le lien canonique et incluant 4 prédicteurs, X_1 et X_2 quantitatifs, X_3 et X_4 qualitatifs à respectivement 3 et 4 modalités avec un terme d'interaction entre X_1 et X_3.
3. De quel(s) test(s) d'adéquation dispose-t-on pour évaluer l'ajustement du modèle ? Mettre en œuvre.
4. De quel(s) test(s) d'adéquation dispose-t-on pour évaluer la nécessité de chacun des régresseurs ? Mettre en œuvre.
5. Comment tester si l'impact de $X^{(1)}$ sur la variable réponse Y est identique à l'impact de $X^{(2)}$?

Exercice 9.
Considérons un modèle de régression de Poisson avec le lien canonique incluant deux variables explicatives quantitatives $X^{(1)}$ et $X^{(2)}$ ainsi que deux variables explicatives qualitatives $X^{(3)}$ et $X^{(4)}$ comportant respectivement 3 et 4 modalités. Considérons dans le modèle une interaction entre $X^{(2)}$ et $X^{(3)}$.

1. Ecrire l'équation et les hypothèses du modèle linéaire généralisé correspondant.
2. De quel(s) test(s) dispose-t-on pour évaluer la qualité de l'ajustement d'un tel modèle à des données ?
3. De quel(s) test(s) dispose-t-on pour évaluer la nécessité de l'inclusion de $X^{(3)}$? Mettre en œuvre.
4. Supposons maintenant qu'à la suite de l'analyse précédente, on a écarté $X^{(3)}$ du modèle. Réécrire le modèle obtenu. De quel(s) test(s) dispose-t-on pour évaluer la nécessité de l'inclusion de $X^{(1)}$? Mettre en œuvre.
5. Supposons maintenant que, dans le cadre du modèle initial, on dispose de 5 observations successives pour chacun des individus. Comment réécrire le modèle ?

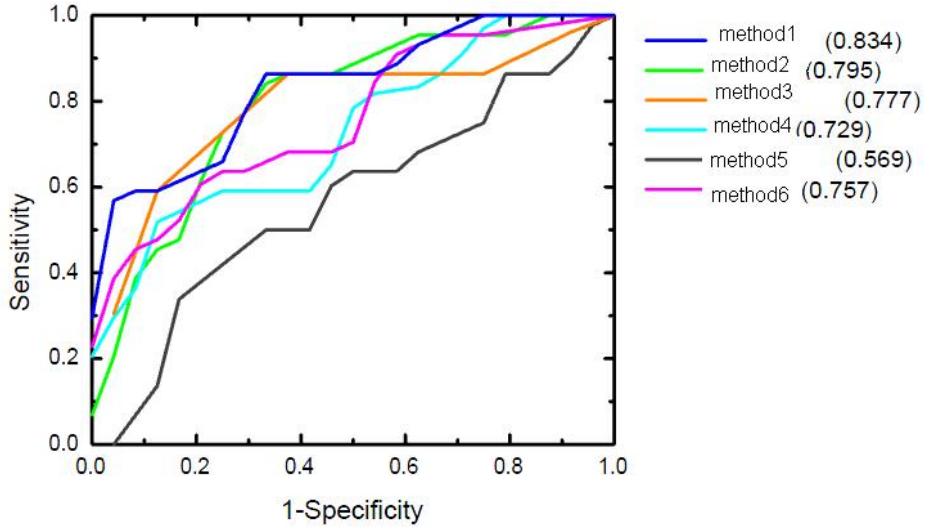
Exercice 10.
Un médecin souhaite essayer de prévoir la réponse ou la non-réponse d'un patient à un traitement sur la base du dosage d'un ou deux taux sanguin(s). On admettra qu'il n'y a pas d'interaction entre ces deux taux sanguins. Le médecin dispose d'un échantillon de patients sur lesquels il effectue un prélèvement sanguin pour mesurer chaque taux et auxquels il administre le traitement. Notons $X^{(1)}$ la variable aléatoire qui représente le premier taux sanguin et $X^{(2)}$ la variable aléatoire qui représente le second taux sanguin. Le médecin souhaite comparer les trois méthodes suivantes :

— prévoir la réponse au traitement sur la base de $X^{(1)}$ seulement,
— prévoir la réponse au traitement sur la base de $X^{(2)}$ seulement,
— prévoir la réponse au traitement sur la base de $X^{(1)}$ et $X^{(2)}$.

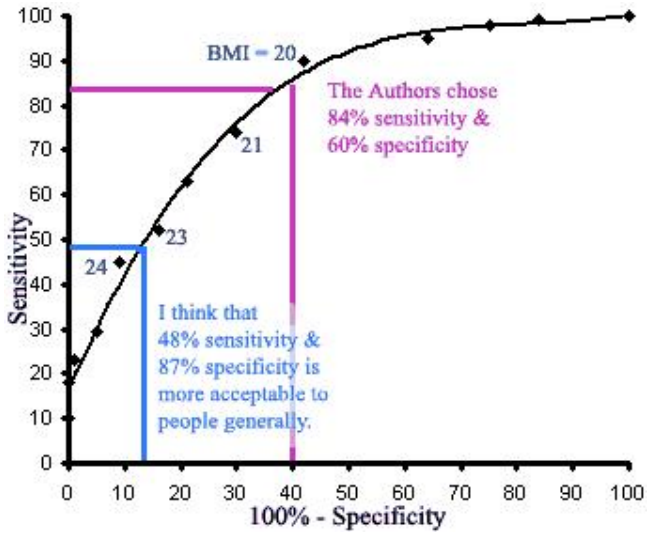
Comment procède-t-il ?

Exercice 11.
D'après le Dictionnaire de Médecine Flammarion, l'obésité est un état caractérisé par un excès de masse adipeuse répartie de façon généralisée dans les diverses zones grasses de l'organisme. L'obésité est souvent appréciée par le poids mais il n'y a pas de stricte équivalence entre poids et obésité puisque dans le poids interviennent la masse grasse mais aussi le tissu osseux, l'eau et le muscle. Il existe différents tests diagnostiques pour caractériser un seuil d'alerte relatif à l'obésité morbide à partir duquel on risque de voir apparaître une morbidité secondaire liée à différents types de complications, par exemple les méthodes BMI (Body Mass Index) ou DEXA (Dual-Energy X-ray Absorptiometry).

1. Afin de comparer ces différentes méthodes, sont tracées sur un même graphique les courbes ROC correspondant aux différentes méthodes. Entre parenthèses, sont indiquées les AUC (Area Under Curve) correspondantes. Y-a-t-il une méthode plus performante que les autres ?



2. Curtin F., Morabia A., Pichard C. & Slosman D.O. dans un article paru en 1997 dans la revue *J. Clin. Epidemiol.* ont travaillé à l'établissement d'un seuil de BMI au delà duquel un patient est considéré obèse. Que penser du commentaire de leurs résultats effectué par un de leur confrère ?



Exercice 12.
Une étude est réalisée afin de voir comment un nombre p de variables cliniques (mesurées par un médecin) affectent le ressenti de l'état de santé de patients diabétiques (pourcentage calculé à partir d'un questionnaire rempli par le patient). Afin de modéliser la loi conditionnelle des pourcentages, le statisticien utilise une loi beta dont il paramètre ainsi la densité par rapport à la mesure de Lebesgue :

$$f_{\mu,\phi}(y) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \quad 0<\mu<1, \phi>0, y\in(0,1),$$

de sorte que $\mathbb{E}[Y]=\mu$ et que $\mathrm{Var}(Y)=\dfrac{\mu(1-\mu)}{1+\phi}$. Il utilise également la fonction de lien *logit* pour relier l'espérance conditionnelle au prédicteur linéaire.

1. En supposant que toutes les variables cliniques recueillies par le médecin sont quantitatives et sans interactions entre elles et en supposant que les patients sont indépendants entre eux, écrire les équations et hypothèses définissant le modèle de régression beta utilisé par le statisticien. S'agit-il d'un modèle GLM ? Comment peut-on faire de l'inférence avec ce modèle ?
2. En réalité, le pourcentage quantifiant le ressenti de l'état de santé des patients est à valeurs dans $(0,1]$. Proposer une solution pour que le modèle puisse prendre en compte la borne droite de l'intervalle $(0,1]$.
3. Le statisticien utilise ensuite son modèle pour une nouvelle étude. Cette fois, les mêmes variables que précédemment sont recueillies lors d'une étude multicentrique sur des patients hospitalisés dans 2 ou 3 services hospitaliers situés dans 3 ou 4 hopitaux dans 4 villes différentes. Que fait le statisticien ?

Exercice 13.
Ecrire les équations et hypothèses définissant les modèles suivants ajustés avec le logiciel `R`.

1. `glm(y~x1+x2+log(x3),family=poisson)`
2. `glm(y~x1*x2+log(x3),family=poisson)`
3. `glm(y~x1+x2+x3,family=quasipoisson)`
4. `glm(y~x1+x2+log(x3),family=binomial)`
5. `glm(y~x1+x2+log(x3),family=binomial(link="probit"))`
6. `glm(y~x1+x2+log(x3),family=binomial(link="cloglog"))`
7. `glm(y~x1+x2+x3,family=quasi(link = "identity", variance = "constant"))`
8. `glm(y~x1+x2+x3,family=quasi(link = "identity", variance = "mu^3"))`
9. `glm(y~x1*x2+x3*x4),family=poisson)`
10. `glm(y~x1+x2+x3,family=quasi(link = "probit", variance = "mu(1-mu)"))`

Exercice 14.
Considérons le jeu de données `dicentric` disponible dans le logiciel `R` qui contient 27 observations indépendantes de 4 variables :

— `ca` : nombre d'anomalies chromosomiques comptées dans les cellules soumises au rayonnement Gamma,

— `cell` : nombre de cellules soumises au rayonnement Gamma,
— `doserate` : taux de rayonnement Gamma,
— `doseamt` : niveau de rayonnement Gamma.

1. Ecrire les équations et hypothèses définissant les modèles ajustés ci-dessous avec le logiciel `R` après avoir introduit les variables aléatoires nécessaires à la formalisation du problème. Quels sont leurs avantages/inconvénients comparés ?
   (a) `m<-lm(ca/cells~log(doserate)*factor(doseamt),data=dicentric)`
   (b) `m<-glm(ca ~log(cells)+log(doserate)*factor(doseamt),family=poisson, data=dicentric)`
   (c) `m<-glm(ca ~offset(log(cells))+log(doserate)*factor(doseamt),family=poisson, data=dicentric)`
2. De quel(s) diagnostics/test(s) dispose-t-on pour évaluer la qualité de l'ajustement de tels modèles aux données ?
3. De quel(s) diagnostics/test(s) dispose-t-on pour évaluer la nécessité de l'inclusion de `doserate` ?
4. Afin d'augmenter la puissance de l'étude, on recommence l'étude en incluant la fratrie des 27 personnes initiales. Comment réécrire les modèles ?

Exercice 15.
Ecrire les équations et hypothèses définissant le modèles ajustés ci-dessous avec le logiciel `R` après avoir introduit les variables aléatoires nécessaires à la formalisation du problème. Quels sont les apports respectifs de ces modèles ?

```
myfit1 <- glm(art~fem+mar+kid5+phd+ment, family=poisson,  data=bioChemists)
myfit2 <- glm(art~fem+mar+kid5+phd+ment, family=quasipoisson,  data=bioChemists)
myfit3 <- glm.nb(art~fem+mar+kid5+phd+ment,data=bioChemists)
myfit4 <- zeroinfl(art~fem+mar+kid5+phd+ment|1, data=bioChemists)
myfit5 <- zeroinfl(art~fem+mar+kid5+phd+ment|1, dist="negbin", data=bioChemists)
myfit6 <- zeroinfl(art~fem+mar+kid5+phd+ment|fem+mar+kid5+phd+ment, data=bioChemists)
myfit7 <- zeroinfl(art~fem+mar+kid5+phd+ment|fem+mar+kid5+phd+ment, dist="negbin",
data=bioChemists)
```

Exercice 16.
Pour $i = 1, ..., n$, on compte Y_i le nombre d'hippopotames observés en un site i donné. On relève en même temps un certain nombre de variables environnementales (extraction d'or ou autre exploitation minière à proximité, fréquence annuelle moyenne des feux de brousse, déversement de substances polluantes dans l'air, le sol ou l'eau, surface de prairie dans un rayon de 1km autour du point d'eau), dans le but d'expliquer Y_i. Supposons que les mesures effectuées en n différents sites sont indépendantes. On suspecte la présence de trop nombreux '0' dans les données due au fait que l'observateur n'a pas vu les hippopotames car ils étaient sous une eau boueuse ou car seuls leurs yeux émergeaient, les rendant indistinguables des crocodiles pour des observateurs distants. Proposer des modèles susceptibles de s'ajuster sur ces données. Quelle est votre démarche ?

Exercice 17.
On cherche à expliquer le nombre de fois par heure, noté Y, où une baleine bleue remonte à la surface pour respirer en fonction de plusieurs variables (température de l'océan au lieu de mesure, présence de perturbateurs hormonaux au lieu de mesure, profondeur maximale moyenne atteinte lors des plongées, présence de période de repos d'un des deux hémisphères du cerveau pendant l'observation). Supposons que les différents individus, observés chacun pendant une heure, sont indépendants. Notons également que la durée maximale de plongée enregistrée pour une baleine bleue est 36 minutes. Quel(s) modèle(s) proposer ? Quelle analyse effectuer ?

Exercice 18.
On réalise une étude sur 282 sujets masculins afin d'expliquer les altérations cellulaires chez des individus exposés à des radiations gamma. L'altération cellulaire est caractérisée par le nombre de micronuclei présents dans une cellule. Une cellule saine n'en comporte pas. A l'inverse, plus le nombre de micronuclei présents dans une cellule est élevé, plus la cellule dysfonctionne. Pour chacun des 282 sujets, sont relevées les variables suivantes : nombre de cellules analysées, nombre de cellules avec présence de micronuclei, dose cumulée de radiation gamma reçue, âge, IMC (indice de masse corporelle), carences nutritionnelles (aucune, moyennement, beaucoup), qualité du sommeil (bonne, passable, mauvaise).

1. Proposer un modèle de régression adapté à ces données en supposant qu'il y a potentiellement interaction entre la dose cumulée de radiations gamma et les carences nutritionnelles.
2. De quel(s) diagnostics(s) dispose-t-on pour évaluer la qualité de l'ajustement d'un tel modèle à des données ?
3. De quel(s) test(s) dispose-t-on pour évaluer l'impact de la dose cumulée de radiations gamma sur l'altération cellulaire ? Mettre en œuvre.
4. Déterminer la dose associée à une probabilité de 50% de risque que la moitié de ses N_0 cellules analysées soient altérées pour un homme de 50 ans, d'IMC égal à 22, sans carences nutritionnelles et bon dormeur.

