LSTAT2100 - Exercices - Série 4

Solutions

Les exercices de ce TP sont en très grande partie tirés d'examens antérieurs.

Excercice 1

Considérons n variables aléatoires indépendantes $Y_1, \dots, Y_n$ telles que Y_i est distribuée selon une loi binomiale négative ($NBin$) avec les paramètres (k, π_i), où $k \geq 2$ est un entier connu et $\pi_i \in (0, 1)$ est inconnu. La loi binomiale négative (k, π_i) modélise, dans le contexte d'une suite d'épreuves indépendantes de Bernoulli, le nombre d'essais nécessaires pour obtenir k succès, π_i représente la probabilité de succès. Nous avons

$$f(y_i; \pi_i) = C_{y_i-1}^{k-1} \pi_i^k (1-\pi_i)^{(y_i-k)}, \;\; y_i = k, k+1, \dots .$$

(a)

Montrer que cette loi appartient à la famille exponentielle. En suivant les notations du cours quant aux familles exponentielles, on explicitera les paramètres θ et ϕ, et la fonction b.

Réponse

Puisque f peut être écrite comme,

$$f(y_i; \pi_i) = C_{y_i-1}^{k-1} \exp\left(y_i \log(1-\pi_i) + k \log \frac{\pi_i}{1-\pi_i} \right),$$

la loi négative binomiale, avec k connu, fait donc partie de la famille exponentielle avec

$$\theta_i = \log(1-\pi_i) \iff \pi_i = 1 - e^{\theta_i}, \quad \phi = 1, \text{ et}$$
$$b(\theta_i) = -k \log \frac{1-e^{\theta_i}}{e^{\theta_i}}.$$

(b)

Calculez $E(Y_i)$ et $Var(Y_i)$.

Réponse

$$E(Y_i) = b' = \frac{k}{1 - e^{\theta_i}} = \frac{k}{\pi_i}$$
$$Var(Y_i) = \phi b'' = k\frac{e^{\theta_i}}{(1 - e^{\theta_i})^2} = k\frac{1 - \pi_i}{\pi_i^2}.$$

En plus du fait que $Y_i \sim NBin(k, \pi_i)$, $i = 1, \ldots, n$, on suppose dans ce qui suit que

$$\log(1 - \pi_i) = \beta_0 + \beta_1 x_i,$$

où x_i est une variable à valeurs dans $\mathbb{R}$.

(c)

Montrez qu'il s'agit d'un modèle linéaire généralisé. La fonction de lien canonique a-t-elle été utilisée ? Justifiez et détaillez votre raisonnement.

Réponse

La loi négative binomiale étant membre de la famille exponentielle, le glm qui en découle, avec la fonction de lien g, est

$$\mu_i = g^{-1}(\beta_0 + \beta_1 x_i)$$
$$\frac{k}{\pi_i} = g^{-1}(\beta_0 + \beta_1 x_i)$$
$$\pi_i = \frac{k}{g^{-1}(\beta_0 + \beta_1 x_i)}.$$

Par définition, la fonction de lien canonique est $g(\theta) = (b')^{-1}(\theta)$. Ainsi, le glm avec le lien canonique ci-dessus correspond au modèle

$$\pi_i = 1 - \exp(\beta_0 + \beta_1 x_i).$$

Donc oui, le modèle stipulé est bien un glm canonique.

(d)

Pour le modèle défini ci-dessus, donnez les équations de vraisemblance qui déterminent les estimateurs du maximum de vraisemblance de β_0 et β_1. Développez et simplifiez vos calculs. Est-il possible d'obtenir des expressions explicites pour ces estimateurs ? Dans l'affirmative, procédez aux calculs qui s'imposent.

Réponse

La log-vraisemblance est donnée par

$$\ell_n = \sum_{i=1}^{n} \{k \log(\pi_i) + (y_i - k) \log(1 - \pi_i)\} + Const$$

Les équations de vraisemblances sont $\partial_{\beta_0} \ell_n = 0$ ***et*** $\partial_{\beta_1} \ell_n = 0$ qui doivent être résolus simultanément pour trouver les EMVs. Étant donné que $\partial_{\beta_0} \pi_i = -(1 - \pi_i)$ et que $\partial_{\beta_1} \pi_i = -(1 - \pi_i) x_i$, on a

$$\begin{aligned}
\partial_{\beta_0} \ell_n &= \sum_i \left\{ k \frac{1}{\pi_i} \partial_{\beta_0} \pi_i - (y_i - k) \frac{1}{1 - \pi_i} \partial_{\beta_0} \pi_i \right\} \\
&= \sum_i \left\{ y_i - k - k \frac{1 - \pi_i}{\pi_i} \right\} = \sum_i (y_i - k) - k \sum_i \frac{\exp(\beta_0 + \beta_1 x_i)}{1 - \exp(\beta_0 + \beta_1 x_i)}.
\end{aligned}$$

De même,

$$\partial_{\beta_1} \ell_n = \sum_i (y_i - k) x_i - k \sum_i \frac{\exp(\beta_0 + \beta_1 x_i)}{1 - \exp(\beta_0 + \beta_1 x_i)} x_i.$$

Une résolution explicite de ces équations n'est clairement pas possible.

(e)

Donnez les expressions de la déviance et de la déviance nulle correspondant au modèle défini ci-dessus. Simplifiez au maximum vos formules.

Réponse

La déviance est $D^2 = 2(\ell_n^S - \ell_n)$, où ℓ_n est le maximum de la log-vraisemblance du modèle considéré, et ℓ_n^S est la log-vraisemblance saturée obtenue en maximisant la vraisemblance (telle que formulée ci-dessus) mais sans imposer de contrainte sur les π_i.

