Feuille de Travaux Dirigés n$^{\text{o}}$ 6
Modèle linéaire généralisé

Les exemples de cette feuille de travaux dirigées sont tirés du livre A Handbook of Statistical Analyses Using R*, de Brian S. Everitt and Torsten Hothorn, CRC Press 2007.*

Exercice VI.1. Sédimentation des globules rouges
Le taux de sédimentation des globules rouges est le taux avec lequel les globules rouges sont en suspension. Il peut être indicateur de certaines maladies si ce taux dépasse $20mm/h$. On étudie ici l'influence de deux protéines, la fibrinogène et la globuline, du plasma sur l'ESR en analysant des données de Collett et Jemain (1985). Nous avons noté par 0 une valeur de ESR inférieure à $20mm/h$ et par 1 une valeur supérieure ou égale à $20mm/h$.

fibrogen	globulin	ESR	fibrogen	globulin	ESR
2,52	38	0	2,56	31	0
2,19	33	0	2,18	31	0
3,41	37	0	2,46	36	0
3,22	38	0	2,21	37	0
3,15	39	0	2,60	41	0
2,29	36	0	2,35	29	0
3,15	36	0	2,68	34	0
2,60	38	0	2,23	37	0
2,88	30	0	2,65	46	0
2,28	36	0	2,67	39	0
2,29	31	0	2,15	31	0
2,54	28	0	3,34	30	0
2,99	36	0	3,32	35	0
5,06	37	1	3,34	32	1
2,38	37	1	3,53	46	1
2,09	44	1	3,93	32	1

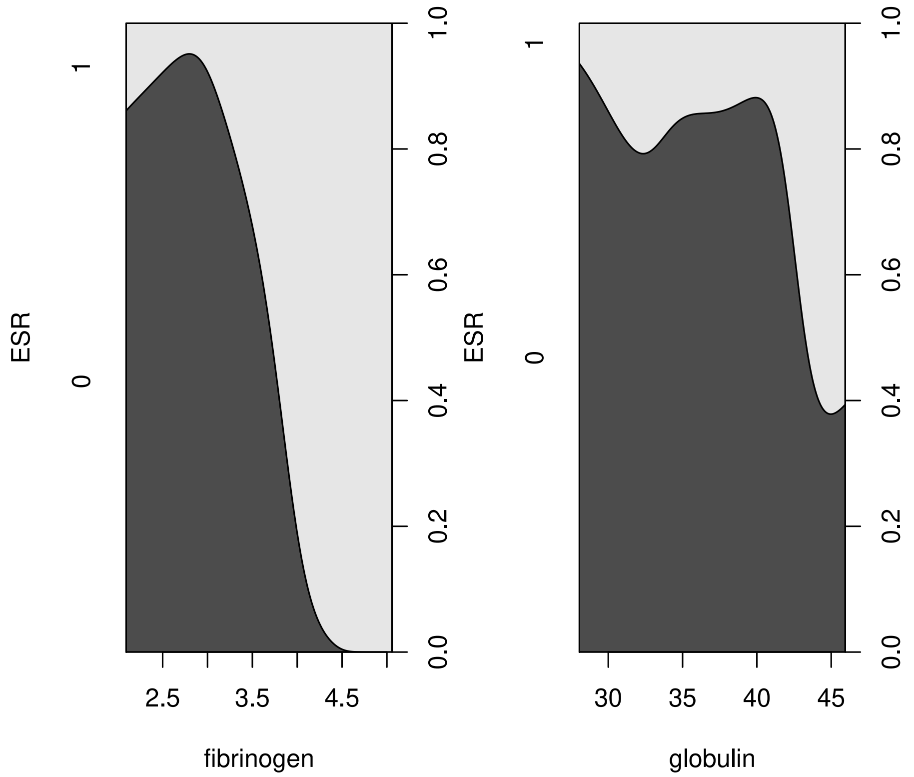
1. Récupérer les données dans R en exécutant les instructions suivantes[1].

```
> Chemin <- "C:\\..."
> plasma <- read.csv(paste(Chemin, "plasma.CSV", sep = ""))
```