Exercice 19.
On dispose d'une solution contenant des bactéries dont la concentration en bactéries, notée ρ_0, est inconnue. On élabore le dispositif suivant. On effectue n dilutions de la solution initiale, en diluant à chaque fois d'un facteur constant égal à deux. A chaque itération, on prélève N volumes de solution, tous égaux à v. Chacun de ces volumes est déposé sur une plaque de gélose contenant un milieu nutritif propice au développement de ces bactéries. Au bout de 72h, on observe si une colonie de bactéries est apparue ou non. Estimer la concentration de la solution initiale.
Indication : s'inspirer du modèle de régression binômiale avec lien `cloglog` et utiliser le fait que la loi du nombre de bactéries déposées par plaque de gélose est une loi de Poisson.

Exercice 20.
Afin de démontrer la toxicité d'un herbicide sur la faune, des scarabées sont soumis pendant 72h à une certaine dose (exprimée en mg, sur une échelle logarithmique) de l'herbicide incriminé. Pour cela, on dispose de 10 boîtes contenant des scarabées. Au bout de 72h, on compte le

nombre de scarabées morts (ou amorphes ie ne répondant pas à l'agitation d'une tapette) dans chacune des boites. On recueille les données suivantes :

log(dose)	nombre de scarabées	nombre de morts
1.3863	56	6
1.4881	62	13
1.5899	63	18
1.6917	60	28
1.7935	61	39
1.8953	59	45
1.9972	61	51
2.0990	61	55
2.2008	60	59
2.3026	64	64

Le travail présenté ci-dessous est effectué sur ces données au moyen du logiciel `R`.
NB : la variable `dose` dans le code ci-dessous est exprimée sur l'échelle logarithmique.

```
> mymodel1<-glm(dead/nb~dose,family=binomial,weights=nb)
> summary(mymodel1)

Call:
glm(formula = dead/nb ~ dose, family = binomial, weights = nb)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.73140  -0.30268   0.06081   0.41759   1.61176

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -11.3027     0.9097  -12.43   <2e-16 ***
dose          6.5969     0.5172   12.76   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 296.1843  on 9  degrees of freedom
Residual deviance:   5.0163  on 8  degrees of freedom
AIC: 43.919

Number of Fisher Scoring iterations: 4

> mymodel2<-glm(dead/nb~dose,family=binomial(link="probit"),weights=nb)
> summary(mymodel2)

Call:
glm(formula = dead/nb ~ dose, family = binomial(link = "probit"),
```

```
    weights = nb)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8284  -0.3673   0.1133   0.3595   1.2369

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -6.6009     0.4823  -13.69   <2e-16 ***
dose          3.8486     0.2716   14.17   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 296.1843  on 9  degrees of freedom
Residual deviance:   3.5314  on 8  degrees of freedom
AIC: 42.434

Number of Fisher Scoring iterations: 4

> mymodel3<-glm(dead/nb~dose,family=binomial(link="cloglog"),weights=nb)
> summary(mymodel3)

Call:
glm(formula = dead/nb ~ dose, family = binomial(link = "cloglog"),
    weights = nb)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1391  -0.4374  -0.1628   0.5424   1.1181

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -7.2348     0.5530  -13.08   <2e-16 ***
dose          3.9378     0.2946   13.37   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 296.1843  on 9  degrees of freedom
Residual deviance:   5.2659  on 8  degrees of freedom
AIC: 44.169

Number of Fisher Scoring iterations: 4

>
```

```
> rp1<-residuals(mymodel1, type="pearson")
> sum(rp1^2)/(10-2)
[1] 0.4520803
> deviance(mymodel1)/(10-2)
[1] 0.6270417
>
> rp2<-residuals(mymodel2, type="pearson")
> sum(rp2^2)/(10-2)
[1] 0.3482475
> deviance(mymodel2)/(10-2)
[1] 0.4414264
>
> rp3<-residuals(mymodel3, type="pearson")
> sum(rp3^2)/(10-2)
[1] 0.6525673
> deviance(mymodel3)/(10-2)
[1] 0.658232
>
>
> mean(rp1)
[1] 0.08782847
> mean(rp2)
[1] 0.05436226
> mean(rp3)
[1] -0.04956657
>
>
> library(statmod)
> rq1<-qresiduals(mymodel1)
> rq2<-qresiduals(mymodel2)
> rq3<-qresiduals(mymodel3)
>
>
> x11()
> par(mfrow=c(1,3))
> plot(rq1,main="mymodel1")
> abline(h=0)
> plot(dose,rq1,main="mymodel1")
> abline(h=0)
> plot(dead/nb,rq1,main="mymodel1")
> abline(h=0)
>
> x11()
> par(mfrow=c(1,3))
> plot(rq2,main="mymodel2")
> abline(h=0)
> plot(dose,rq2,main="mymodel2")
> abline(h=0)
```

```
> plot(dead/nb,rq2,main="mymodel2")
> abline(h=0)
>
> x11()
> par(mfrow=c(1,3))
> plot(rq3,main="mymodel3")
> abline(h=0)
> plot(dose,rq3,main="mymodel3")
> abline(h=0)
> plot(dead/nb,rq3,main="mymodel3")
> abline(h=0)
>
>
> library(MuMIn)
> AICc(mymodel1)
[1] 45.63359
> AICc(mymodel2)
[1] 44.14867
> AICc(mymodel3)
[1] 45.88311
```

Le tracé effectué dans le code est présenté en figures 1, 2 et 3. On appelle dose létale médiane

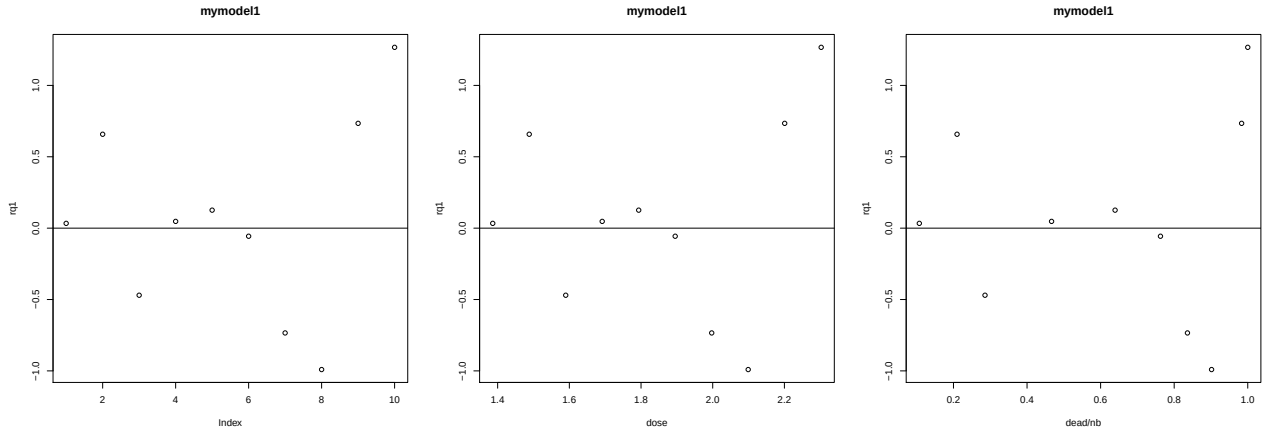


FIGURE 1 – Tracé des résidus quantiles randomisés obtenus avec le 1er modèle ajusté.

la dose de produit associée à une probabilité de décès de 50%. On la note ED_{50}. Estimer ED_{50}.

Exercice 21.
Cet exercice est librement inspiré d'un article de Deutsch et Piegorsch, paru dans *Biometrics* en 2012. On souhaite évaluer l'impact du DDT et des nanoparticules de dioxide de titane sur la proportion de cellules hépatiques humaines présentant des micronuclei après exposition à une certaine dose de DDT (pesticide organochloré largement utilisé et suspecté d'être cancérigène) et à une certaine dose de nanoparticules de dioxide de titane (agent fortement oxydant, suspecté

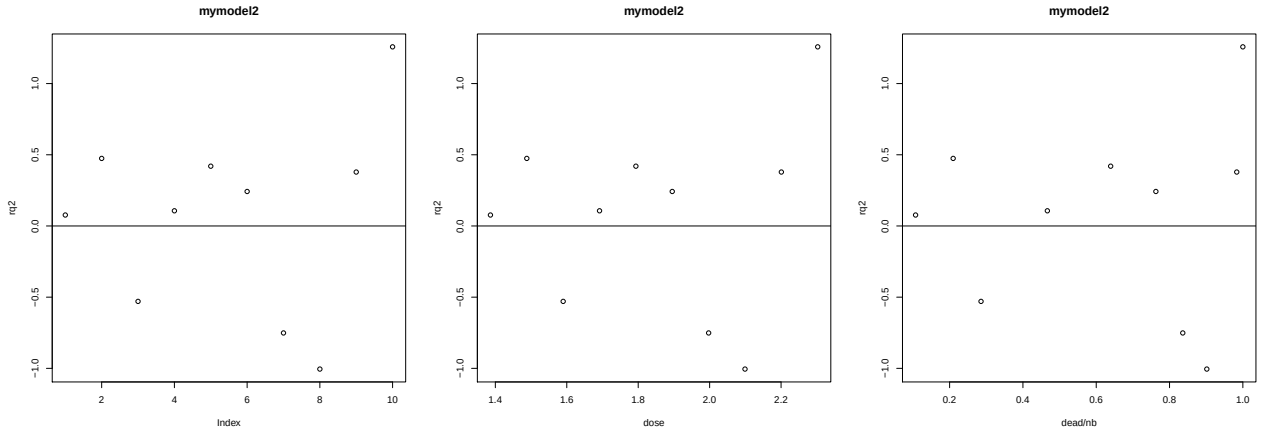


FIGURE 2 – Tracé des résidus quantiles randomisés obtenus avec le 2ème modèle ajusté.

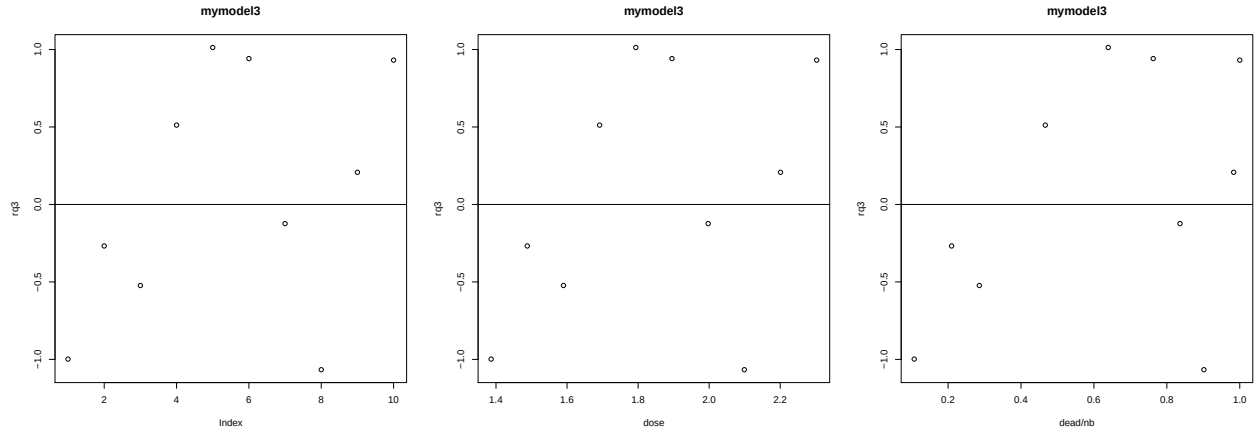


FIGURE 3 – Tracé des résidus quantiles randomisés obtenus avec le 3ème modèle ajusté.

de pouvoir réagir avec un grand nombre de molécules organiques). Les données sont présentées dans le tableau ci-dessous :

		Dose de nanoparticules de dioxide de titane en μg/L			
		0	0.01	0.1	1.0
Dose de DDT (μmol/L)	0	59/3000	65/3000	70/3000	67/3000
	0.001	67/3000	75/3000	83/3000	84/3000
	0.01	76/3000	87/3000	96/3000	83/3000
	0.1	94/3000	107/3000	110/3000	117/3000

1. Ecrire les équations et hypothèses d'un modèle permettant d'étudier l'influence du DDT et des nanoparticules de dioxide de titane sur sur la proportion de cellules hépatiques humaines présentant des micronuclei. Pour cela, on incluera un terme d'interaction entre la dose de DDT et la dose de nanoparticules de dioxide de titane, ainsi que des termes d'ordre deux, à la fois pour la dose de DDT et pour la dose de nanoparticules de dioxide de titane.
2. Comment tester la nécessité de l'inclusion du terme d'interaction ? Mettre en oeuvre.

3. On note $ED_{50}(x_0^{(1)})$ la dose de nanoparticules de dioxide de titane associée à une probabilité d'apparition de micronuclei de 50% pour une dose $x_0^{(1)}$ de DDT. Comment estimer $ED_{50}(x_0^{(1)})$ lorsqu'on a conservé le modèle initial de la question 1. ? Mettre en oeuvre.
4. Que pourriez-vous proposer pour garantir la positivité de l'estimation de $ED_{50}(x_0^{(1)})$?

Exercice 22.