Il est facile de voir que maximiser ℓ_n par rapport aux π_i, correspond à remplacer π_i par $\hat{\pi}_i^S = \frac{k}{y_i}$. Autrement dit,

$$\ell_n^S = \sum_{i=1}^{n} \left\{ k \log(\hat{\pi}_i^S) + (y_i - k) \log(1 - \hat{\pi}_i^S) \right\} + Const$$

Et donc

$$D^2 = 2 \sum_i \left\{ k \log \frac{k}{\hat{\pi}_i y_i} + (y_i - k) \log \frac{y_i - k}{(1 - \hat{\pi}_i) y_i} \right\},$$

où $\hat{\pi}_i = 1 - \exp\left(\hat{\beta}_0 - \hat{\beta}_1 x_i\right)$.

La déviance nulle est $D_0^2 = 2(\ell_n^S - \ell_n^0)$, où ℓ_n^0 est le maximum de la log-vraisemblance du modèle nulle, càd du modèle avec un π_i constant (càd $\pi_i = \pi_1$, $\forall i$). Dans ce cas, Il est facile de voir que l'EMV de π_1 n'est rien d'autre que

$$\hat{\pi}_1^0 = \frac{nk}{\sum_i y_i}$$

Et donc

$$D_0^2 = 2 \sum_i \left\{ k \log \frac{\sum_i y_i}{n y_i} + (y_i - k) \log \frac{(y_i - k) \sum_i y_i}{y_i \sum_i (y_i - k)} \right\}$$

Excercice 2

Le tableau suivant donne le nombre de candidats (admis ou rejetés) à une certaine prestigieuse université américaine pour six grands départements (A, B, ..., F). La question de la discrimination à l'égard des femmes dans l'accès à l'enseignement supérieur étant au cœur de cette étude, le genre est rapporté.

Gender	Male		Female	
Admit	Rejected	Admitted	Rejected	Admitted
Dept				
A	313	512	19	89
B	207	353	8	17
C	205	120	391	202
D	279	138	244	131
E	138	53	299	94
F	351	22	317	24

(a)

Calculez la proportion de candidats admis à cette université (tous départements confondus) par genre. Représentez ces chiffres à l'aide d'un graphique approprié. Que constatez-vous ?

Réponse

Commençons par enregistrer les données dans R.

```
DF <- data.frame(Dept = factor(rep(c("A", "B", "C", "D", "E", "F"), each = 4)),
    Gender = factor(c("Male", "Female", "Male", "Female"), levels = c("Male", "Female")),
    Admit = factor(c("Admitted", "Admitted", "Rejected", "Rejected"), levels = c("Rejected",
        "Admitted")), Freq = c(512, 89, 313, 19, 353, 17, 207, 8, 120, 202, 205,
        391, 138, 131, 279, 244, 53, 94, 138, 299, 22, 24, 351, 317))
```

Voici la table de contingence *Gender* × *Admit*

```
TB <- xtabs(Freq ~ Gender + Admit + Dept, data = DF)
tbm <- TB |> margin.table(margin = c("Gender", "Admit")) |>
  print()
#or tbm <- xtabs(Freq ~ Gender + Admit, DF)
```

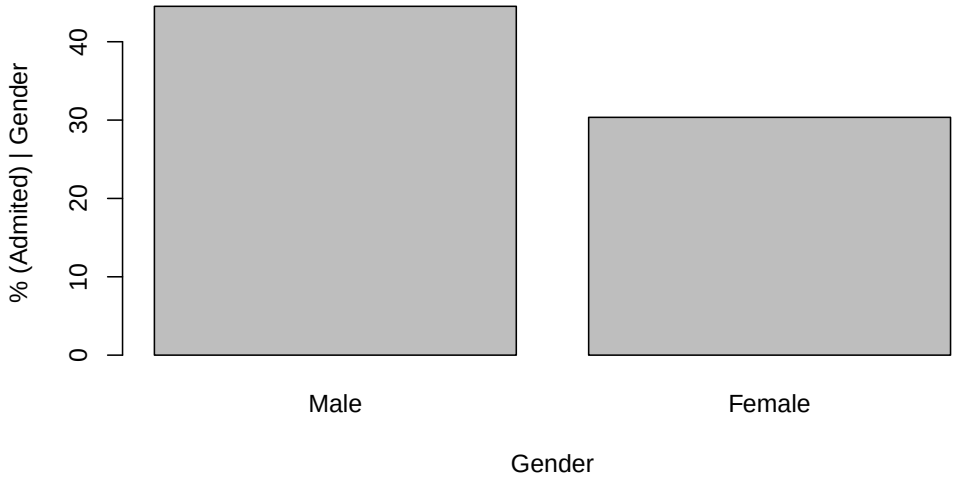
```
        Admit
Gender   Rejected Admitted
  Male       1493     1198
  Female     1278      557
```

Et voici ce qui est recherché

```
tbp <- tbm |>
  proportions("Gender") |>
  transform(Prc = round(100 * Freq, 2), Freq = NULL) |>
  subset(Admit == "Admitted") |>
  transform(Admit = NULL) |>
  print()
```

```
  Gender   Prc
3   Male 44.52
4 Female 30.35
```

```
tbp |> barplot(Prc ~ Gender, ylab = "% (Admited) | Gender", data = _)
#or tbm |> spineplot()
```



Dans l'ensemble, les filles ont un taux d'admission de 30.35%, ce qui est bien inférieur à celui des garçons, qui est de 44.52%. Cela suggère qu'il y a peut-être une discrimination à l'égard des femmes.