Les lignes de commandes suivantes permettent de visualiser l'évolution de la proportion de ESR $> 20mm/h$ et de ESR $< 20mm/h$ en fonction des deux concentrations des protéines :

[1] Il faut remplacer `"C :\\..."` par le répertoire dans lequel vous avez enregistré le fichier que vous souhaitez ouvrir.

```
> layout(matrix(1:2, ncol = 2))
> cdplot(as.factor(ESR) ~ fibrinogen, data = plasma, ylab = "ESR")
> cdplot(as.factor(ESR) ~ globulin, data = plasma, ylab = "ESR")
> layout(1)
```



2. Justifier l'utilisation d'un modèle de régression logistique. L'intruction suivante permet de l'ajustement avec R d'un modèle avec comme seule variable explicative la mesure de fibrinogen :

```
> plasma_glm_1 <- glm(ESR ~ fibrinogen, data = plasma,
+     family = binomial())
> summary(plasma_glm_1)
```

3. Nous calculons un intervalle de confiance à 95 % du paramètre associé à la fibrinogène dans le modèle puis une estimation du rapport des côtes de succés et un intervalle de confiance à 95 % autour de celui-ci :

```
> confint(plasma_glm_1, parm = "fibrinogen")
> exp(coef(plasma_glm_1))["fibrinogen"]
> exp(confint(plasma_glm_1, parm = "fibrinogen"))
```

4. Nous nous intéressons maintenant à l'influence conjointe du fibrinogène et de la globuline :

```
> plasma_glm_2 <- glm(ESR ~ fibrinogen + globulin, data = plasma,
+     family = binomial())
> summary(plasma_glm_2)
```

5. Nous procédons au test des déviances. Il n'y a a priori pas d'avantage pour le modèle plus complexe. Nous verrons comment trancher ce problème à la question **6.**.

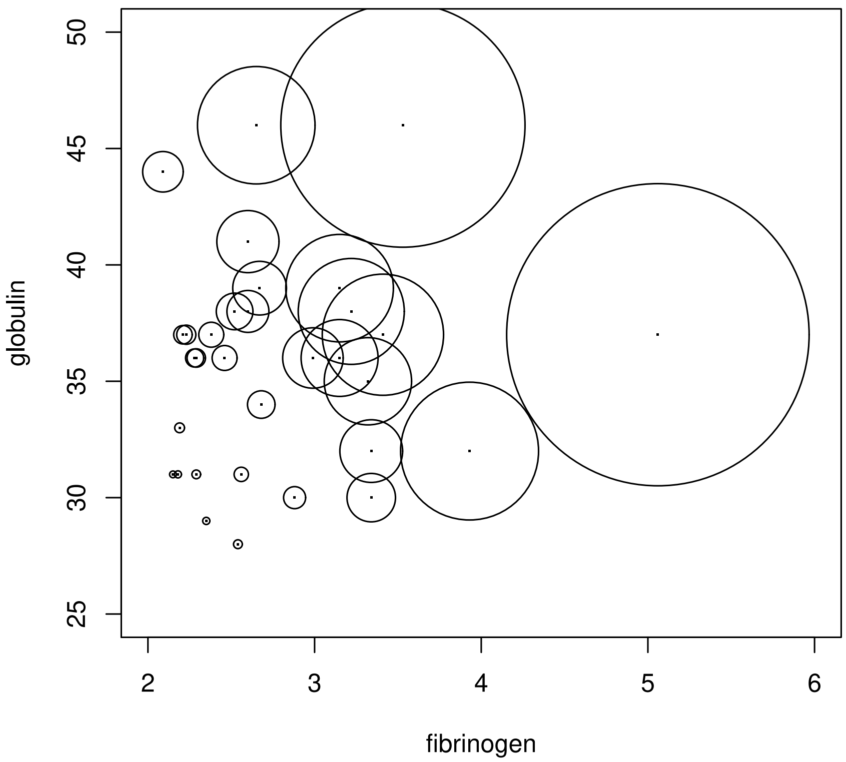
```
> anova(plasma_glm_1, plasma_glm_2, test = "Chisq")
```