Une étude est réalisée sur la rémission à 5 ans chez des patients atteints de cancer de la prostate à partir d'un jeu de données stocké dans `mydata`. La variable réponse `y` est binaire et indique la rémission à 5 ans (codée par 1) ou l'absence de rémission à 5 ans (codée par 0). Sont également disponibles 5 prédicteurs candidats qui sont les suivants :

— `age` : âge du patient (en années) au moment du diagnostic
— `aps` : taux d'acide phosphatase sérique (APS) (en ng/ml)
— `stade` : stade du cancer évalué à parties des données d'imagerie et/ou de biopsies (1 = localisé = limité à la prostate, 2 = localement avancé = étendu aux organes adjacents mais sans atteinte ganglionnaire ni métastase, 3 = atteinte ganglionnaire pelvienne, 4 = cancer métastatique)
— `gleason` : score de Gleason (grades de 4 à 10 : plus le grade est élevé, plus la tumeur est agressive)
— `trt` : traitement attribué à l'individu (1= ablation chirurgicale de la prostate, 2 = radiothérapie externe, 3 = curiethérapie, 4 = hormonothérapie)

Un graphique exploratoire est présenté en figure 4 tandis qu'un début d'analyse avec le logiciel `R` est reproduit ci-dessous et en figures 5 et 6.

```
> head(mydata)
  y  age  aps stade gleason trt
1 1 67.8  5.0     3       7   4
2 1 75.9 12.4     3       7   3
3 1 80.7  7.0     3       6   3
4 1 72.4 24.0     1       5   4
5 0 74.1 44.1     3       5   3
6 1 63.4 11.2     3       7   3
> str(mydata)
'data.frame': 568 obs. of  6 variables:
 $ y      : num  1 1 1 1 0 1 1 0 1 1 ...
 $ age    : num  67.8 75.9 80.7 72.4 74.1 63.4 77.9 47.6 70.7 50.6 ...
 $ aps    : num  5 12.4 7 24 44.1 11.2 26.3 40.2 11.5 26.7 ...
 $ stade  : Factor w/ 4 levels "1","2","3","4": 3 3 3 1 3 3 3 2 2 4 ...
 $ gleason: num  7 7 6 5 5 7 6 7 5 7 ...
 $ trt    : Factor w/ 4 levels "1","2","3","4": 4 3 3 4 3 3 2 1 4 4 ...
> summary(mydata)
       y                age              aps         stade      gleason     trt
 Min.   :0.0000   Min.   :45.10   Min.   : 5.00   1: 70   Min.   :4.00   1:128
 1st Qu.:0.0000   1st Qu.:54.90   1st Qu.:14.00   2:196   1st Qu.:5.00   2:172
 Median :1.0000   Median :65.55   Median :23.75   3:232   Median :6.00   3:125
 Mean   :0.6285   Mean   :64.81   Mean   :24.13   4: 70   Mean   :5.84   4:143
 3rd Qu.:1.0000   3rd Qu.:74.30   3rd Qu.:33.92           3rd Qu.:7.00
 Max.   :1.0000   Max.   :85.00   Max.   :44.90           Max.   :9.00
```

```
> x11()
> par(mfrow=c(2,3))
> scatter.smooth(y~age)
> scatter.smooth(y~aps)
> scatter.smooth(y~stade)
There were 20 warnings (use warnings() to see them)
> scatter.smooth(y~gleason)
There were 20 warnings (use warnings() to see them)
> scatter.smooth(y~trt)
There were 23 warnings (use warnings() to see them)

> myfit<- glm(y~age+aps+stade+gleason+trt+age:trt+stade:trt ,family="binomial",
+              data=mydata)
> summary(myfit)

Call:
glm(formula = y ~ age + aps + stade + gleason + trt + age:trt +
    stade:trt, family = "binomial", data = mydata)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.90900  -0.27831   0.00011   0.19798   2.80209

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    2.70633    2.17557   1.244 0.213512
age            0.03058    0.02412   1.268 0.204951
aps           -0.08290    0.01677  -4.942 7.71e-07 ***
stade2         0.54786    0.82262   0.666 0.505417
stade3        -0.74601    0.82125  -0.908 0.363676
stade4       -18.18893 2256.60465  -0.008 0.993569
gleason       -0.75402    0.17137  -4.400 1.08e-05 ***
trt2          -5.51720    3.38476  -1.630 0.103099
trt3          -8.06279    3.17305  -2.541 0.011053 *
trt4          19.25608 2336.37768   0.008 0.993424
age:trt2       0.18690    0.05384   3.471 0.000518 ***
age:trt3       0.21781    0.05135   4.242 2.21e-05 ***
age:trt4       0.03906    0.06660   0.586 0.557604
stade2:trt2   -0.82334    1.76035  -0.468 0.639989
stade3:trt2   -0.26466    1.73862  -0.152 0.879009
stade4:trt2   13.79822 2256.60530   0.006 0.995121
stade2:trt3   -5.09313    1.50455  -3.385 0.000711 ***
stade3:trt3   -5.38474    1.56515  -3.440 0.000581 ***
stade4:trt3    9.00579 2256.60542   0.004 0.996816
stade2:trt4  -16.84330 2336.37373  -0.007 0.994248
stade3:trt4    1.18152 2682.36208   0.000 0.999649
stade4:trt4   18.04340 4117.98527   0.004 0.996504
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 749.46  on 567  degrees of freedom
Residual deviance: 248.69  on 546  degrees of freedom
AIC: 292.69

Number of Fisher Scoring iterations: 18


> library(statmod)
> rq<-qresiduals(myfit)
>
> x11()
> par(mfrow=c(2,3))
> plot(age,rq)
> plot(aps,rq)
> plot(stade,rq)
> plot(gleason,rq)
> plot(trt,rq)
> plot(y, rq)

> drop1(myfit,test="LRT",trace=TRUE)
Single term deletions

Model:
y ~ age + aps + stade + gleason + trt + age:trt + stade:trt
          Df Deviance    AIC    LRT  Pr(>Chi)
<none>         248.69 292.69
aps        1   279.04 321.04 30.352 3.604e-08 ***
gleason    1   271.27 313.27 22.583 2.012e-06 ***
age:trt    3   275.88 313.88 27.193 5.364e-06 ***
stade:trt  9   281.31 307.31 32.623 0.0001553 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> p.hat<-predict(myfit,type="response")
>
> pc<-sort(unique(p.hat))
> mat<-matrix(NA,length(pc),2)
> for(i in 1:length(pc))
+ {
+     t1<-table(factor(p.hat>pc[i],levels=c(F,T)),y)
+     mat[i,]<-c(1-t1[1,1]/sum(t1[,1]),t1[2,2]/sum(t1[,2]))
+ }
> x11()
```

```
> plot(mat[,1],mat[,2],type="l", xlim=c(0,1),ylim=c(0,1),xlab="1-Specificity",
+      ylab="Sensibility",main="ROC curve", col="orangered")
> abline(0,1)
> abline(1,-1)
>
> identify(mat[,1],mat[,2],n=1)
[1] 223
> 1- mat[223,1]
[1] 0.9146919
> mat[223,2]
[1] 0.9159664
> pc[223]
[1] 0.5214698
```

1. Ecrire les équations et hypothèses définissant le modèle ajusté après avoir introduit les variables aléatoires nécessaires à la formalisation du problème.
2. Donner l'expression de la variance conditionnelle de la réponse d'un individu de l'échantillon.
3. Ecrire la log-vraisemblance des données ainsi que la déviance.
4. Comment sont estimés les paramètres ? Donner la valeur des estimations.
5. Citer trois méthodes de construction d'intervalle de confiance bilatéral pour les paramètres du modèle.
6. Quelle est la taille de l'échantillon ? Ecrire les hypothèses testées, la statistique de test employée et sa loi asymptotique.
7. Quelles conclusions pouvez-vous tirer des éléments présentés ?
8. Grâce à la courbe ROC, on choisit le seuil 0.5214698 pour prédire la rémission à 5 ans. Quels sont alors le taux de faux négatifs et le taux de faux positifs de la méthode ?
9. Un nouveau patient (que l'on numérotera par `new`) se présente. Les valeurs de ses prédicteurs sont les suivantes :
   — `age` : 51
   — `aps` : 11
   — `stade` : 3
   — `gleason` : 6
   — `trt` : 4
   Estimer sa probabilité de rémission à 5 ans.
10. Comment effectuer la prédiction de $\widehat{Y}_{\texttt{new}}$?
11. Estimer la variance conditionnelle de la prédiction de la réponse de cet individu.

Exercice 23.

1. Détailler les scénarios utiles à une étude par simulations de Monte-Carlo de l'impact du nombre de prédicteurs dans un modèle de régression de Poisson.
2. Détailler les scénarios utiles à une étude par simulations de Monte-Carlo de la sensibilité au choix de la fonction de lien dans un modèle de régression pour variables réponses binaires.

Exercice 24.
Ci-dessous est présentée une étude par simulations de Monte-Carlo au moyen du logiciel R.

1. Ecrire les équations et hypothèses définissant le(s) modèle(s) simulé(s) et ajusté(s) après avoir introduit les variables aléatoires nécessaires à la formalisation du problème.
2. Ecrire les hypothèses testées.
3. Interpréter les résultats obtenus.
4. Quelle(s) conclusion(s) en tirer ?

```
> library(lmtest)
>
> M<-1000
> n<-200
> p<-3
> sigma2<- 1
> sigma2_X<- 2
> beta<- rep(1,p+1)
>
> X<-matrix(NA,n,p+1)
> X[,1]<-rep(1,n)
> X[,2]<- seq(from=1,to=11,length.out=n)
> X[,3]<- c(rep(1,floor(n/5)),seq(from=1,to=11,length.out=n-floor(n/5)))
> X[,4]<- c(seq(from=1,to=11,length.out=n-floor(n/5)),rep(11,floor(n/5)))
>
> beta_hat_m1<- matrix(NA,M,p+1)
> sd_hat_beta_hat_m1<- matrix(NA,M,p+1)
> pvalue_beta_hat_m1<- matrix(NA,M,p+1)
> pvalue_bptest_m1<- rep(NA,M)
> pvalue_Ftest_m1<- rep(NA,M)
> pvalue_Shapiro_res_test_m1<- rep(NA,M)
>
> beta_hat_m2<- matrix(NA,M,p+2)
> sd_hat_beta_hat_m2<- matrix(NA,M,p+2)
> pvalue_beta_hat_m2<- matrix(NA,M,p+2)
> pvalue_bptest_m2<- rep(NA,M)
> pvalue_Ftest_m2<- rep(NA,M)
> pvalue_Shapiro_res_test_m2<- rep(NA,M)
>
> beta_hat_m3<- matrix(NA,M,p)
> sd_hat_beta_hat_m3<- matrix(NA,M,p)
> pvalue_beta_hat_m3<- matrix(NA,M,p)
> pvalue_bptest_m3<- rep(NA,M)
> pvalue_Ftest_m3<- rep(NA,M)
> pvalue_Shapiro_res_test_m3<- rep(NA,M)
>
>
> for (m in 1:M)
+ {
```

```
+     Xsup<- rnorm(n,mean=7,sd=sqrt(sigma2_X))
+     eps<-rnorm(n,mean=0,sd=sqrt(sigma2))
+     y<-X%*%beta+eps
+
+     myfit_m1<- lm(y~X[,2]+X[,3]+X[,4])
+     myout_m1<- summary(myfit_m1)
+     beta_hat_m1[m,]<- myout_m1$coeff[,1]
+     sd_hat_beta_hat_m1[m,]<- myout_m1$coeff[,2]
+     pvalue_beta_hat_m1[m,]<-myout_m1$coeff[,4]
+     pvalue_bptest_m1[m]<- bptest(myout_m1)$p.value
+     pvalue_Ftest_m1[m]<-  anova(myfit_m1,lm(y~1))$Pr[2]
+     pvalue_Shapiro_res_test_m1[m]<- shapiro.test(rstudent(myfit_m1))$p.value
+
+     myfit_m2<- lm(y~X[,2]+X[,3]+X[,4]+Xsup)
+     myout_m2<- summary(myfit_m2)
+     beta_hat_m2[m,]<- myout_m2$coeff[,1]
+     sd_hat_beta_hat_m2[m,]<- myout_m2$coeff[,2]
+     pvalue_beta_hat_m2[m,]<-myout_m2$coeff[,4]
+     pvalue_bptest_m2[m]<- bptest(myout_m2)$p.value
+     pvalue_Ftest_m2[m]<-  anova(myfit_m2,lm(y~1))$Pr[2]
+     pvalue_Shapiro_res_test_m2[m]<- shapiro.test(rstudent(myfit_m2))$p.value
+
+     myfit_m3<- lm(y~X[,2]+X[,3])
+     myout_m3<- summary(myfit_m3)
+     beta_hat_m3[m,]<- myout_m3$coeff[,1]
+     sd_hat_beta_hat_m3[m,]<- myout_m3$coeff[,2]
+     pvalue_beta_hat_m3[m,]<-myout_m3$coeff[,4]
+     pvalue_bptest_m3[m]<- bptest(myout_m3)$p.value
+     pvalue_Ftest_m3[m]<-  anova(myfit_m3,lm(y~1))$Pr[2]
+     pvalue_Shapiro_res_test_m3[m]<- shapiro.test(rstudent(myfit_m3))$p.value
+ }
>
>
> apply(beta_hat_m1,2,mean)
[1] 0.9882808 1.0008971 0.9939572 1.0051894
> apply(beta_hat_m1,2,sd)
[1] 0.3083708 0.3006661 0.1590674 0.1617416
> apply(sd_hat_beta_hat_m1,2,mean)
[1] 0.3093331 0.3010303 0.1619161 0.1619161
>
> for (k in 1:(p+1))
+ { print(mean(ifelse(pvalue_beta_hat_m1[,k]<=0.05,1,0)))
+ }
[1] 0.885
[1] 0.907
[1] 1
[1] 1
> mean(ifelse(pvalue_Ftest_m1<=0.05,1,0))
```

```
[1] 1
> mean(ifelse(pvalue_bptest_m1<=0.05,1,0))
[1] 0.047
> mean(ifelse(pvalue_Shapiro_res_test_m1<=0.05,1,0))
[1] 0.064
>
>
> apply(beta_hat_m2,2,mean)
[1] 0.965452227 1.003465847 0.992333493 1.004562055 0.002842125
> apply(beta_hat_m2,2,sd)
[1] 0.5030776 0.3033114 0.1609565 0.1620920 0.0509500
> apply(sd_hat_beta_hat_m2,2,mean)
[1] 0.51517427 0.30458865 0.16455134 0.16231687 0.05128845
>
> for (k in 1:(p+2))
+ { print(mean(ifelse(pvalue_beta_hat_m2[,k]<=0.05,1,0)))
+ }
[1] 0.463
[1] 0.898
[1] 1
[1] 1
[1] 0.05
> mean(ifelse(pvalue_Ftest_m2<=0.05,1,0))
[1] 1
> mean(ifelse(pvalue_bptest_m2<=0.05,1,0))
[1] 0.044
> mean(ifelse(pvalue_Shapiro_res_test_m2<=0.05,1,0))
[1] 0.062
>
>
> apply(beta_hat_m3,2,mean)
[1] 0.6387101 2.5439938 0.6194204
> apply(beta_hat_m3,2,sd)
[1] 0.3018648 0.1669849 0.1467632
> apply(sd_hat_beta_hat_m3,2,mean)
[1] 0.3325116 0.1856373 0.1642598
>
> for (k in 1:p)
+ { print(mean(ifelse(pvalue_beta_hat_m3[,k]<=0.05,1,0)))
+ }
[1] 0.477
[1] 1
[1] 0.977
> mean(ifelse(pvalue_Ftest_m3<=0.05,1,0))
[1] 1
> mean(ifelse(pvalue_bptest_m3<=0.05,1,0))
[1] 0.32
> mean(ifelse(pvalue_Shapiro_res_test_m3<=0.05,1,0))
```

```
[1] 0.064
```

Exercice 25.
Ci-dessous est présentée une étude par simulations de Monte-Carlo.