(b)

Au vu de ces données (tous départements confondus), peut-on affirmer que les femmes font l'objet d'une discrimination significative ? Pour répondre à cette question, vous devrez réaliser *un test LR sur un rapport de cotes*. Formulez vos hypothèses H_0 et H_1, et décrivez clairement votre démarche. Proposez une autre approche, toujours basée sur un test LR de rapport de cotes, en utilisant cette fois une modélisation différente. Parvenez-vous à la même conclusion ? Commentez.

Réponse

Soit $p_F = P(Admitted|Female)$ et $p_M = P(Admitted|Male)$. Il faut tester

$$H_0 : p_F = p_M \text{ vs } p_F \neq p_M$$

Soit $o_F = p_F/(1-p_F)$ et $o_M = p_M/(1-p_M)$ les cotes correspondantes. Tester H_0 est équivalent à tester que le rapport des cotes $or = o_F/o_M$ est égal à 1 ou que $\log(or) = 0$. Voici deux façons de tester cette hypothèse à travers un test LR, la première basée sur un modèle de Poisson et la seconde sur un modèle logistique. Les deux approches sont équivalentes.

```
glm(Freq ~ Gender * Admit, family = poisson, data = DF) |> drop1(test = "LR")
```

```
Single term deletions

Model:
Freq ~ Gender * Admit
             Df Deviance    AIC    LRT  Pr(>Chi)
<none>            2163.7 2330.8
Gender:Admit  1   2257.2 2422.2 93.449 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
DF2 <- DF |>
  tidyr::pivot_wider(names_from = Admit, values_from = Freq) |>
  transform(Applicant = Rejected + Admitted)

glm(Admitted / Applicant ~ Gender, family = binomial,
    weights = Applicant, data = DF2) |> drop1(test = "LR")
```

```
Single term deletions

Model:
Admitted/Applicant ~ Gender
       Df Deviance    AIC    LRT  Pr(>Chi)
<none>      783.61 856.55
Gender  1   877.06 948.00 93.449 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Les deux méthodes donnent exactement les mêmes $p-$valeurs et donc la même conclusion, à savoir une discrimination significative (à 5%).

(c)

Répétez la question (a), mais cette fois-ci par département. Peut-on parler d'une *discrimination significative*, fondée sur le genre, au sein de l'un ou l'autre des départements considérés ? Justifiez. Comparez et mettez en contraste les deux analyse (celle du point (a) et la présente).

Réponse

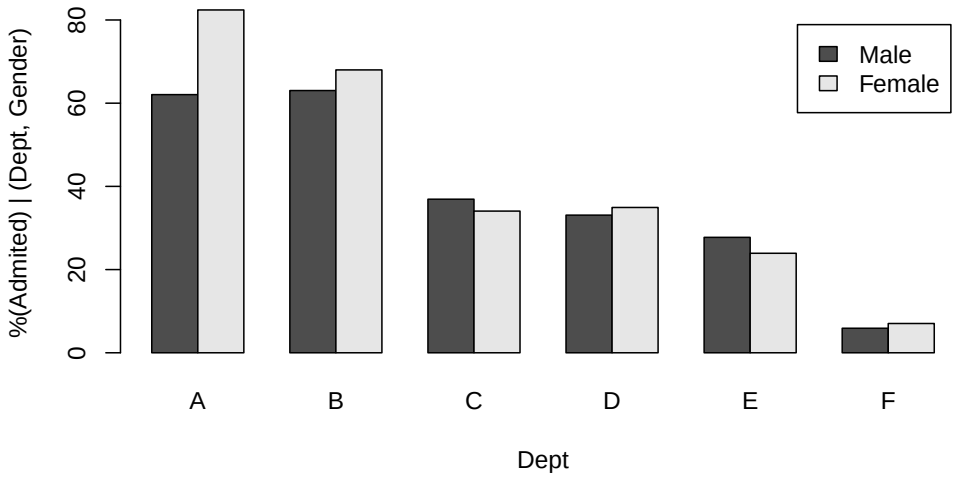
Voici les proportions d'admissions par sexe et par département, ainsi qu'une représentation graphique de ces chiffres.

```r
tbj <- TB |>
  proportions(c("Gender", "Dept")) |>
  as.data.frame() |>
  transform(Prc = round(Freq * 100, 2), Freq = NULL) |>
  subset(Admit == "Admitted") |>
  transform(Admit = NULL)

tbj |> tidyr::pivot_wider(names_from = Gender, values_from = Prc) |> print()
```

```
# A tibble: 6 x 3
  Dept   Male Female
  <fct> <dbl>  <dbl>
1 A      62.1  82.4
2 B      63.0  68
3 C      36.9  34.1
4 D      33.1  34.9
5 E      27.8  23.9
6 F       5.9   7.04
```

```r
tbj |> barplot(Prc ~ Gender + Dept, beside = TRUE,
               legend.text = TRUE, ylab = "%(Admited) | (Dept, Gender)", data = _)
#or TB |> vcd::doubledecker(Admit ~ Dept + Gender, data = _)
```



À l'exception des départements C et E, les proportions de filles admises sont plus élevées que celles des garçons. L'écart le plus important est observé dans le département A, où les filles ont un taux d'admission de 82.41%, ce qui est bien supérieur à celui des garçons, qui est de 62.06%.

C'est le seul département où l'écart entre les garçons et les filles en termes d'admission peut être qualifié de significatif. En effet,