6. Le package `Design` permet de réaliser des régressions logistiques avec la fonction `lrm` puis un test d'adéquation du modèle ainsi que de calculer plusieurs mesures d'associations avec la function `residuals.lrm` et l'option `'gof'`.

```
> library(Design)
> lrm_plasma_1 <- lrm(ESR ~ fibrinogen, data = plasma,
+     x = TRUE, y = TRUE)
> print(lrm_plasma_1)
> residuals.lrm(lrm_plasma_1, "gof")
> lrm_plasma_2 <- lrm(ESR ~ fibrinogen + globulin, data = plasma,
+     x = TRUE, y = TRUE)
> print(lrm_plasma_2)
> residuals.lrm(lrm_plasma_2, "gof")
```

7. L'ajustement du second modèle est bien meilleur que celui du premier, par conséquent nous le conservons et représentons les valeurs prédites par celui-ci en fonction des deux variables explicatives :

```
> prob <- predict(plasma_glm_2, type = "response")
> plot(globulin ~ fibrinogen, data = plasma, xlim = c(2,
+     6), ylim = c(25, 50), pch = ".")
> symbols(plasma$fibrinogen, plasma$globulin, circles = prob,
+     add = TRUE)
```



Exercice VI.2. Rôle des femmes dans la société
Lors d'une enquête réalisée de 1974 à 1975, les enquêteurs ont demandé à chacune des personnes interrogées si elle approuvait ou désapprouvait la phrase suivante : “ Les femmes devraient se préocupper des problèmes domestiques et laisser aux hommes la gestion des intérêts de la nation ”. Les réponses ont été récapitulées dans le tableau ci-dessous. Le sexe de la personne interrogée ou le nombre d'années d'étude affectent-ils la réponse ?

education	sex	agree	disagree	education	sex	agree	disagree
0	Male	4	2	0	Female	4	2
1	Male	2	0	1	Female	1	0
2	Male	4	0	2	Female	0	0
3	Male	6	3	3	Female	6	1
4	Male	5	5	4	Female	10	0
5	Male	13	7	5	Female	14	7
6	Male	25	9	6	Female	17	5
7	Male	27	15	7	Female	26	16
8	Male	75	49	8	Female	91	36
9	Male	29	29	9	Female	30	35
10	Male	32	45	10	Female	55	67
11	Male	36	59	11	Female	50	62
12	Male	115	245	12	Female	190	403
13	Male	31	70	13	Female	17	92
14	Male	28	79	14	Female	18	81
15	Male	9	23	15	Female	7	34
16	Male	15	110	16	Female	13	115
17	Male	3	29	17	Female	3	28
18	Male	1	28	18	Female	0	21
19	Male	2	13	19	Female	1	2
20	Male	3	20	20	Female	2	4

1. Récupérer les données dans R en exécutant les instructions suivantes[2].

```
> Chemin <- "C:\\..."
> rolefemmes <- read.csv(paste(Chemin, "rolefemmes.CSV",
+     sep = ""))
```