1. Ecrire les équations et hypothèses définissant le modèle étudié ci-dessous avec le logiciel R après avoir introduit les variables aléatoires nécessaires à la formalisation du problème.
2. Ecrire les hypothèses testées, les statistiques de test employées et leur loi asymptotique.
3. Interpréter les résultats obtenus.
4. Quelle(s) conclusion(s) en tirer ?
5. Présenter les alternatives au modèle de Poisson en expliquant ce qui est alors spécifiquement modélisé.

```
> library(moments)
> library(AER)
>
> M<-1000
> n<-200
> p<-3
> beta<- c(0,1,1,1)
> sigma2_X<-2
>
> beta_hat<- matrix(NA,M,p+1)
> sd_hat_beta_hat<- matrix(NA,M,p+1)
> pvalue_beta_hat<- matrix(NA,M,p+1)
> mean_rpstud<-rep(NA,M)
> skew_rpstud<- rep(NA,M)
> dispers_rpstud<-  matrix(NA,M,2)
> pvalue_Ptest<- rep(NA,M)
> pvalue_dtest1<- rep(NA,M)
> pvalue_dtest2<- rep(NA,M)
>
> for (m in 1:M)
+   {
+    X<-matrix(0,n,p+1)
+    X[,1] <- rep(1,n)
+    if (p>=1)  { X[,2] <- runif(n,0,4) }
+    if (p>=2)
+     {
+       for (k in 3:(p+1)) { X[,k] <- rnorm(n,mean=0,sd=sqrt(sigma2_X)) }
+       }
+    y<-rpois(n,lambda=exp(X%*%beta))
+    myfit <- glm(y~X[,2]+X[,3]+X[,4],family=poisson)
+    beta_hat[m,] <- summary(myfit)$coeff[,1]
+    sd_hat_beta_hat[m,] <- summary(myfit)$coeff[,2]
+    pvalue_beta_hat[m,] <- summary(myfit)$coeff[,4]
+    rp <- residuals(myfit,type="pearson")
```

```
+    rps <- residuals(myfit,type="pearson")/sqrt(1-hatvalues(myfit))
+    mean_rpstud[m] <- mean(rps)
+    skew_rpstud[m] <- skewness(rps)
+    dispers_rpstud[m,] <- quantile(rps,probs=c(0.025,0.975),names=F)
+    pvalue_Ptest[m] <- pchisq(sum(rp^2),df=length(y)-(p+1),lower.tail=F)
+    pvalue_dtest1[m]<- dispersiontest(myfit)$p.value
+    pvalue_dtest2[m]<- dispersiontest(myfit,trafo=2)$p.value
+   }
>
> apply(beta_hat,2,mean)
[1] 0.000122407 0.999895678 0.999826647 1.000174989
>
> summary(mean_rpstud)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-0.125438 -0.021638  0.001616  0.001125  0.024814  0.138481
> summary(dispers_rpstud)
       V1                V2
 Min.   :-2.161   Min.   :1.438
 1st Qu.:-1.730   1st Qu.:2.001
 Median :-1.641   Median :2.160
 Mean   :-1.644   Mean   :2.169
 3rd Qu.:-1.547   3rd Qu.:2.311
 Max.   :-1.263   Max.   :3.164
> summary(skew_rpstud)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-0.02508  0.45369  0.61346  0.67984  0.81057  3.96821
>
> for (k in 1:(p+1))
+  {
+   print(mean(ifelse(pvalue_beta_hat[,k]<=0.05,1,0)))
+   }
[1] 0.051
[1] 1
[1] 1
[1] 1
> mean(ifelse(pvalue_Ptest<=0.05,1,0))
[1] 0.097
> mean(ifelse(pvalue_dtest1<=0.05,1,0))
[1] 0.011
> mean(ifelse(pvalue_dtest2<=0.05,1,0))
[1] 0.004
>
>
> beta_hat<- matrix(NA,M,p)
> sd_hat_beta_hat<- matrix(NA,M,p)
> pvalue_beta_hat<- matrix(NA,M,p)
> mean_rpstud<-rep(NA,M)
> skew_rpstud<- rep(NA,M)
```

```
> dispers_rpstud<-  matrix(NA,M,2)
> pvalue_Ptest<- rep(NA,M)
> pvalue_dtest1<- rep(NA,M)
> pvalue_dtest2<- rep(NA,M)
>
> for (m in 1:M)
+ {
+     myfit <- glm(y~X[,2]+X[,3],family=poisson)
+     beta_hat[m,] <- summary(myfit)$coeff[,1]
+     sd_hat_beta_hat[m,] <- summary(myfit)$coeff[,2]
+     pvalue_beta_hat[m,] <- summary(myfit)$coeff[,4]
+     rp <- residuals(myfit,type="pearson")
+     rps <- residuals(myfit,type="pearson")/sqrt(1-hatvalues(myfit))
+     mean_rpstud[m] <- mean(rps)
+     skew_rpstud[m] <- skewness(rps)
+     dispers_rpstud[m,] <- quantile(rps,probs=c(0.025,0.975),names=F)
+     pvalue_Ptest[m] <- pchisq(sum(rp^2),df=length(y)-p,lower.tail=F)
+     pvalue_dtest1[m]<- dispersiontest(myfit)$p.value
+     pvalue_dtest2[m]<- dispersiontest(myfit,trafo=2)$p.value
+ }
>
> apply(beta_hat,2,mean)
[1] 0.9285931 1.0033414 0.9430411
>
> summary(mean_rpstud)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
 -1.3557  -0.3550  -0.1004   1.2198   0.3276 709.9236
> summary(dispers_rpstud)
       V1                V2
 Min.   :-73.674   Min.   :  11.02
 1st Qu.:-18.422   1st Qu.:  25.71
 Median :-14.877   Median :  31.50
 Mean   :-16.041   Mean   :  38.74
 3rd Qu.:-12.737   3rd Qu.:  38.96
 Max.   : -7.693   Max.   :2997.66
> summary(skew_rpstud)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.430   3.759   4.968   5.477   6.712  13.477
>
> for (k in 1:p)
+ {
+     print(mean(ifelse(pvalue_beta_hat[,k]<=0.05,1,0)))
+ }
[1] 0.984
[1] 1
[1] 1
> mean(ifelse(pvalue_Ptest<=0.05,1,0))
[1] 1
```

```
> mean(ifelse(pvalue_dtest1<=0.05,1,0))
[1] 0.812
> mean(ifelse(pvalue_dtest2<=0.05,1,0))
[1] 0.837
>
```

Exercice 26.
Ci-dessous est présentée une étude par simulations de Monte-Carlo.

1. Ecrire les équations et hypothèses définissant le modèle étudié ci-dessous avec le logiciel R après avoir introduit les variables aléatoires nécessaires à la formalisation du problème.
2. Ecrire les hypothèses testées, les statistiques de test employées (du moins celle(s) qui ont été présentées en cours) et leur loi asymptotique, le cas échéant. Pourquoi le test d'adéquation de Pearson n'est-il pas employé ?
3. Qu'illustrent ces simulations ? Quelle(s) conclusion(s) en tirer ?

```
> library(moments)
> library(lmtest)
>
> M<-1000
> n<-200
> p<-3
> beta<- c(6,rep(1,p))
> sigma2_X<-1
>
> beta_hat_fit1 <- matrix(NA,M,p+1)
> beta_hat_fit2 <- matrix(NA,M,p+1)
> beta_hat_fit3 <- matrix(NA,M,p+1)
>
> sd_hat_beta_hat_fit1 <- matrix(NA,M,p+1)
> sd_hat_beta_hat_fit2 <- matrix(NA,M,p+1)
> sd_hat_beta_hat_fit3 <- matrix(NA,M,p+1)
>
> pvalue_beta_hat_fit1 <- matrix(NA,M,p+1)
> pvalue_beta_hat_fit2 <- matrix(NA,M,p+1)
> pvalue_beta_hat_fit3 <- matrix(NA,M,p+1)
>
> mean_rpstud1 <-rep(NA,M)
> mean_rpstud2 <-rep(NA,M)
> mean_rpstud3 <-rep(NA,M)
>
> skew_rpstud1 <- rep(NA,M)
> skew_rpstud2 <- rep(NA,M)
> skew_rpstud3 <- rep(NA,M)
>
> dispers_rpstud1 <-  matrix(NA,M,5)
> dispers_rpstud2 <-  matrix(NA,M,5)
```

```
> dispers_rpstud3 <-  matrix(NA,M,5)
>
> pvalue_stest1<- rep(NA,M)
> pvalue_stest2<- rep(NA,M)
> pvalue_stest3<- rep(NA,M)
>
> pvalue_bptest1 <- rep(NA,M)
> pvalue_bptest2 <- rep(NA,M)
>
> AIC_fit <- matrix(NA,M,3)
>
>
> for (m in 1:M)
+ {
+   X<-matrix(0,n,p+1)
+   X[,1] <- rep(1,n)
+   if (p>=1)  { X[,2] <- runif(n,-3,3) }
+   if (p>=2)
+   {
+     for (k in 3:(p+1)) { X[,k] <- rnorm(n,mean=0,sd=sqrt(sigma2_X)) }
+   }
+
+   y<-rgamma(n,shape=exp(1),scale=exp(X%*%beta))
+
+   myfit1 <- lm(y~X[,2]+X[,3]+X[,4])
+   myfit2 <- lm(log(y)~X[,2]+X[,3]+X[,4])
+   myfit3 <- glm(y~X[,2]+X[,3]+X[,4],family=Gamma(link="log"))
+
+   beta_hat_fit1[m,] <- summary(myfit1)$coeff[,1]
+   beta_hat_fit2[m,] <- summary(myfit2)$coeff[,1]
+   beta_hat_fit3[m,] <- summary(myfit3)$coeff[,1]
+
+   sd_hat_beta_hat_fit1[m,] <- summary(myfit1)$coeff[,2]
+   sd_hat_beta_hat_fit2[m,] <- summary(myfit2)$coeff[,2]
+   sd_hat_beta_hat_fit3[m,] <- summary(myfit3)$coeff[,2]
+
+   pvalue_beta_hat_fit1[m,] <- summary(myfit1)$coeff[,4]
+   pvalue_beta_hat_fit2[m,] <- summary(myfit2)$coeff[,4]
+   pvalue_beta_hat_fit3[m,] <- summary(myfit3)$coeff[,4]
+
+   rp1 <- residuals(myfit1,type="pearson")
+   rps1 <- residuals(myfit1,type="pearson")/sqrt((sum(rp1^2)/(length(y)-(p+1)))*(1-hatval
+   rp2 <- residuals(myfit2,type="pearson")
+   rps2 <- residuals(myfit2,type="pearson")/sqrt((sum(rp2^2)/(length(y)-(p+1)))*(1-hatval
+   rp3 <- residuals(myfit3,type="pearson")
+   rps3 <- residuals(myfit3,type="pearson")/sqrt((sum(rp3^2)/(length(y)-(p+1)))*(1-hatval
+
+   mean_rpstud1[m] <- mean(rps1)
```

```
+   mean_rpstud2[m] <- mean(rps2)
+   mean_rpstud3[m] <- mean(rps3)
+
+   skew_rpstud1[m] <- skewness(rps1)
+   skew_rpstud2[m] <- skewness(rps2)
+   skew_rpstud3[m] <- skewness(rps3)
+
+   dispers_rpstud1[m,] <- quantile(rps1,probs=c(0,0.025,0.5,0.975,1),names=F)
+   dispers_rpstud2[m,] <- quantile(rps2,probs=c(0,0.025,0.5,0.975,1),names=F)
+   dispers_rpstud3[m,] <- quantile(rps3,probs=c(0,0.025,0.5,0.975,1),names=F)
+
+   pvalue_stest1[m]<- shapiro.test(rps1)$p.value
+   pvalue_stest2[m]<- shapiro.test(rps2)$p.value
+   pvalue_stest3[m]<- shapiro.test(rps3)$p.value
+
+   pvalue_bptest1[m] <- bptest(myfit1)$p.value
+   pvalue_bptest2[m] <- bptest(myfit2)$p.value
+
+   AIC_fit[m,1] <- AIC(myfit1)
+   AIC_fit[m,2] <- AIC(myfit2)
+   AIC_fit[m,3] <- AIC(myfit3)
+ }
>
> apply(beta_hat_fit1,2,mean)
[1] 9670.715 6527.400 9492.644 9685.919
> apply(beta_hat_fit2,2,mean)
[1] 6.8063786 0.9991937 0.9988683 1.0004768
> apply(beta_hat_fit3,2,mean)
[1] 6.9971113 0.9988255 1.0007760 1.0003221
>
> summary(mean_rpstud1)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
0.0007075 0.0014025 0.0016739 0.0017182 0.0019471 0.0035018
> summary(mean_rpstud2)
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
-1.714e-03 -2.830e-04 -1.803e-05 -9.996e-06  2.586e-04  1.225e-03
> summary(mean_rpstud3)
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
-1.490e-03 -3.026e-04 -3.756e-05 -2.935e-05  2.185e-04  1.303e-03
>
> summary(skew_rpstud1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.232   4.727   6.275   6.734   8.279  12.837
> summary(skew_rpstud2)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-2.01381 -0.74231 -0.58411 -0.60552 -0.44425  0.05299
> summary(skew_rpstud3)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
 0.5166  0.9371  1.0911  1.1328  1.2958  2.6415
>
> summary(dispers_rpstud1)
       V1                V2                V3                 V4               V5
 Min.   :-1.9969   Min.   :-1.1855   Min.   :-0.29860   Min.   :0.4339   Min.   : 4.863
 1st Qu.:-1.1761   1st Qu.:-0.9005   1st Qu.:-0.20675   1st Qu.:1.0265   1st Qu.: 8.101
 Median :-1.0293   Median :-0.8075   Median :-0.17842   Median :1.4304   Median : 9.833
 Mean   :-1.0444   Mean   :-0.8062   Mean   :-0.17764   Mean   :1.5730   Mean   : 9.883
 3rd Qu.:-0.8954   3rd Qu.:-0.7139   3rd Qu.:-0.15022   3rd Qu.:2.0469   3rd Qu.:11.597
 Max.   :-0.5489   Max.   :-0.4271   Max.   :-0.03722   Max.   :3.9233   Max.   :13.988
> summary(dispers_rpstud2)
       V1               V2               V3                  V4              V5
 Min.   :-6.591   Min.   :-2.929   Min.   :-0.09275   Min.   :1.288   Min.   :1.649
 1st Qu.:-3.904   1st Qu.:-2.301   1st Qu.: 0.06123   1st Qu.:1.576   1st Qu.:2.008
 Median :-3.443   Median :-2.167   Median : 0.09767   Median :1.650   Median :2.161
 Mean   :-3.532   Mean   :-2.182   Mean   : 0.09588   Mean   :1.652   Mean   :2.182
 3rd Qu.:-3.083   3rd Qu.:-2.052   3rd Qu.: 0.13033   3rd Qu.:1.724   3rd Qu.:2.328
 Max.   :-2.223   Max.   :-1.558   Max.   : 0.23142   Max.   :2.037   Max.   :3.081
> summary(dispers_rpstud3)
       V1               V2               V3                  V4              V5
 Min.   :-1.919   Min.   :-1.670   Min.   :-0.33592   Min.   :1.862   Min.   :2.702
 1st Qu.:-1.598   1st Qu.:-1.402   1st Qu.:-0.22643   1st Qu.:2.250   1st Qu.:3.528
 Median :-1.531   Median :-1.347   Median :-0.19453   Median :2.376   Median :3.928
 Mean   :-1.533   Mean   :-1.347   Mean   :-0.19392   Mean   :2.383   Mean   :4.038
 3rd Qu.:-1.463   3rd Qu.:-1.287   3rd Qu.:-0.16377   3rd Qu.:2.506   3rd Qu.:4.442
 Max.   :-1.235   Max.   :-1.107   Max.   :-0.04017   Max.   :3.076   Max.   :7.460
>
> for (k in 1:(p+1))
+ {
+     print(paste("k=",k))
+     print(paste("results for fit 1:",mean(ifelse(pvalue_beta_hat_fit1[,k]<=0.05,1,0))))
+     print(paste("results for fit 2:",mean(ifelse(pvalue_beta_hat_fit2[,k]<=0.05,1,0))))
+     print(paste("results for fit 3:",mean(ifelse(pvalue_beta_hat_fit3[,k]<=0.05,1,0))))
+ }
[1] "k= 1"
[1] "results for fit 1: 0.979"
[1] "results for fit 2: 1"
[1] "results for fit 3: 1"
[1] "k= 2"
[1] "results for fit 1: 0.997"
[1] "results for fit 2: 1"
[1] "results for fit 3: 1"
[1] "k= 3"
[1] "results for fit 1: 0.982"
[1] "results for fit 2: 1"
[1] "results for fit 3: 1"
[1] "k= 4"
[1] "results for fit 1: 0.986"
```

```
[1] "results for fit 2: 1"
[1] "results for fit 3: 1"
>
> mean(ifelse(pvalue_stest1<=0.05,1,0))
[1] 1
> mean(ifelse(pvalue_stest2<=0.05,1,0))
[1] 0.852
> mean(ifelse(pvalue_stest3<=0.05,1,0))
[1] 1
>
> mean(ifelse(pvalue_bptest1<=0.05,1,0))
[1] 0.971
> mean(ifelse(pvalue_bptest2<=0.05,1,0))
[1] 0.054
>
> table(apply(AIC_fit,1,which.min))
   2
1000
```