```
TB |> vcd::loddsratio() |> confint()
```

```
       2.5 %    97.5 %
A  0.5371775 1.5669744
B -0.6376433 1.0776883
C -0.4070436 0.1572003
D -0.2124158 0.3763902
E -0.5926552 0.1922812
F -0.4092137 0.7870054
```

Il apparaît donc que la quasi totalité des départements ne présente pas de discrimination fondée sur le genre. La seule exception est le département A, où l'on observe une discrimination en faveur des femmes. Cela semble contredire l'analyse du point (a).

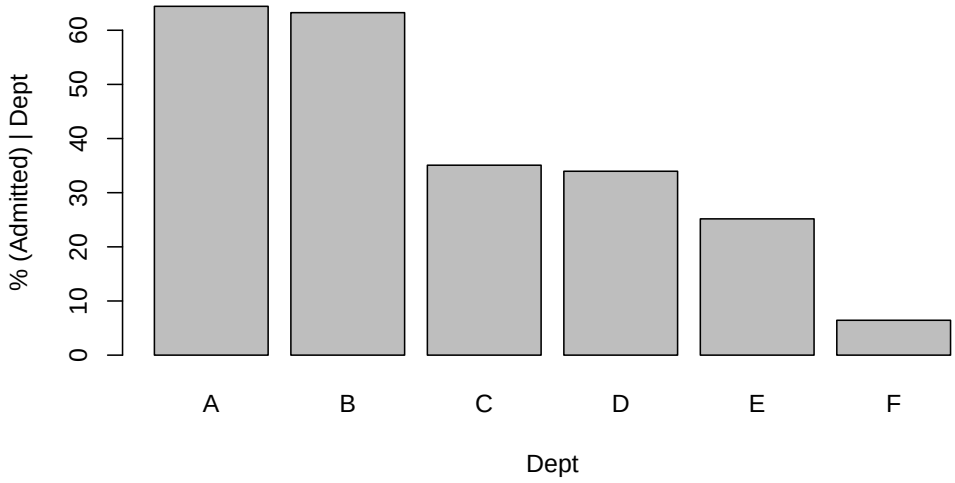
(d)

Ces données relèvent du paradoxe de Simpson. Expliquez brièvement sa manifestation ici. Identifiez la cause de ce paradoxe, càd un (ou des) élément(s) dans les données qui l'explique(nt) clairement. Pour répondre à cette question, vous devez fournir un ou plusieurs graphiques, accompagnés d'une argumentation.

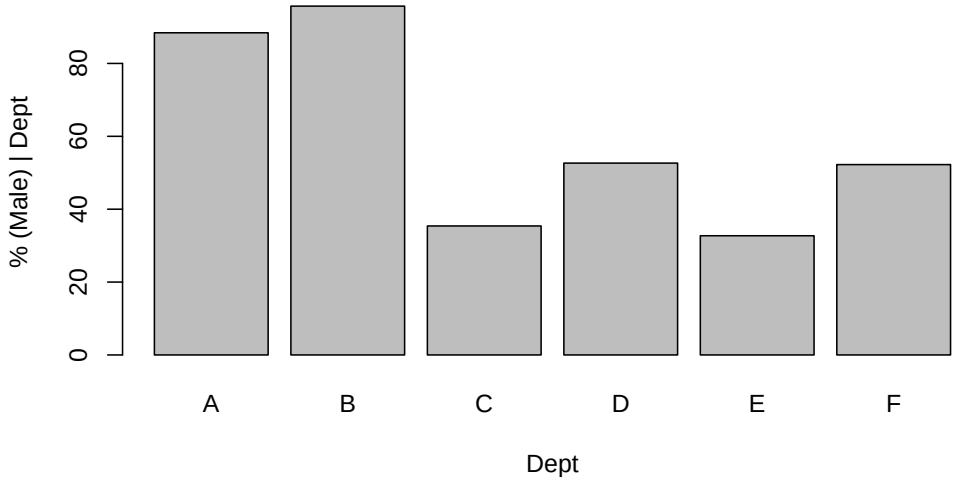
Réponse

La contradiction relevée dans la question précédente est la manifestation du paradoxe de Simpson. Pour l'expliquer il faut examiner l'association entre *Dept* et les deux autres variables (*Gender* et *Admit*).

```
xtabs(Freq ~ Dept + Admit, DF) |>
  proportions("Dept") |>
  as.data.frame() |>
  subset(Admit == "Admitted") |>
  transform(Admit = NULL, Prc = round(Freq * 100, 2), Freq = NULL) |>
  barplot(data = _, Prc ~ Dept, ylab = "% (Admitted) | Dept")
```




```
xtabs(Freq ~ Dept + Gender, DF) |>
  proportions("Dept") |>
  as.data.frame() |>
  subset(Gender == "Male") |>
  transform(Gender = NULL, Prc = round(Freq * 100, 2), Freq = NULL) |>
  barplot(data = _, Prc ~ Dept, ylab = "% (Male) | Dept")
```



Le premier graphique montre que les départements A et B sont de loin les plus faciles d'accès. En effet, le taux d'admission dans ces départements est d'environ 64%, alors que dans les autres il ne dépasse pas 35%.
Le deuxième graphique montre surtout que les départements A et B reçoivent une majorité écrasante de candidats masculins (88.42% pour A, et 95.73% pour B).

En résumé, les filles ont tendance à postuler dans des départements compétitifs avec des taux d'admission (très) faibles, tandis que les graçons ont tendance à postuler dans des départements moins compétitifs avec des taux d'admission (très) élevés.

C'est ce qui conduit au pardoxe constaté lorsque le facteur département est ignoré.

(e)

En prenant "Male" et "A" comme catégories de référence (pour *Gender* et *Dept*, respectivement), ajustez un modèle de régression logistique pour modéliser la probabilité d'admission en fonction du genre et du département. Peut-on conclure à une association homogène entre les trois variables considérées ? Si nous nous concentrons sur la ligne `GenderFemale:DeptB` dans `summary()`, comment pouvez-vous interpréter précisément les valeurs figurant dans les colonnes `Estimate` et `Pr(>|z|)` ?

Réponse