2. Quel est le type de modèle utilisé dans les instructions suivantes ? Pourquoi son utilisation est-elle judicieuse ?

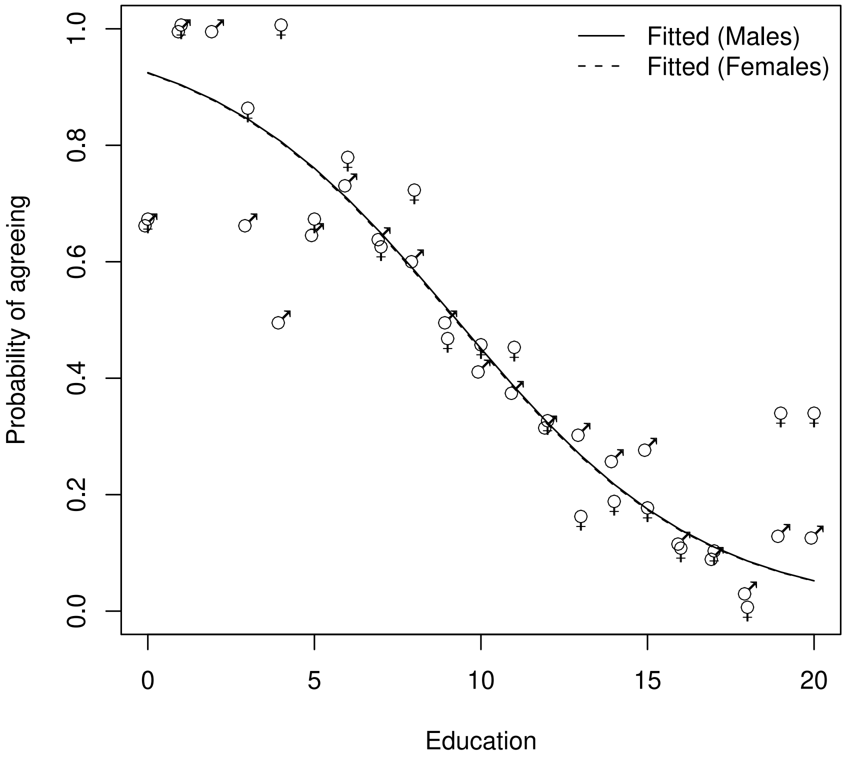
```
> womensrole_glm_1 <- glm(cbind(agree, disagree) ~ sex +
+     education, data = rolefemmes, family = binomial())
> summary(womensrole_glm_1)
```

3. Nous souhaitons représenter les probabilités prédites par le modèle avec celles observées sur l'échantillon afin de déterminer graphiquement si le modèle proposé est en adéquation avec les données qu'il est sensé modéliser.

```
> role.fitted1 <- predict(womensrole_glm_1, type = "response")
> dessin <- function(role.fitted) {
+     f <- rolefemmes$sex == "Female"
+     plot(rolefemmes$education, role.fitted, type = "n",
+         ylab = "Probability of agreeing", xlab = "Education",
+         ylim = c(0, 1))
+     lines(rolefemmes$education[!f], role.fitted[!f],
+         lty = 1)
```

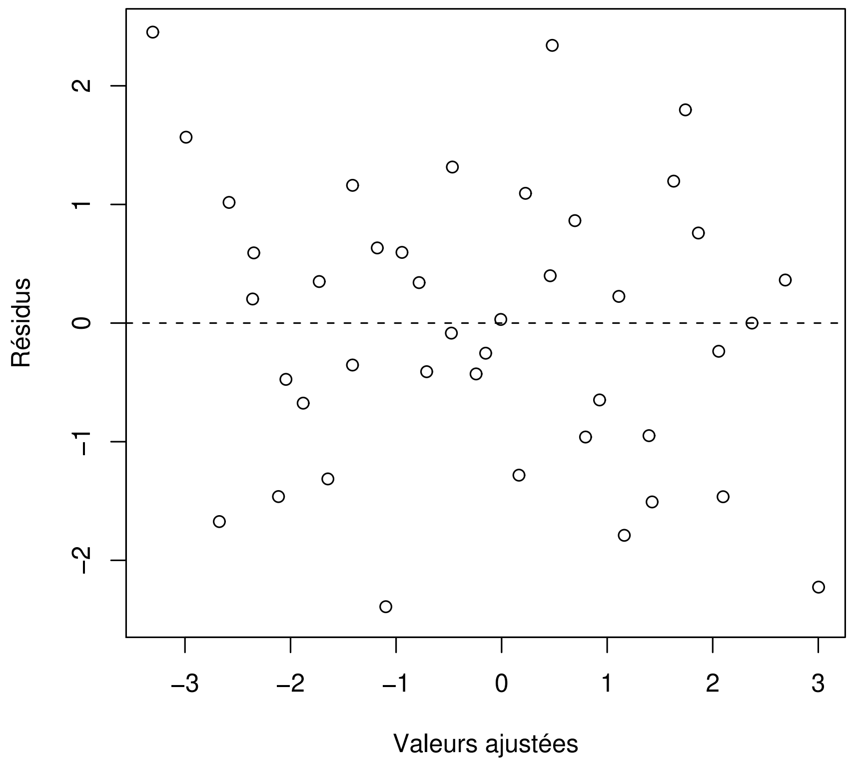
[2]Il faut remplacer `"C :\\..."` par le répertoire dans lequel vous avez enregistré le fichier que vous souhaitez ouvrir.

```
+      lines(rolefemmes$education[f], role.fitted[f], lty = 2)
+      lgtxt <- c("Fitted (Males)", "Fitted (Females)")
+      legend("topright", lgtxt, lty = 1:2, bty = "n")
+      y <- rolefemmes$agree/(rolefemmes$agree + rolefemmes$disagree)
+      text(rolefemmes$education, y, ifelse(f, "\\VE",
+          "\\MA"), vfont = c("serif", "plain"), cex = 1.25)
+ }
> dessin(role.fitted1)
```