Indication : dans le logiciel `R`, la loi gamma est paramétrée au moyen de 2 paramètres nommés *shape* et *scale* de sorte qu'une variable aléatoire Y tirée selon cette loi a pour espérance :

$$\mathbb{E}[Y] = \text{shape} \times \text{scale}$$

et pour variance :

$$\text{Var}[Y] = \text{shape} \times \text{scale}^2.$$

Si l'on note *shape* par a et *scale* par s, alors la densité de Y en un point $y > 0$ s'écrit :

$$f(y) = \frac{1}{s^a \Gamma(a)} y^{a-1} e^{-y/s}.$$

Exercice 27.

Le jeu de données `heart.data` est disponible dans le package `glmpath` du logiciel `R`. Ce jeu de données a été présenté dans le manuel de Hastie T., Tibshirani R. et Friedman J. (2001) intitulé *Elements of Statistical Learning ; Data Mining, Inference, and Prediction*. Les données sont issues d'une étude portant sur le risque de développer une coronaropathie chez des hommes blancs en Afrique du Sud (qui est une population particulièrement touchée par cette pathologie). La variable réponse `y` est binaire et indique la présence (codée par 1) ou l'absence de coronaropathie (codée par 0). Sont également disponibles 9 prédicteurs candidats qui sont les suivants :

- `sbp` : pression sanguine systolique
- `tobacco` : prise cumulée de tabac exprimée en kg
- `ldl` : taux sanguin de lipoprotéines de basse densité (appelé aussi mauvais cholestérol)
- `famhist` : indique la présence ou non d'antécédents familiaux
- `age` : âge de l'individu
- `alcohool` : consommation actuelle d'alcool

— `obesity` : mesure de l'obésité basée sur l'IMC (poids en kg rapporté au carré de la taille en m d'un individu)
— `adiposity` : mesure du taux de masse grasse basée sur le tour de hanche rapporté à la taille d'un individu
— `typea` : mesure du stress psycho-social

Un graphique exploratoire est présenté en figures 7 et 8 tandis qu'un début d'analyse est reproduit ci-dessous.

1. Quelles conclusions pouvez-vous tirer des graphiques exploratoires ?
2. Ecrire les équations et hypothèses correspondant au modèle ajusté.
3. Ecrire la log-vraisemblance des données ainsi que la déviance.
4. Comment sont estimés les paramètres ? Donner la valeur des estimations.
5. Citer trois méthodes de construction d'intervalle de confiance bilatéral pour les paramètres du modèle.
6. Comment tester la nécessité de l'inclusion des différentes covariables introduits ? Mettre en oeuvre.
7. Comment évaluer la qualité de l'ajustement du modèle aux données ?
8. Que pensez-vous des premiers résultats obtenus ? Sont-ils en accord avec vos conclusions issues de l'étape exploratoire ? D'où peuvent venir les problèmes ?
9. Quelle est la variance de la réponse d'un individu de l'échantillon, conditionnellement aux prédicteurs. Comment l'estimer ?
10. Estimer la probabilité de développer une coronaropathie pour un homme de 50 ans, ayant une consommation d'alcool à 10, une mesure de `typea` à 60, une consommation cumulée de tabac à 6, une pression artérielle systolique à 140, des antécédents familiaux, un taux de `ldl` à 5, un taux de masse grasse évalué par `adiposity`=7 et un IMC évalué par `obesity`= 18.
11. Comment est la prédiction par le modèle de la réponse de cet individu ?
12. Comment évaluer la qualité prédictive du modèle ?

```
> my_fit1<-glm(y~sbp+tobacco+ldl+adiposity +famhist +typea +obesity +alcohol
  +age,family=binomial)
> summary(my_fit1)

Call:
glm(formula = y ~ sbp + tobacco + ldl + adiposity + famhist +
    typea + obesity + alcohol + age, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7781  -0.8213  -0.4387   0.8889   2.5435

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.1507209  1.3082600  -4.701 2.58e-06 ***
sbp          0.0065040  0.0057304   1.135 0.256374
tobacco      0.0793764  0.0266028   2.984 0.002847 **
```

```
ldl          0.1739239  0.0596617   2.915 0.003555 **
adiposity    0.0185866  0.0292894   0.635 0.525700
famhist      0.9253704  0.2278940   4.061 4.90e-05 ***
typea        0.0395950  0.0123202   3.214 0.001310 **
obesity     -0.0629099  0.0442477  -1.422 0.155095
alcohol      0.0001217  0.0044832   0.027 0.978350
age          0.0452253  0.0121298   3.728 0.000193 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 596.11  on 461  degrees of freedom
Residual deviance: 472.14  on 452  degrees of freedom
AIC: 492.14

Number of Fisher Scoring iterations: 5
```

Exercice 28.

1. Une étude est réalisée sur le lien entre l'exposition à un composé organochloré et l'apparition de cancer chez le médaka (poisson de 2 à 4 cm de long, originaire des rizières des régions côtières d'Asie du Sud). On dispose de 70 poissons (sans lien de parenté) équirépartis dans 7 aquariums. Dans chaque aquarium, on répand une certaine dose du composé organochloré. Les poissons sont euthanasiés au bout de 6 mois d'exposition. Une analyse histopathologique de leurs tissus est effectuée afin de détecter la présence ou non de cellules cancéreuses dans l'organisme. On relève également le dosage sanguin d'un marqueur noté m indiquant le niveau de fonctionnement du système immunitaire. Proposer un modèle adapté à cette étude (on ne prendra pas en compte d'éventuelle source de corrélation dû au confinement dans un même aquarium). Quels diagnostics effectuer pour évaluer la qualité de l'ajustement du modèle proposé à un tel jeu de données ? Comment tester le lien entre l'exposition à un composé organochloré et l'apparition de cancer ? Mettre en oeuvre. On note $ED_{50}(m_0)$ la dose de composé organochloré associée à une probabilité d'apparition de cancer de 50% pour un niveau m_0 du marqueur immunitaire. Comment estimer $ED_{50}(m_0)$ lorsqu'on a conclu à l'existence d'un lien entre cancer et exposition au composé organochloré ?

Exercice 29.

1. Ecrire l'équation et les hypothèses définissant un modèle de régression pour la variable réponse Y représentant le taux de chômage dans une agglomération donnée (Londres, Manchester, Bristol, Birmingham, Glasgow, Barcelone, Madrid, Séville, Saragosse, Paris, Lyon, Marseille, Lille, Montpellier, Bordeaux, Nice, Strasbourg, Genève, Lausanne, Berne, Neuchâtel, Berlin, Munich, Bonn, Karlsruhe, Cologne, Francfort, Stuttgart, Düsseldorf, Hambourg, Brême) en fonction du volume des exportations (en euros), de la taille de l'agglomération, SMIC horaire (en euros), formation des demandeurs d'emploi (pourcentage de personnes sans diplôme, avec CAP/BEP, avec Baccalauréat, avec Bac+3, avec Bac+5 et plus).

2. Une étude est réalisée chez la souris sur le lien entre l'exposition à un pesticide pendant la gestation et l'issue de la gestation. Pour cela, 12 femelles (sans lien de parenté) sont exposées à ce pesticide à différentes doses pendant toute la durée de leur gestation. A l'issue de la gestation, peu de temps avant la délivrance, on compte le nombre de souriceaux morts, le nombre de souriceaux viables mais malformés et le nombre de souriceaux vivants sans malformation. Montrer que la loi multinômiale appartient à la famille exponentielle multivariée. Comment pourrait-on utiliser la loi multinômiale ici ?
   On note $\pi_1(d)$ la probabilité pour un souriceau d'être mort à l'issue de la gestation lorsque sa mère a reçu la dose d, $\pi_2(d)$ la probabilité pour un souriceau d'être viable mais mal formé à l'issue de la gestation lorsque sa mère a reçu la dose d, et $\pi_3(d)$ la probabilité pour un souriceau d'être viable sans malformation à l'issue de la gestation lorsque sa mère a reçu la dose d. On propose maintenant de spécifier ces trois probabilités comme suit :

$$\begin{aligned}\pi_1(d) &= 1 - \exp\left(-(\beta_0 + \beta_1 d)^{\alpha_1}\right) ,\\ \pi_2(d) &= 1 - \exp\left(-(\beta_2 + \beta_3 d)^{\alpha_2}\right) ,\\ \pi_3(d) &= 1 - \pi_1(d) - \pi_2(d) .\end{aligned}$$

   Le modèle de régression qui en résulte est-il un GLM si α_1 et α_2 sont tous deux inconnus ? tous deux connus ?

Exercice 30.

Une étude est réalisée sur le lien entre l'exposition à un pesticide et la présence de troubles endocriniens. Pour cela, on recrute 85 exploitants agricoles (sans lien de parenté) n'ayant jamais utilisé ce pesticide ou bien ayant utilisant ce pesticide depuis plus de 10 ans (variable `pest`, 1 = non utilisation de pesticide, 2 = utilisation de pesticide depuis plus de 10 ans). On effectue chez chaque exploitant un dosage d'une hormone thyroïdienne (variable `y`), un taux élevé étant révélateur d'hypothyroïdie. On recueille également les informations suivantes pour chaque exploitant : âge (variable `age`), sexe (variable `sex`, 1 = homme, 2 = femme) et prise de médicaments (variable `med`, 1 = non, 2 = prise de médicaments au lithium, 3 = prise de médicaments à l'iode, 4 = prise de médicaments à l'iode et au lithium). Un tracé exploratoire est représenté en figure 9. Puis, l'analyse suivante est effectuée au moyen du logiciel `R`.

```
> mylm<-lm(y~age+factor(sex)+factor(pest)+factor(med)+factor(pest)*factor(med))
> summary(mylm)

Call:
lm(formula = y ~ age + factor(sex) + factor(pest) + factor(med) +
    factor(pest) * factor(med))

Residuals:
     Min       1Q   Median       3Q      Max
-0.54479 -0.15644 -0.04940  0.08538  1.07164

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                0.479816   0.263141   1.823  0.07222 .
age                       -0.003843   0.004401  -0.873  0.38530
```

```
factor(sex)2                -0.125442   0.072877  -1.721  0.08932 .
factor(pest)2                0.084216   0.134716   0.625  0.53378
factor(med)2                -0.118718   0.135176  -0.878  0.38261
factor(med)3                -0.084797   0.131043  -0.647  0.51955
factor(med)4                -0.114807   0.127836  -0.898  0.37202
factor(pest)2:factor(med)2   0.683122   0.215312   3.173  0.00219 **
factor(pest)2:factor(med)3   0.421803   0.193414   2.181  0.03233 *
factor(pest)2:factor(med)4   0.163323   0.189965   0.860  0.39266
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3121 on 75 degrees of freedom
Multiple R-squared:  0.3847,    Adjusted R-squared:  0.3109
F-statistic:  5.21 on 9 and 75 DF,  p-value: 1.719e-05
>
> res1<-rstudent(mylm)
>
> x11()
> par(mfrow=c(2,2))
> plot(med,res1)
> abline(h=c(0,-2,2))
> plot(age,res1)
> abline(h=c(0,-2,2))
> plot(pest,res1)
> abline(h=c(0,-2,2))
> plot(sex,res1)
> abline(h=c(0,-2,2))
> x11()
> qqnorm(res1)
> qqline(res1)
> shapiro.test(res1)

        Shapiro-Wilk normality test

data:  res1
W = 0.7854, p-value = 8.513e-10

> library(lmtest)
> bptest(mylm)

        studentized Breusch-Pagan test

data:  mylm
BP = 9.9095, df = 1, p-value = 0.001644
>
```