```
mlg <- glm(Admitted / Applicant ~ Gender * Dept,
           family = binomial, weights = Applicant, data = DF2)
drop1(mlg, test = "LR")
```

```
Single term deletions

Model:
Admitted/Applicant ~ Gender * Dept
            Df Deviance    AIC    LRT Pr(>Chi)
<none>           0.000  92.94
Gender:Dept  5   20.204 103.14 20.204 0.001144 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

L'association homogène est rejetée (à 5%).

```
mlg |> summary() |> coef()
```

```
                    Estimate Std. Error     z value     Pr(>|z|)
(Intercept)         0.49212143 0.07174966   6.8588682 6.940825e-12
GenderFemale        1.05207596 0.26270810   4.0047336 6.208742e-05
DeptB               0.04162783 0.11318919   0.3677721 7.130431e-01
DeptC              -1.02763967 0.13549685  -7.5842331 3.344593e-14
DeptD              -1.19607953 0.12640656  -9.4621632 3.016374e-21
```

```
DeptE               -1.44908321 0.17681152  -8.1956378 2.492678e-16
DeptF               -3.26186520 0.23119594 -14.1086615 3.358928e-45
GenderFemale:DeptB -0.83205342 0.51039480  -1.6302153 1.030560e-01
GenderFemale:DeptC -1.17699758 0.29955796  -3.9291147 8.525915e-05
GenderFemale:DeptD -0.97008876 0.30261874  -3.2056467 1.347593e-03
GenderFemale:DeptE -1.25226298 0.33032201  -3.7910371 1.500195e-04
GenderFemale:DeptF -0.86318013 0.40266653  -2.1436600 3.206014e-02
```

Dans le ligne `GenderFemale:DeptB`, la valeur dans `Estimate` (`-0.83205342`) n'est rien d'autre que

$$log\frac{\widehat{or}^{Admit\times Gender|Dept=B}}{\widehat{or}^{Admit\times Gender|Dept=A}}$$

C'est ce que nous pouvons vérifier à l'aide de la sortie suivante

```
TB |> vcd::loddsratio()
```

```
log odds ratios for Gender and Admit by Dept

          A           B           C           D           E           F
 1.05207596  0.22002254 -0.12492163  0.08198719 -0.20018702  0.18889583
```

En effet, $0.220 - 1.052 = -0.832$. Il s'ensuit que la valeur indiquée dans `Pr(>|z|)` n'est rien d'autre que la $p-$valeur du test de Wald pour

$$H_0 : \ or^{Admit\times Gender|Dept=B} = or^{Admit\times Gender|Dept=A}.$$

Excercice 3

Le fichier lung_cancer_data contient des données sur le nombre de cas de cancer du poumon (variable `cases`) dans quatre villes du Danemark (Fredericia, Horsens, Kolding, Vejle) et pour différentes catégories d'âge ($40-54$, $55-59$, $60-64$, $65-69$, $70-74$, ≥ 75). La taille de la population de chaque groupe d'âge de chaque ville est rapportée dans la variable `city`. Et pour chaque tranche d'âge, `age_midpt` donne le point central, sauf pour la dernière tranche où 75 est utilisé.

Pour charger des données et les enregistrer dans l'objet `lc`, vous pouvez saisir la commande suivante.

```r
lc <- read.table("Data/lung_cancer_data.txt")
```

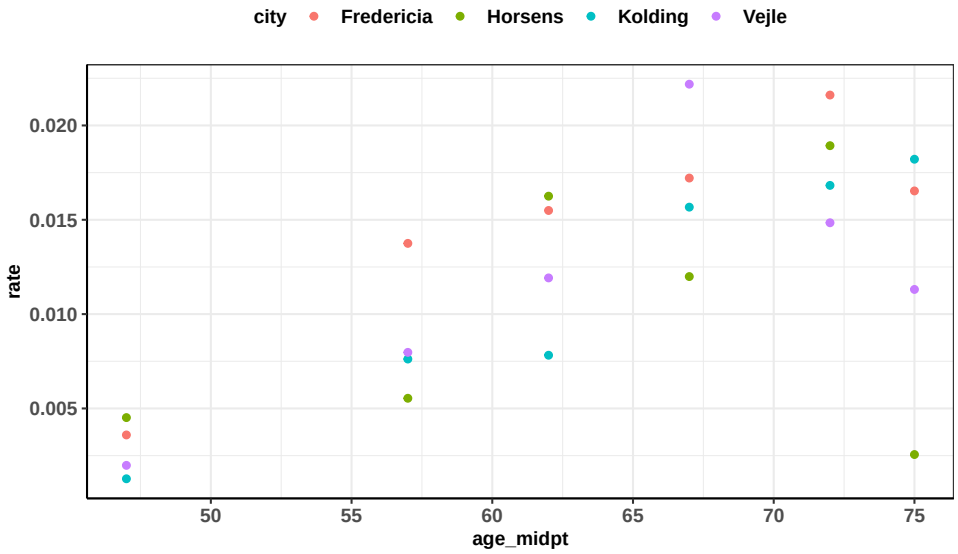
(a)

Faites un (seul) graphique montrant l'évolution du taux de cancer (*cas/population*) en fonction de l'âge. Utilisez une couleur différente pour chaque ville.