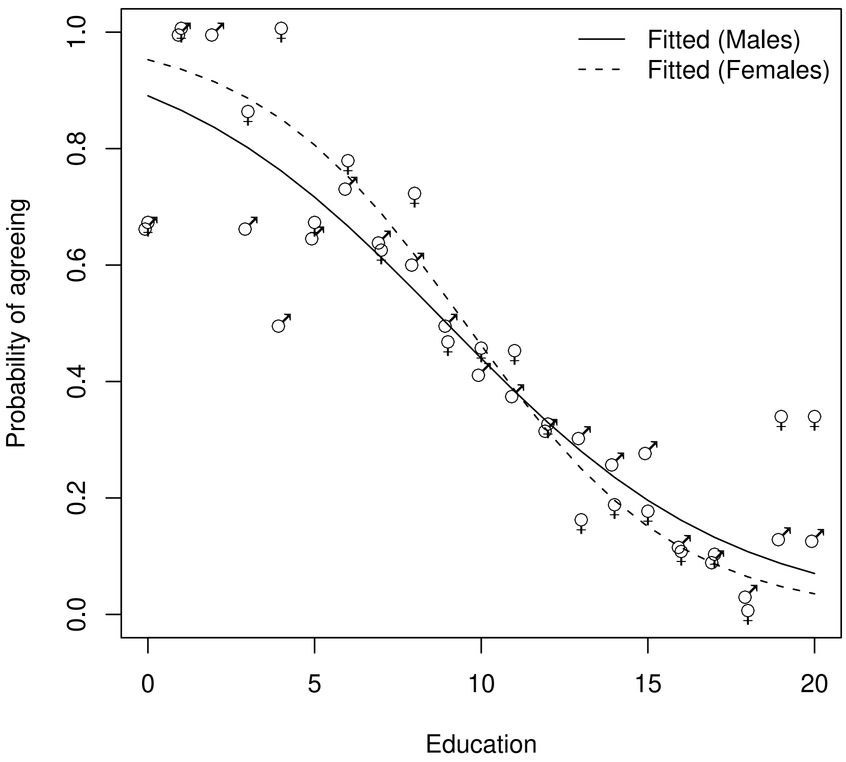
4. Qu'est-ce que le modèle suivant apporte comme modification par rapport au modèle introduit dans la question **2.** ?

```
> womensrole_glm_2 <- glm(cbind(agree, disagree) ~ sex +
+      education + sex * education, data = rolefemmes,
+      family = binomial())
> summary(womensrole_glm_2)
> res <- residuals(womensrole_glm_2, type = "deviance")
> plot(predict(womensrole_glm_2), res, xlab = "Valeurs ajustées",
+      ylab = "Résidus", ylim = max(abs(res)) * c(-1, 1),
+      cex = 1)
> abline(h = 0, lty = 2)
```



5. Comparer le graphique obtenu par le code suivant avec celui produit au **3.**. Que constatez-vous ?

```
> role.fitted2 <- predict(womensrole_glm_2, type = "response")
> dessin(role.fitted2)
```



Exercice VI.3. Polypes

Giardiello *et al.* (1993) and Piantadosi (1997) ont décrit les résultats d'une étude avec un groupe de contrôle recevant un placebo d'un médicament anti-inflammatoire dans le traitement de polypes (FAP). Le tableau ci-dessus donne le nombre de polypes après 12 mois de traitement.