Les graphiques effectués ci-dessus sont présentés en figures 10 et 11. L'analyse se poursuit ci-dessous.

```
> myglm<-glm(y~age+factor(sex)+factor(pest)+factor(med)+factor(pest)*factor(med),
+ family=Gamma(link = "log"))
> summary(myglm)

Call:
glm(formula = y ~ age + factor(sex) + factor(pest) + factor(med) +
    factor(pest) * factor(med), family = Gamma(link = "log"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8302  -0.9345  -0.4359   0.3193   2.0101

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                -2.090961   0.856099  -2.442   0.0169 *
age                         0.009071   0.014318   0.634   0.5283
factor(sex)2               -0.405240   0.237098  -1.709   0.0916 .
factor(pest)2               0.694325   0.438282   1.584   0.1174
factor(med)2               -0.694740   0.439779  -1.580   0.1184
factor(med)3               -0.255490   0.426334  -0.599   0.5508
factor(med)4               -0.549625   0.415901  -1.322   0.1903
factor(pest)2:factor(med)2  1.587565   0.700494   2.266   0.0263 *
factor(pest)2:factor(med)3  0.847847   0.629252   1.347   0.1819
factor(pest)2:factor(med)4  0.399647   0.618028   0.647   0.5198
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 1.031266)

    Null deviance: 135.508  on 84  degrees of freedom
Residual deviance:  83.909  on 75  degrees of freedom
AIC: -81.679

Number of Fisher Scoring iterations: 9

> rps<-residuals(myglm,type="pearson")/sqrt(1-hatvalues(myglm))
> mean(rps)
[1] 0.0003389486
> library(moments)
> skewness(rps)
[1] 1.579206
>
> rds<-residuals(myglm,type="deviance")/sqrt(1-hatvalues(myglm))
> mean(rds)
[1] -0.3182297
> skewness(rds)
[1] 0.1209898
>
```

```
> x11()
> par(mfrow=c(2,2))
> plot(rps)
> abline(h=c(0,-2,2))
> plot(age,rps)
> abline(h=c(0,-2,2))
> plot(pest,rps)
> abline(h=c(0,-2,2))
> plot(sex,rps)
> abline(h=c(0,-2,2))
>
> x11()
> par(mfrow=c(2,2))
> plot(rds)
> abline(h=c(0,-2,2))
> plot(age,rds)
> abline(h=c(0,-2,2))
> plot(pest,rds)
> abline(h=c(0,-2,2))
> plot(sex,rds)
> abline(h=c(0,-2,2))
>
> drop1(myglm,test="Chisq")
Single term deletions

Model:
y ~ age + factor(sex) + factor(pest) + factor(med) + factor(pest) *
    factor(med)
                         Df Deviance     AIC scaled dev. Pr(>Chi)
<none>                        83.909 -81.679
age                       1   84.239 -83.359      0.3203  0.57142
factor(sex)               1   87.003 -80.679      3.0002  0.08326 .
factor(pest):factor(med)  3   89.924 -81.847      5.8326  0.12005
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> myglm<-glm(y~factor(sex)+factor(pest)+factor(med)+factor(pest)*factor(med),
+ family=Gamma(link = "log"))
> drop1(myglm,test="Chisq")
Single term deletions

Model:
y ~ factor(sex) + factor(pest) + factor(med) + factor(pest) *
    factor(med)
                         Df Deviance     AIC scaled dev. Pr(>Chi)
<none>                        84.239 -83.294
factor(sex)               1   87.015 -82.637      2.6577  0.10305
factor(pest):factor(med)  3   91.365 -82.471      6.8229  0.07776 .
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> myglm<-glm(y~factor(pest)+factor(med)+factor(pest)*factor(med),
+ family=Gamma(link = "log"))
> drop1(myglm,test="Chisq")
Single term deletions

Model:
y ~ factor(pest) + factor(med) + factor(pest) * factor(med)
                         Df Deviance     AIC scaled dev. Pr(>Chi)
<none>                        87.015 -82.110
factor(pest):factor(med)  3   93.977 -81.599      6.5116  0.08921 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> myglm<-glm(y~factor(pest)+factor(med),family=Gamma(link = "log"))
> drop1(myglm,test="Chisq")
Single term deletions

Model:
y ~ factor(pest) + factor(med)
             Df Deviance     AIC scaled dev.  Pr(>Chi)
<none>            93.977 -80.501
factor(pest)  1  129.422 -53.265     29.2363 6.407e-08 ***
factor(med)   3   97.246 -83.805      2.6962    0.4409
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> myglm<-glm(y~factor(pest),family=Gamma(link = "log"))
> drop1(myglm,test="Chisq")
Single term deletions

Model:
y ~ factor(pest)
             Df Deviance     AIC scaled dev. Pr(>Chi)
<none>            97.246 -83.098
factor(pest)  1  135.508 -52.979      32.119 1.45e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> summary(myglm)

Call:
glm(formula = y ~ factor(pest), family = Gamma(link = "log"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
```

```
-2.9144  -1.0697  -0.4641   0.2034   2.3146

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    -2.1419     0.1592 -13.454  < 2e-16 ***
factor(pest)2   1.3678     0.2381   5.744 1.48e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 1.191261)

    Null deviance: 135.508  on 84  degrees of freedom
Residual deviance:  97.246  on 83  degrees of freedom
AIC: -83.098

Number of Fisher Scoring iterations: 6
```

Les graphiques effectués ci-dessus sont présentés en figures 12 et 13.

1. Ecrire les équations et hypothèses définissant les modèles ayant été ajustés après avoir introduit les variables aléatoires nécessaires à la formalisation du problème.
2. Donner les estimations réalisées. Préciser les hypothèses nulles des tests effectués. Interpréter les graphiques et les résultats produits.
3. Détailler le scénario d'une étude par simulations du comportement des différents résidus utilisés dans l'analyse précédente lorsqu'on a ajusté un modèle de régression Gamma.

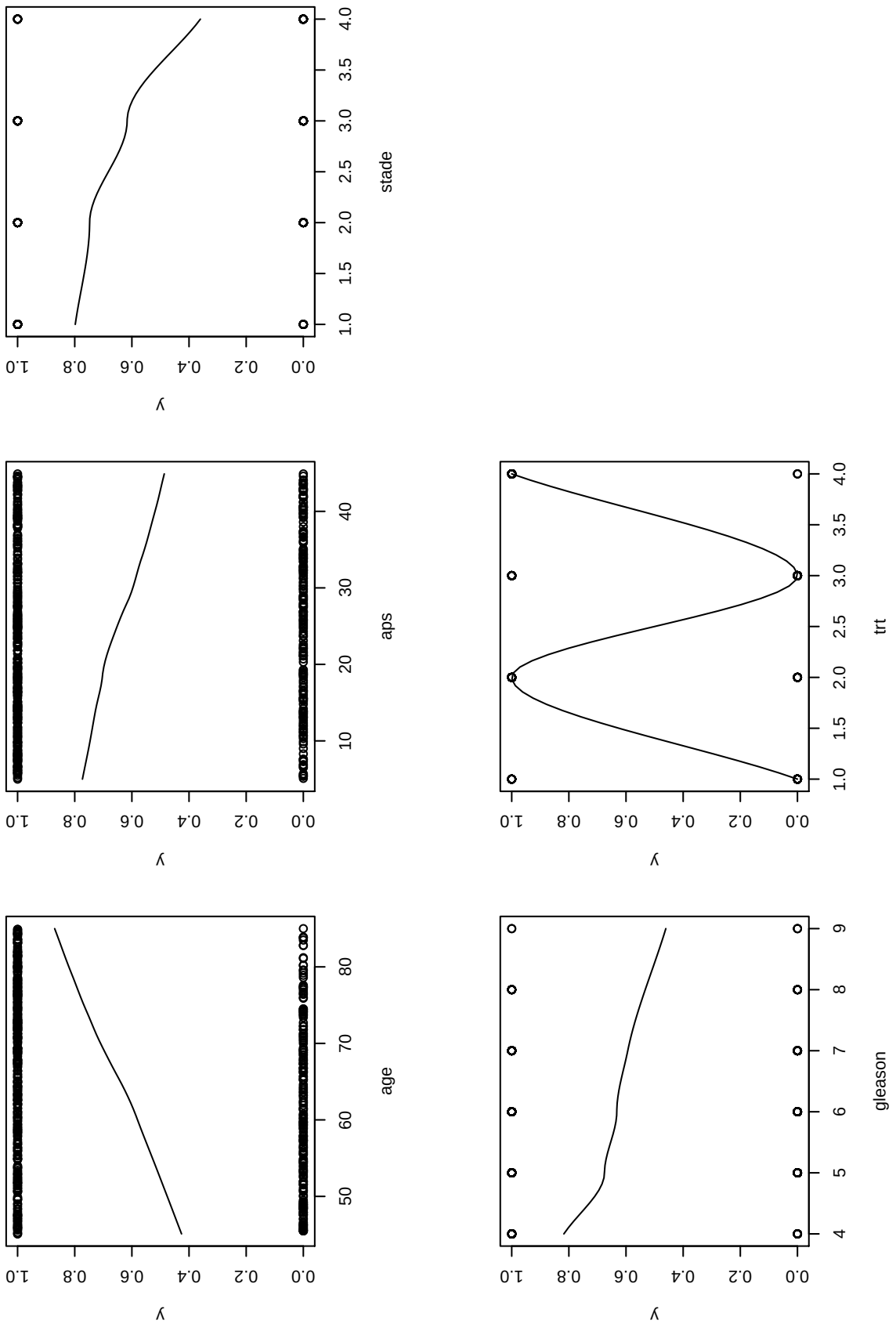


FIGURE 4 – Tracé relatif aux données sur le cancer de la prostate de l'exercice 22.

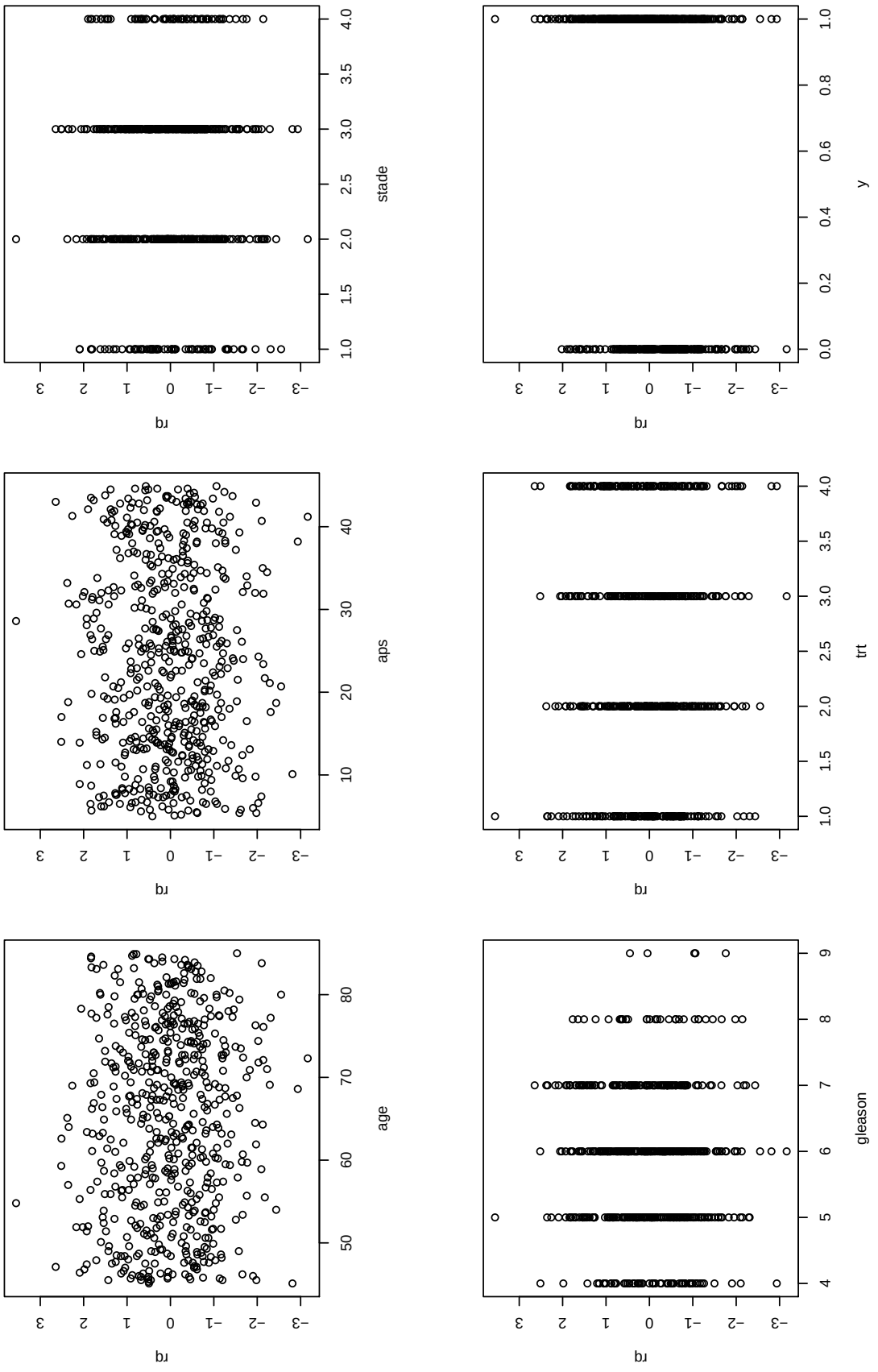


FIGURE 5 – Tracé relatif aux données sur le cancer de la prostate de l'exercice 22.