Réponse

```r
lc <- lc |> transform(rate = cases/population, age = factor(age), city = factor(city))

ggplot(data = lc, aes(x = age_midpt, y = rate)) +
  geom_point(aes(color = city))
```



(b)

Le modèle suivant est proposé pour expliquer le nombre de cas de cancer en fonction des variables disponibles. Ce modèle présente une déficience qui doit être corrigée. Expliquez cette déficience et proposez un remède qui prenne mieux compte les données disponibles *tout en utilisant la régression de Poisson.*

```r
glm(cases ~ city * age_midpt, data = lc, family = poisson)
```

Réponse

Dans le modèle ci-dessus, la taille de chaque sous-population n'est pas prise en compte, ce qui rend les nombres de cas incomparables. Ceci peut être corrigé en intégrant $\log(population)$, dans le modèle, comme `offset`.

(c)

Considérons le modèle de Poisson (corrigé), cette fois avec les variables explicatives *city*, *age_midpt*, l'interaction entre les deux et age_midpt^2. Simplifiez ce modèle en ne conservant que les termes *significatifs à* 1%. En précisant vos notations, écrivez l'équation mathématique de votre modèle final. Diagnostiquez sa qualité d'ajustement et concluez.

Réponse

```
md <- glm(cases ~ city * age_midpt + I(age_midpt^2) + offset(log(population)),
          data = lc, family = poisson)

drop1(md, test = "LR")
```

```
Single term deletions

Model:
cases ~ city * age_midpt + I(age_midpt^2) + offset(log(population))
               Df Deviance    AIC    LRT  Pr(>Chi)
<none>              16.005 130.39
I(age_midpt^2)  1   38.339 150.73 22.334 2.292e-06 ***
city:age_midpt  3   26.017 134.41 10.012   0.01846 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
md <- glm(cases ~ city + age_midpt + I(age_midpt^2) + offset(log(population)),
          data = lc, family = poisson)

drop1(md, test = "LR")
```

```
Single term deletions

Model:
cases ~ city + age_midpt + I(age_midpt^2) + offset(log(population))
               Df Deviance    AIC    LRT  Pr(>Chi)
```

```
<none>              26.017 134.41
city            3   30.966 133.35  4.949    0.1756
age_midpt       1   51.829 158.22 25.811 3.765e-07 ***
I(age_midpt^2)  1   46.448 152.84 20.431 6.182e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
md <- glm(cases ~ age_midpt + I(age_midpt^2) + offset(log(population)),
          data = lc, family = poisson)

drop1(md, test = "LR")
```

```
Single term deletions

Model:
cases ~ age_midpt + I(age_midpt^2) + offset(log(population))
               Df Deviance    AIC    LRT  Pr(>Chi)
<none>              30.966 133.35
age_midpt       1   55.971 156.36 25.005 5.720e-07 ***
I(age_midpt^2)  1   50.698 151.09 19.732 8.911e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(md) |> coef()
```

```
                    Estimate   Std. Error   z value     Pr(>|z|)
(Intercept)    -21.434384043 3.0887571330 -6.939485 3.935317e-12
age_midpt        0.501159509 0.1020193966  4.912394 8.997086e-07
I(age_midpt^2)  -0.003633982 0.0008289412 -4.383884 1.165821e-05
```

L'équation du modèle est la suivante

$$\hat{\mu}_i = population_i \times \exp(-21.43 + 0.501 age\_midpt_i - 0.004 age\_midpt_i^2),$$

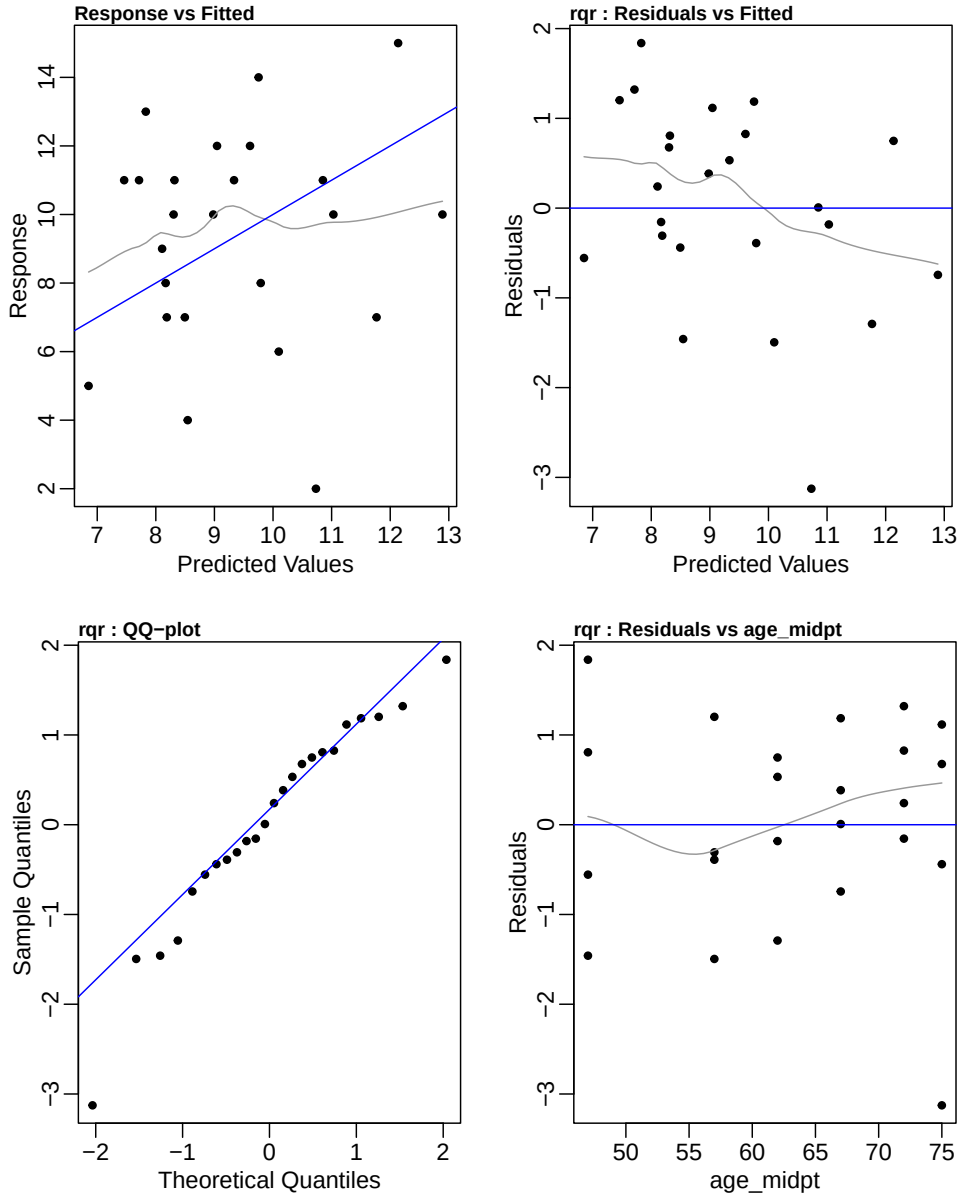
où $\hat{\mu}_i$ est le nombre de cas attendu pour $(population_i, age\_midpt_i, city_i)$ donnés.