number	treat	age	number	treat	age
63	P	20	3	M	23
2	M	16	28	P	22
28	P	18	10	P	30
17	M	22	40	P	27
61	P	13	33	M	23
1	M	23	46	P	22
7	P	34	50	P	34
15	P	50	3	M	23
44	P	19	1	M	22
25	M	17	4	M	42

1. Récupérer les données dans R en exécutant les instructions suivantes[3].

```
> Chemin <- "C:\\..."
> polyps <- read.csv(paste(Chemin, "polyps.CSV", sep = ""))
> attach(polyps)
```

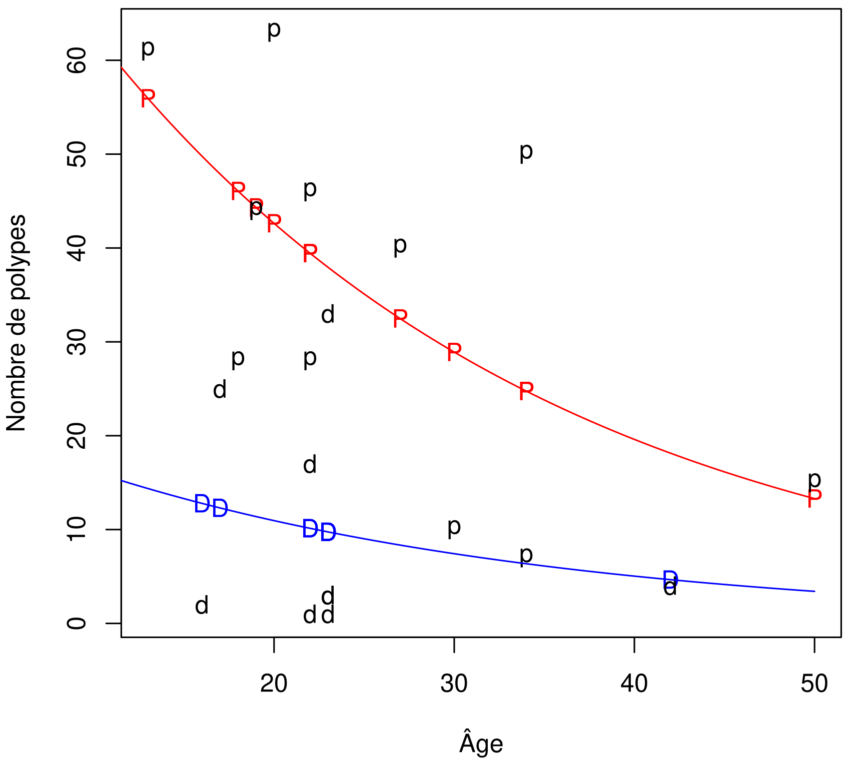
2. Le nombre de polype est-il lié à l'âge des patients et/ou au traitement reçu ? Commentez les résultats obtenus lorsque vous exécutez les lignes de commande suivantes.

```
> polyps_glm_1 <- glm(number ~ treat + age, family = poisson())
> summary(polyps_glm_1)
> confint(polyps_glm_1)
> polyps_glm_2 <- glm(number ~ treat + age, family = quasipoisson())
> summary(polyps_glm_2)
> confint(polyps_glm_2)
```

La représentation graphique obtenue en exécutant les instructions suivantes permet de visualiser le second modèle et sert d'aide à l'interprétation des résultats.

```
> plot(number ~ age, type = "n", ylab = "Nombre de polypes",
+     xlab = "Âge")
> points(age[treat == "placebo"], fitted(polyps_glm_2)[treat ==
+     "placebo"], pch = "P", col = "red")
> xv1 <- seq(0, 50, 0.05)
> yv1 <- predict(polyps_glm_2, list(treat = as.factor(rep("placebo",
+     length(xv1))), age = xv1))
> lines(xv1, exp(yv1), col = "red")
> points(age[treat == "placebo"], number[treat == "placebo"],
+     pch = "p")
> points(age[treat == "drug"], fitted(polyps_glm_2)[treat ==
+     "drug"], pch = "D", col = "blue")
> points(age[treat == "drug"], number[treat == "drug"],
+     pch = "d")
> yv2 <- predict(polyps_glm_2, list(treat = as.factor(rep("drug",
+     length(xv1))), age = xv1))
> lines(xv1, exp(yv2), col = "blue")
```