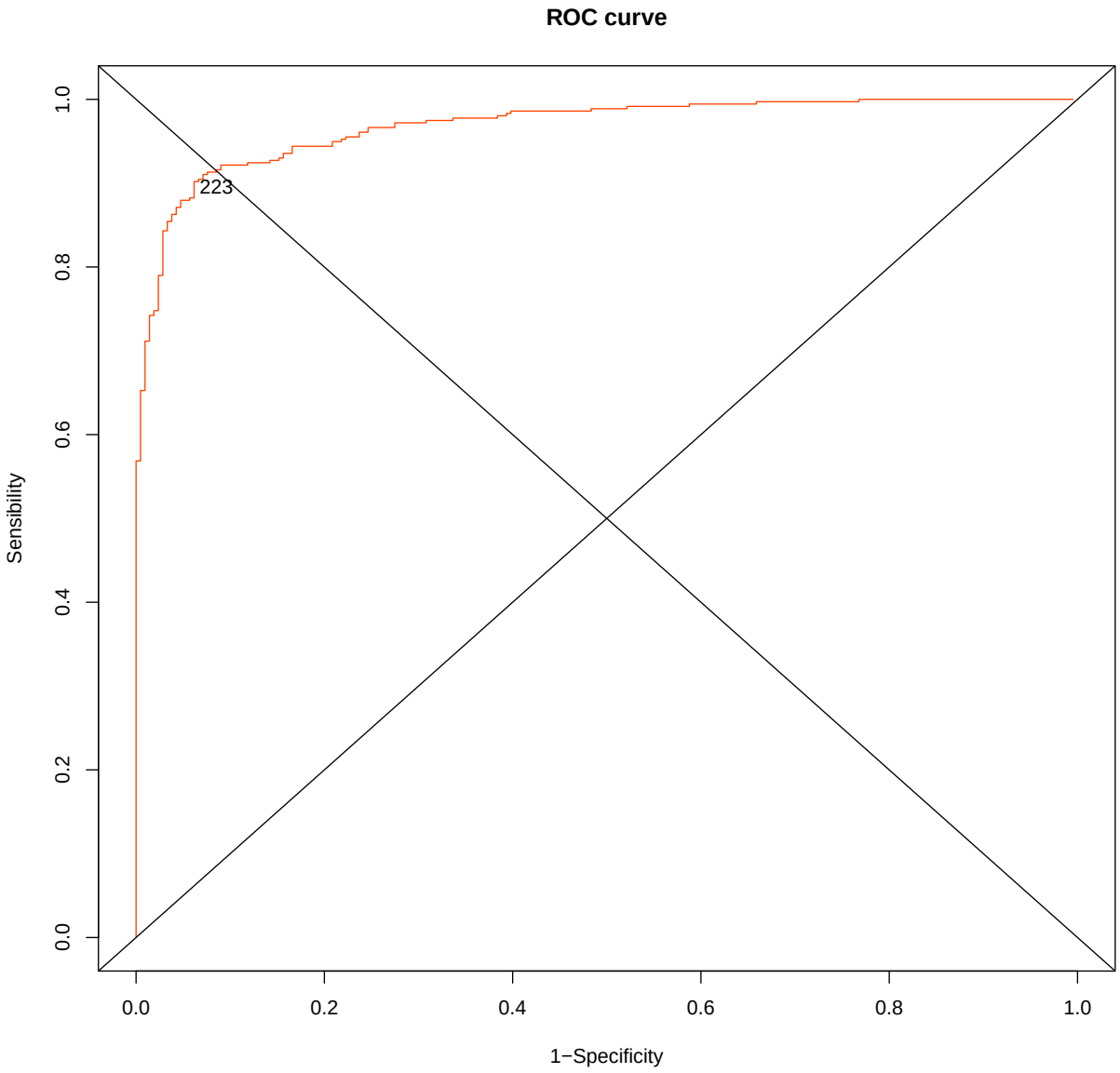


FIGURE 6 – Tracé relatif aux données sur le cancer de la prostate de l'exercice 22.

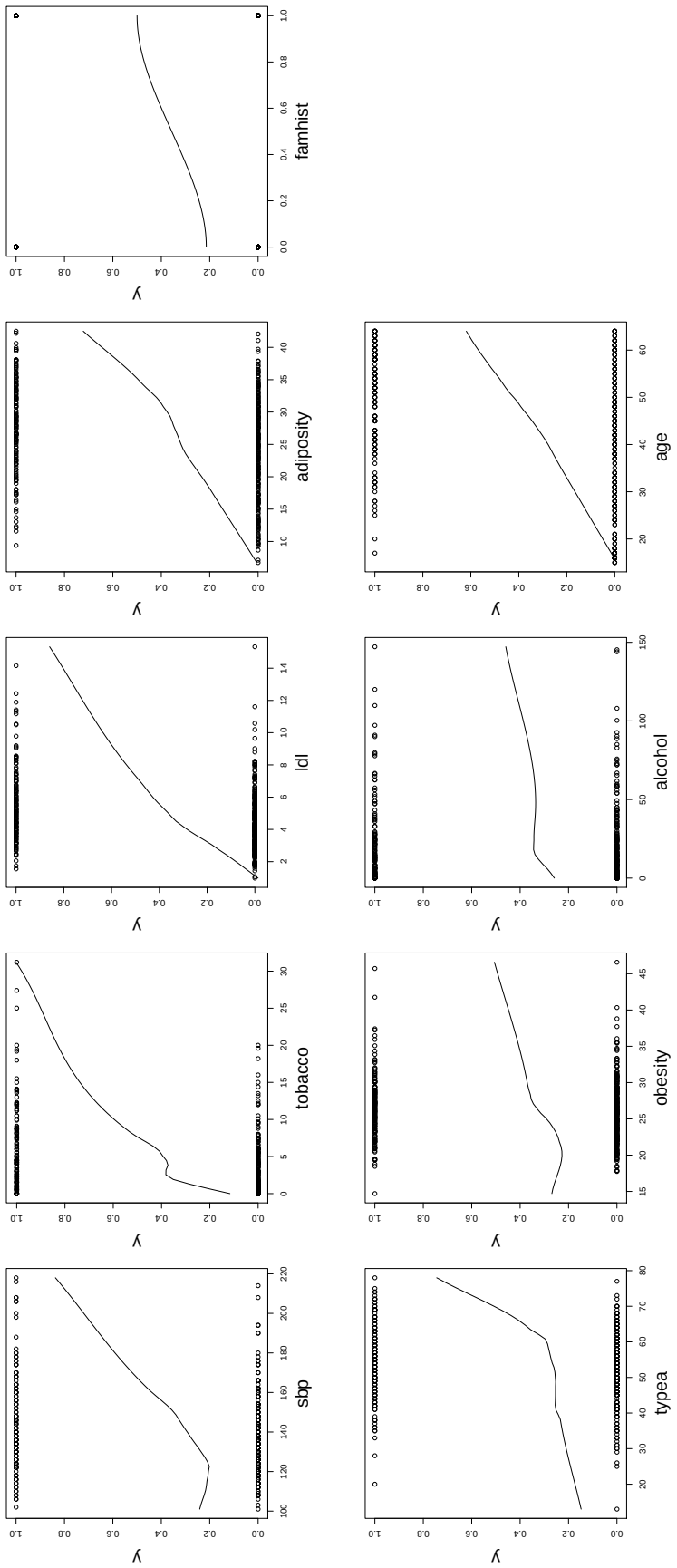


FIGURE 7 – Tracé exploratoire relatif aux données `heart.data` de l'exercice 27.

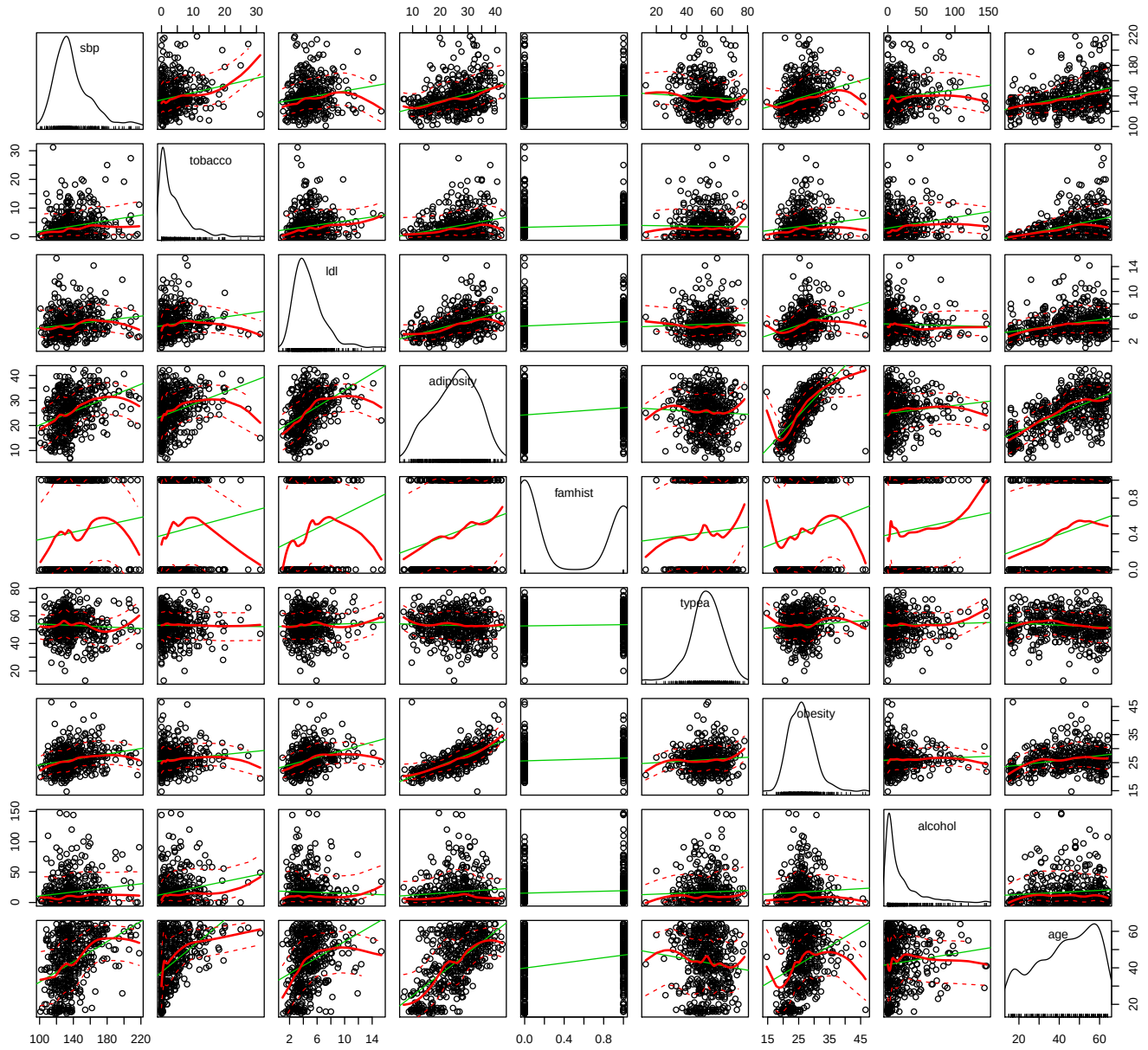


FIGURE 8 – Tracé exploratoire relatif aux données `heart.data` de l'exercice 27.

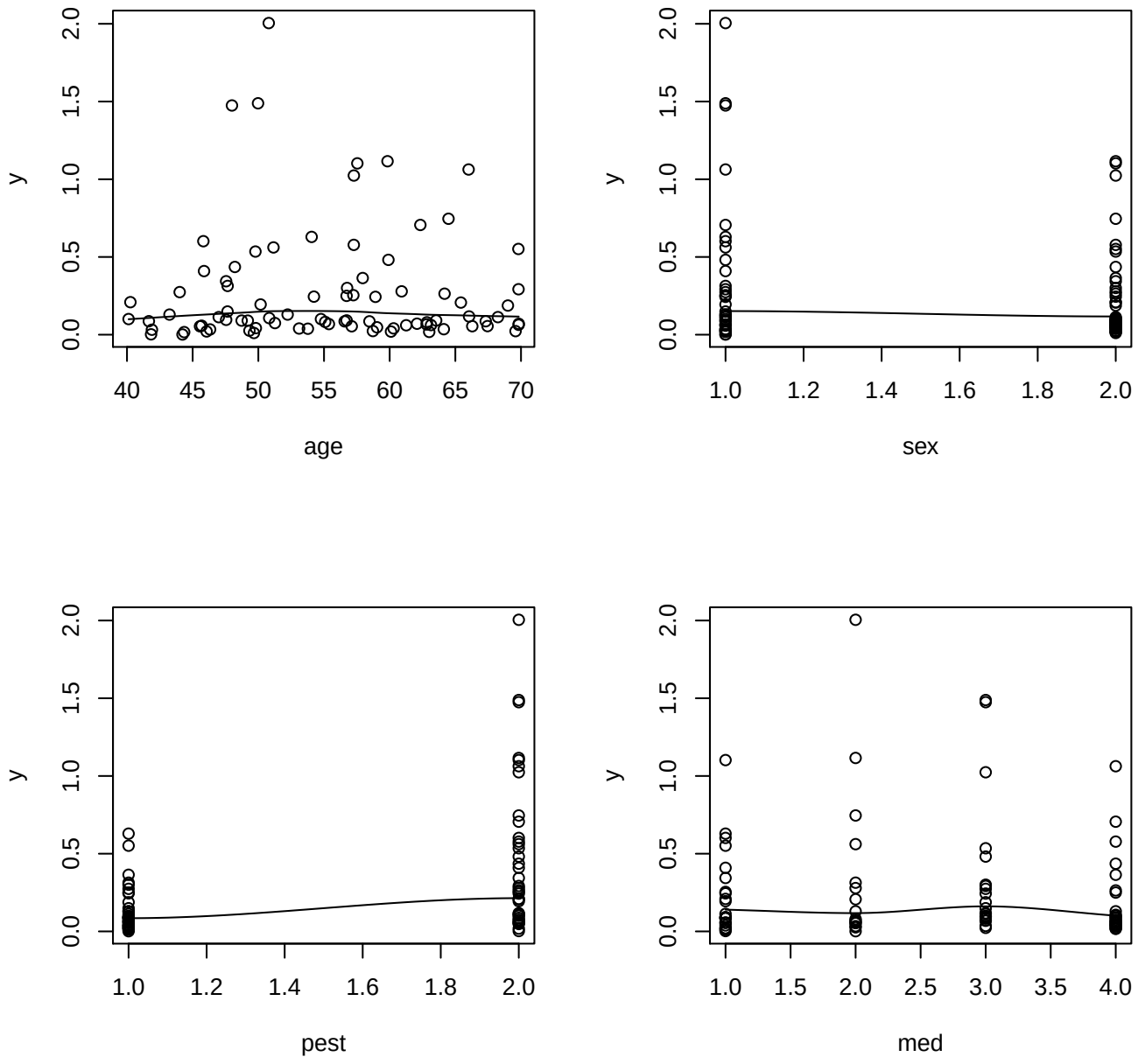


FIGURE 9 – Tracé exploratoire relatif aux données de l'exercice 30.