Nous pouvons effectuer un "diagnostique" à l'aide des fonction `pr2()` et `diagnosis()`.

```r
pr2(md)
```

```
[1] 76.16279
```

```
diagnost(md, plots = c("response", "fitted", "qqplot", "age_midpt"))
```



Dans l'ensemble, le modèle semble satisfaisant (à l'exception d'une valeur extrême).

(d)

Refaites l'analyse précédente mais cette fois-ci en utilisant une régression logistique. Comparez les deux modèles (Poisson et logistique).

Réponse

```
mg <- glm(cases / population ~ city * age_midpt + I(age_midpt^2),
          data = lc, family = binomial, weights = population)

drop1(mg, test = "LR")
```

```
Single term deletions

Model:
cases/population ~ city * age_midpt + I(age_midpt^2)
               Df Deviance    AIC    LRT  Pr(>Chi)
<none>              16.191 130.29
I(age_midpt^2)  1   38.570 150.67 22.379 2.239e-06 ***
city:age_midpt  3   26.273 134.37 10.081   0.01789 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mg <- glm(cases / population ~ city + age_midpt + I(age_midpt^2),
          data = lc, family = binomial, weights = population)

drop1(mg, test = "LR")
```

```
Single term deletions

Model:
cases/population ~ city + age_midpt + I(age_midpt^2)
               Df Deviance    AIC     LRT  Pr(>Chi)
<none>              26.273 134.37
city            3   31.284 133.38  5.0116    0.1709
age_midpt       1   52.166 158.27 25.8931 3.608e-07 ***
I(age_midpt^2)  1   46.751 152.85 20.4788 6.030e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mg <- glm(cases / population ~ age_midpt + I(age_midpt^2),
          data = lc, family = binomial, weights = population)

drop1(mg, test = "LR")
```

```
Single term deletions

Model:
cases/population ~ age_midpt + I(age_midpt^2)
               Df Deviance    AIC    LRT  Pr(>Chi)
<none>              31.284 133.38
age_midpt       1   56.361 156.46 25.076 5.510e-07 ***
I(age_midpt^2)  1   51.056 151.16 19.772 8.726e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(mg) |> coef()
```

```
                    Estimate   Std. Error   z value     Pr(>|z|)
(Intercept)    -21.531007906 3.1021237896 -6.940731 3.900751e-12
age_midpt        0.504307109 0.1024948689  4.920316 8.640475e-07
I(age_midpt^2)  -0.003656187 0.0008330136 -4.389108 1.138165e-05
```