[3] Il faut remplacer `"C :\\..."` par le répertoire dans lequel vous avez enregistré le fichier que vous souhaitez ouvrir.




```
> detach(polyps)
```

Remarque
Nous avons utilisé dans cet exercice une régression de Poisson avec un lien log pour laquelle la distribution des erreurs est une loi de Poisson définie par

$$P_\lambda(y) = \frac{e^{-\lambda}\lambda^y}{y!}.$$

La fonction de variance associée est :

$$V(\mu) = \mu,$$

c'est-à-dire que la variance est égale à la moyenne.
Dans certains cas, une distribution quasipoisson peut être utile pour tenir compte d'une sousdispersion ou d'une surdispersion des données. La nécessité d'un recours à une distribution quasipoisson se détecte en comparant la valeur de la déviance résiduelle à celle du nombre de ses degrés de liberté. Si ces deux nombres sont proches alors une distribution de Poisson classique peut convenir pour ajuster les données, par contre si la déviance résiduelle est sensiblement plus grande (ou plus petite) que le nombre de ses degrés de liberté alors nous sommes face à un cas de surdispersion (sousdispersion). On introduit alors un paramètre supplémentaire ϕ, dit de dispersion, et la fonction de variance associée devient :

$$V(\mu) = \phi\mu,$$

ce qui permet de modéliser une dépendance linéaire en la moyenne μ.
Dans le premier modèle `glm_polyps_1` nous avions une valeur de la `Residual deviance` égale à 179.541 donc largement supérieure au nombre de `degrees of freedom` égal à 17. Le second modèle introduit le paramètre ϕ de dispersion dont la valeur est estimée à 10.728. Remarquez que le critère d'information AIC n'est plus calculable dans le cas d'une famille quasiPoisson.
Il existe également une famille quasibinomiale pour modéliser les problèmes de surdispersion ou de sous-dispersion d'un modèle basé sur la famille binomiale. Elle s'utilise dans les mêmes conditions.

Exercice VI.4. Cancers et centrale nucléaire
Le tableau suivant contient le nombre de cas de cancer par an par clinique ainsi que la distance de la clinique à la centrale nucléaire la plus proche.