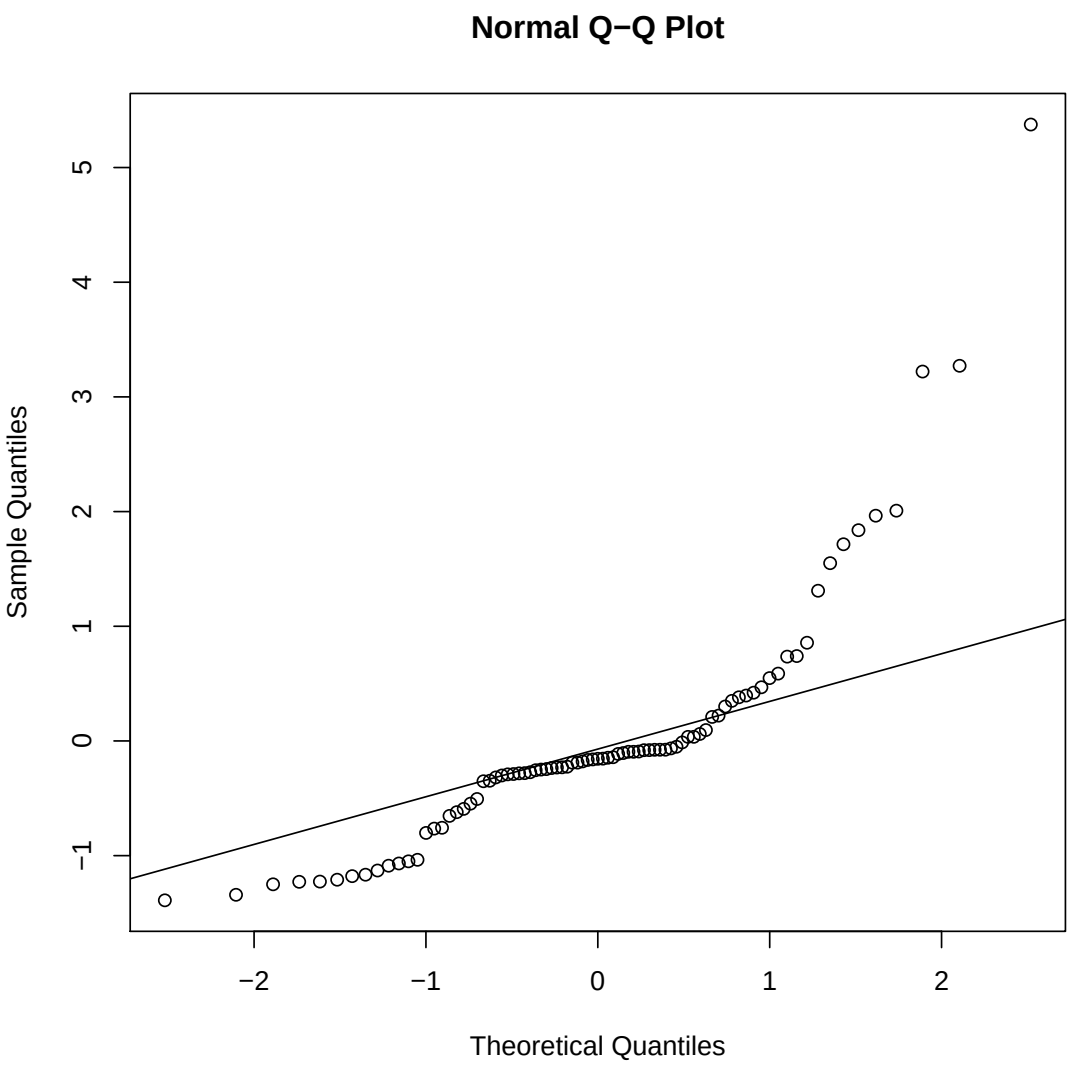


FIGURE 10 – Tracé relatif à l'exercice 30 effectué avec les instructions `qqnorm(res1)` et `qqline(res1)`.

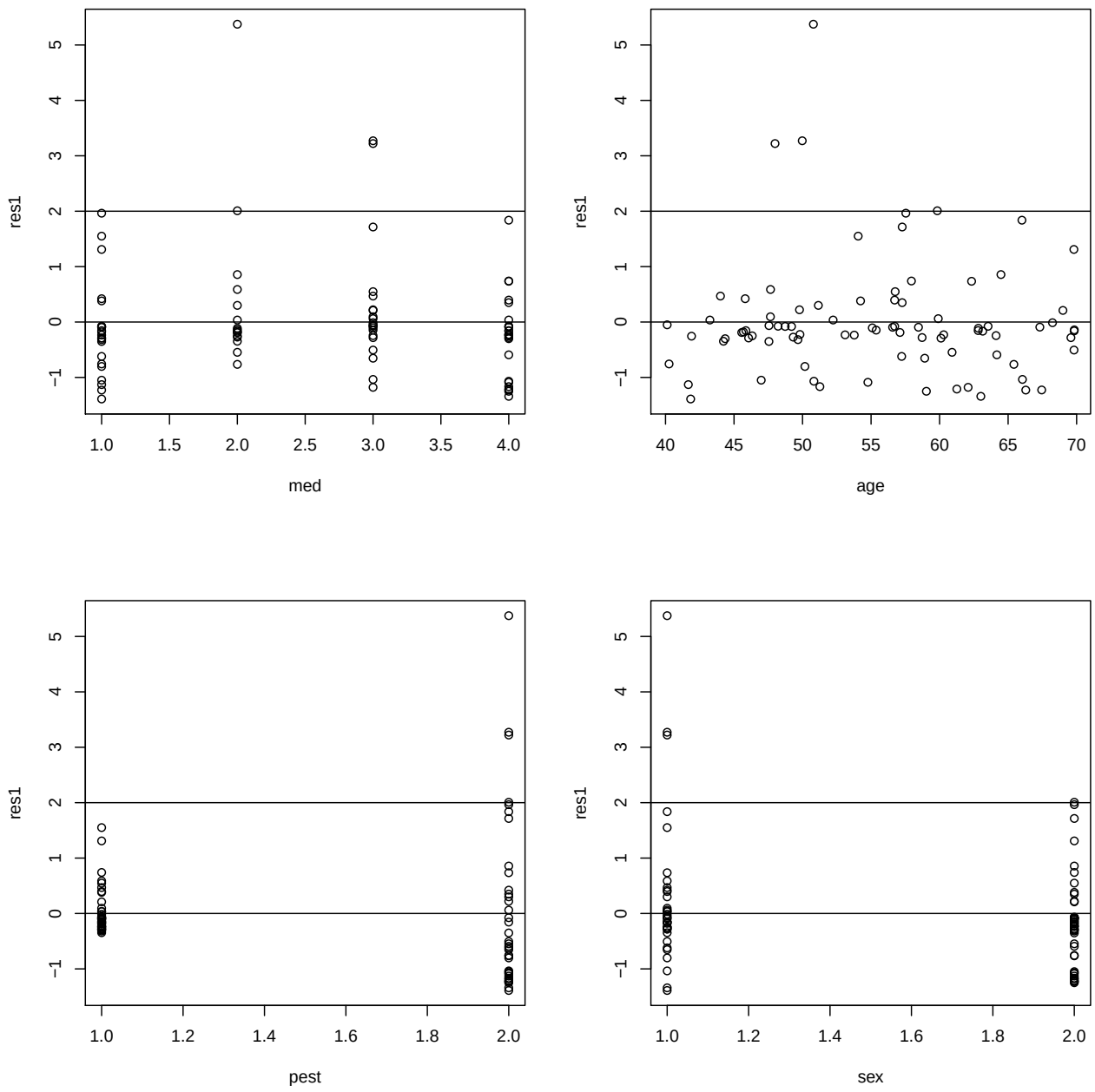


FIGURE 11 – Tracé relatif à l'exercice 30 effectué avec les valeurs stockées dans `res1`.

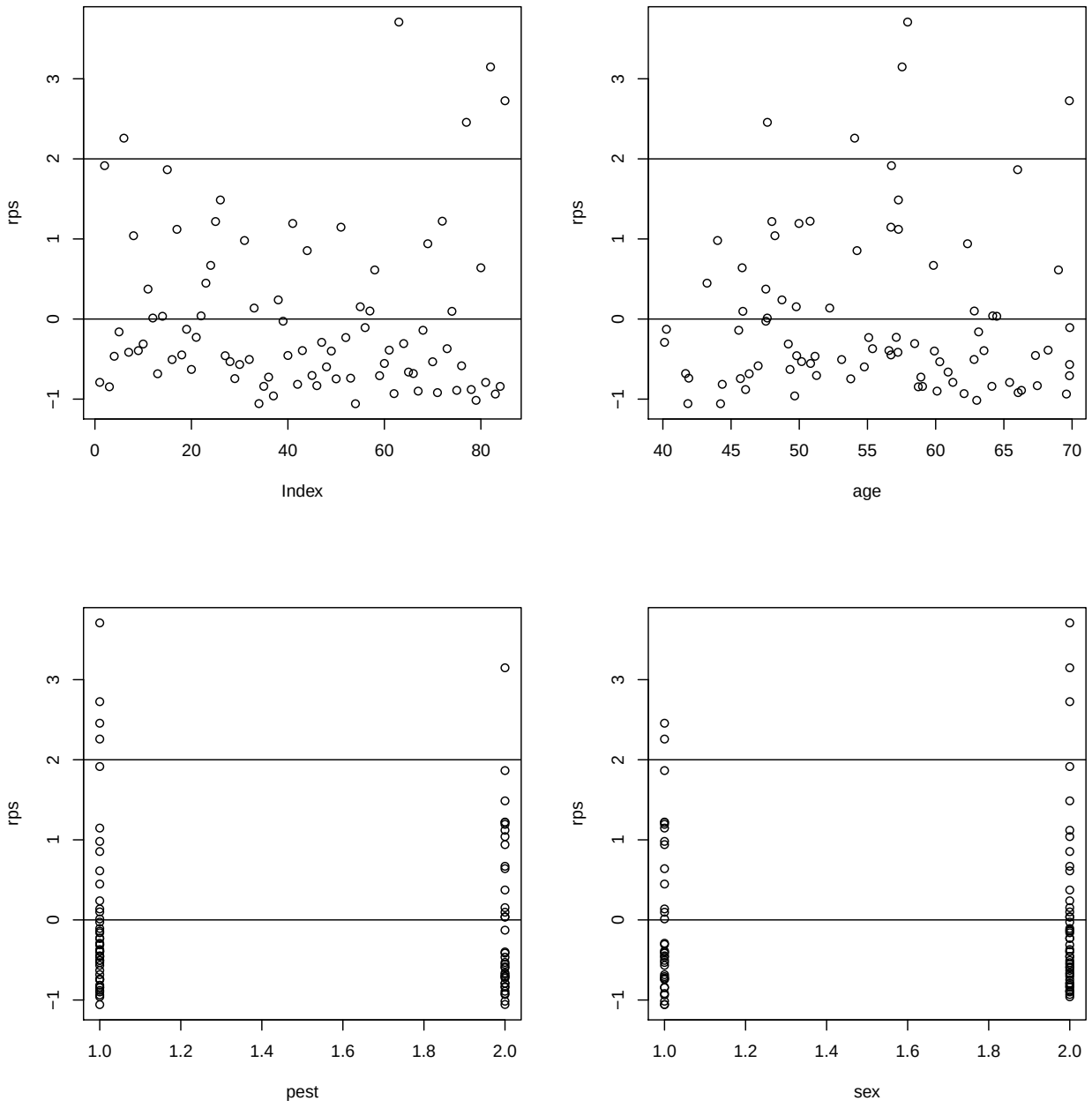


FIGURE 12 – Tracé relatif à l'exercice 30 effectué avec les valeurs stockées dans `rps`.

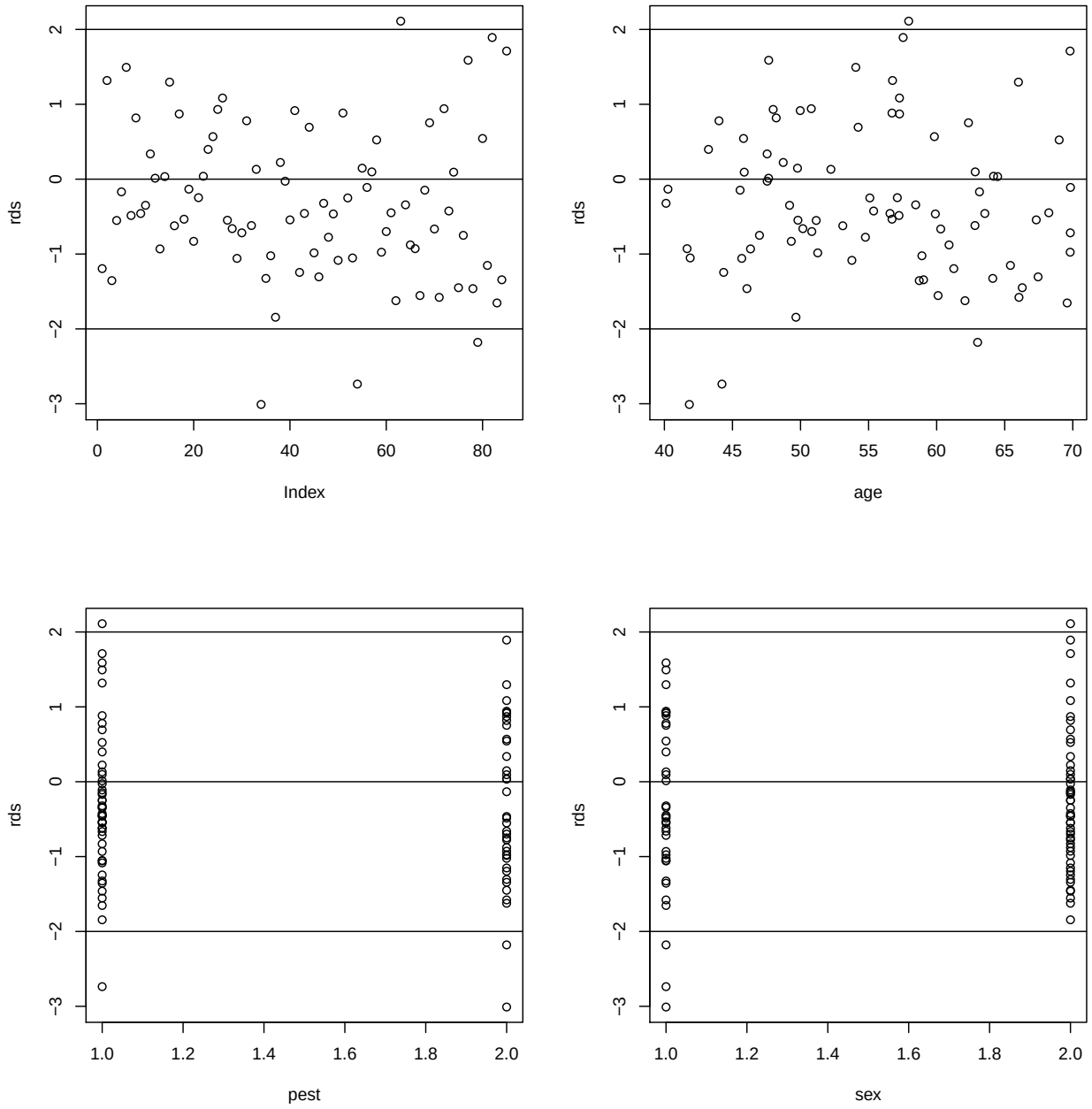


FIGURE 13 – Tracé relatif à l'exercice 30 effectué avec les valeurs stockées dans `rds`.