L'équation du modèle est la suivante

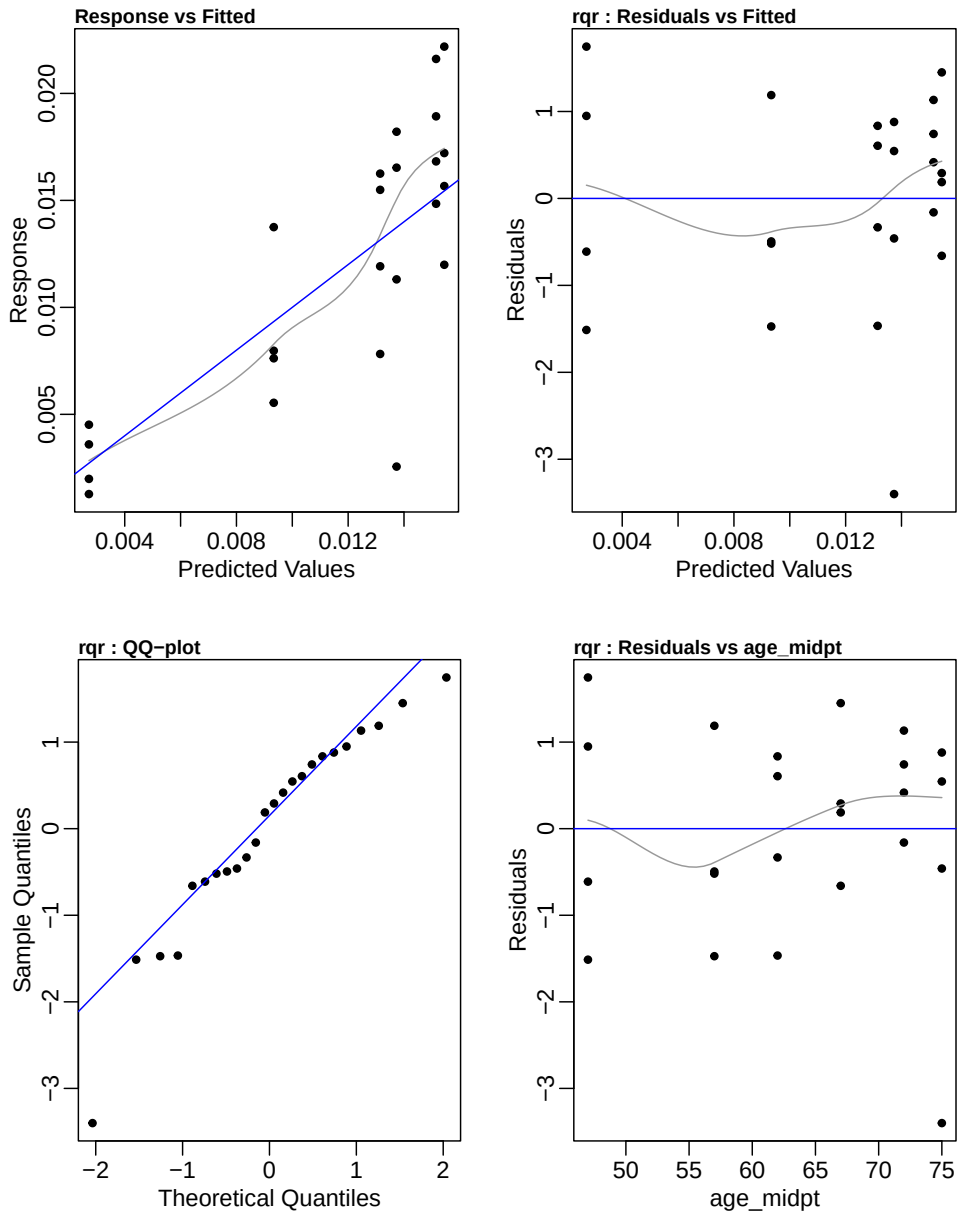
$$\widehat{p}(case_i|age\_midpt_i, city_i) = \frac{1}{1+\exp(21.531 - 0.504age\_midpt_i + 0.004age\_midpt_i^2)},$$

où $\widehat{p}(case|age\_midpt, city)$ est la probabilité d'être un cas (càd avoir le cancer) sachant $(age\_midpt, city)$.

```
pr2(mg)
```

```
[1] 76.11869
```

```
diagnost(mg, plots = c("response", "fitted", "qqplot", "age_midpt"))
```

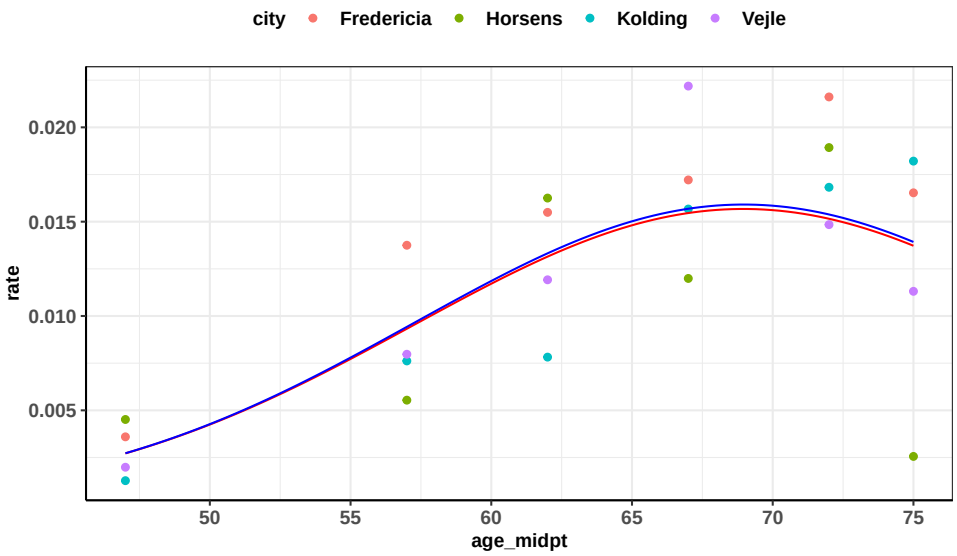


Dans l'ensemble, le modèle semble satisfaisant.

Les deux modèles présentent des performances presque identiques et il serait très difficile de prétendre que l'un est meilleur que l'autre, comme l'attestent, à titre d'exemple, les résultats suivants.

```
ggplot(data = lc, aes(x = age_midpt, y = rate)) +
  geom_point(aes(color = city)) +
```

```
  stat_function(fun = \(x) exp(coef(md)[1] + coef(md)[2] * x +
                                 coef(md)[3] * x^2), color = "red") +
  stat_function(fun = \(x) 1 / exp(-(coef(mg)[1] + coef(mg)[2] * x +
                                       coef(mg)[3] * x^2)), color = "blue")
```




```
resid(md, type = "pearson") |> abs() |> mean()
```

```
[1] 0.8878487
```

```
resid(mg, type = "pearson") |> abs() |> mean()
```

```
[1] 0.8932687
```