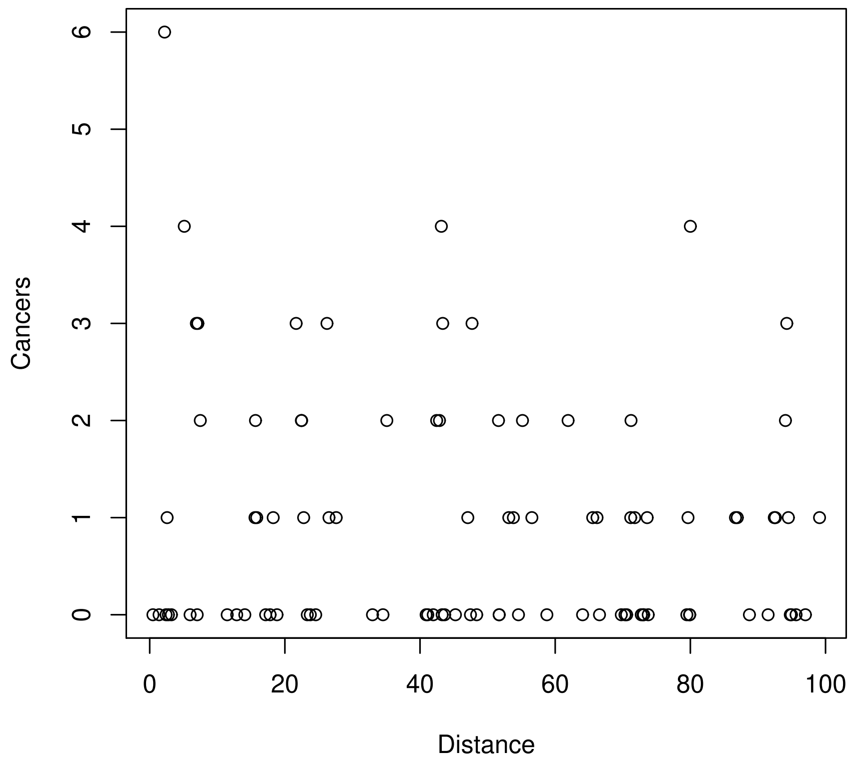
Cancers	Distance	Cancers	Distance	Cancers	Distance
0	11.46952498	0	79.4807305	1	53.81101526
0	66.55394864	0	1.386907856	1	99.11360431
0	47.46230003	0	41.95475353	1	15.84067672
0	48.38128833	0	58.76711801	1	92.41880212
0	73.76534394	0	72.99865268	1	66.18831233
0	70.57555322	0	40.90071201	1	56.54806228
0	43.64083897	0	7.025356296	1	79.6523586
0	17.81218119	0	17.15687836	1	86.94647677
0	54.56414901	0	18.83508269	2	22.48104627
0	95.64008653	0	0.479771068	2	7.488115026
0	2.422537722	0	32.96435415	2	61.91036983
0	12.87567112	0	64.05478868	2	42.87475137
0	73.09881144	0	3.212366303	2	94.06591258
0	24.55984011	0	69.78215864	2	42.46016785
0	23.32238998	0	51.71269425	2	22.44114861
0	70.32468964	0	91.50296577	2	15.65324062
0	79.90967735	1	71.75949918	2	51.60880618
0	14.07392477	1	71.1741827	2	71.21820854
0	41.14788517	1	22.78097983	2	55.15873665
0	34.51294189	1	53.13556856	2	35.08534093
0	72.73501428	1	65.55151643	3	94.26593737
0	51.70798334	1	2.575546962	3	6.90362832
0	45.23712686	1	15.57941073	3	21.66642883
0	94.9555794	1	26.52545168	3	7.130288492
0	79.87797925	1	27.59364034	3	47.69598044
0	2.766892529	1	18.27444924	3	43.35659292
0	94.78422627	1	94.49867119	3	26.23175014
0	97.01885148	1	92.58247404	4	43.1507975
0	43.27307209	1	73.61021651	4	5.114980886
0	5.973689643	1	47.06909093	6	2.2
0	88.73868531	1	86.67474578	4	80
0	23.76115956				

1. Récupérer les données dans R en exécutant les instructions suivantes[4].

```
> Chemin <- "C:\\..."
> cancers <- read.csv(paste(Chemin, "cancers.CSV", sep = ""))
> attach(cancers)
> names(cancers)
```

2. Représenter graphiquement les données. Semble-t-il y avoir une dépendance du nombre de cancers à la distance de la clinique à la centrale nucléaire la plus proche ?

```
> plot(Distance, Cancers)
```



3. À la vue des résultats des lignes de commande ci-dessous, le recours à une distribution quasiPoisson est-il justifié ? Le jeu de données permet-il de mettre en évidence une influence significative, au seuil $\alpha = 5\%$ de distance de la clinique à la centrale nucléaire la plus proche ?

```
> model1 <- glm(Cancers ~ Distance, poisson)
> summary(model1)
> model2 <- glm(Cancers ~ Distance, quasipoisson)
> summary(model2)
```

4. La graphique ci-dessous représente le jeu de données et la courbe de régression ajustée par le second modèle. Quels commentaires pouvez-vous formuler ?

[4]Il faut remplacer `"C :\\..."` par le répertoire dans lequel vous avez enregistré le fichier que vous souhaitez ouvrir.

```
> plot(Distance, Cancers)
> xv <- seq(0, 100, 0.1)
> yv <- predict(model2, list(Distance = xv))
> lines(xv, exp(yv))
```