LSTAT2100 - Exercices - Série 2

Solutions

Exercice 1

Dans cet exercice, nous allons reprendre le premier exercice de la Série 1, et le réaliser ***en utilisant uniquement des modèles log-linéaires***.

Nous nous intéressons à la couleur des yeux d'une certaine population. Pour $n = 300$ individus, pris au hasard, nous avons observé les chiffres suivants.

bleu	marron	noir	vert
48	122	95	35

(a) Ces couleurs sont-elles toutes réparties de manière uniforme (équiprobables) ? Répondez à l'aide d'un test de Pearson et un test LR (les deux à réaliser via une régression log-linéaire).

Solution:

```
Couleur <- factor(c("bleu", "vert", "marron", "noir"))
Freq <- c(48, 35, 122, 95)
dt <- data.frame(Couleur, Freq)

dt$Couleur <- relevel(dt$Couleur, "bleu")   # Ceci n'est pas obligatoire.
fit <- glm(Freq ~ Couleur, data = dt, family = poisson())

# test de Pearson
fit |> drop1(test = "Rao")  #i.e. xtabs(Freq ~ Couleur, data = dt) |> chisq.test()
```

```
## Single term deletions
##
## Model:
## Freq ~ Couleur
##         Df Deviance  AIC Rao score Pr(>Chi)
## <none>           0.0 32.1
## Couleur  3     67.4 93.6      65.8  3.3e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# test LR
fit |> drop1(test = "LRT")
```

```
## Single term deletions
##
## Model:
## Freq ~ Couleur
##         Df Deviance  AIC  LRT Pr(>Chi)
## <none>           0.0 32.1
```

```
## Couleur  3     67.4 93.6 67.4  1.5e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(b) Peut-on dire que les yeux foncés (marron et noir) sont deux fois plus probables que les yeux clairs (bleu et vert) ? Réaliser le test via une régression log-linéaire.

Solution:

Pour répondre à cette question, nous allons commencer par créer la variable binaire `clair` qui prend la valeur `T` (`TRUE`) pour les yeux bleus ou verts, et `F` (`FALSE`) pour marron ou noir.

```
dt$clair <- dt$Couleur %in% c("bleu", "vert")
dt
```

Couleur	Freq	clair
bleu	48	TRUE
vert	35	TRUE
marron	122	FALSE
noir	95	FALSE

Soit $\mu(clair) = E(Freq|clair) = n \times Prob(clair)$. Et soit le modèle de Poisson

```
fitb <- glm(Freq ~ clair, data = dt, family = poisson)
```

L'équation mathématique de ce modèle peut s'écrire comme

$$\log \mu(clair) = \beta_0 + \beta_1 I(clair = T)$$

Nous avons donc que $\mu(T) = \exp(\beta_0 + \beta_1)$, pour les yeux clairs, et $\mu(F) = \exp(\beta_0)$, pour les yeux foncés.

Nous cherchons à tester si $\mu(F) = 2\mu(T) \iff \exp(\beta_1) = 0.5 \iff \beta_1 = \log(0.5)$.

Il s'agit donc d'effectuer le test suivant.

$$H_0 : \beta_1 = \log(0.5) \text{ vs } H_1 : \beta_1 \neq \log(0.5)$$

Il existe différentes méthodes pour réaliser ce test (dans R).

- Utiliser un intervalle de confiance

  ```
  IC <- confint(fitb, parm = "clairTRUE") |> print()
  ```

  ```
  ##  2.5 % 97.5 %
  ## -1.219 -0.713
  ```

  Comme $log(0.5) = -0.693 \notin IC$, nous rejetons l'hypothèse nulle, au niveau 5%, et nous pouvons dès lors affirmer que les yeux foncés ne sont pas deux fois plus probables que les yeux clairs.

- Via un `offset()`

  ```
  fitc <- glm(Freq ~ offset(log(0.5)*clair), data = dt, family = poisson)
  ```

  ```
  anova(fitc, fitb, test = "LRT")$"Pr(>Chi)"[2]
  ```

  ```
  ## [1] 0.0344
  ```

- Utiliser la fonction `car::linearHypothesis()`; voir la fiche d'aide pour plus d'informations concernant cette fonction *fort utile*.

```
car::linearHypothesis(fitb, "clairTRUE = -0.693")
```

Res.Df	Df	Chisq	Pr(>Chisq)
3	NA	NA	NA
2	1	4.31	0.038

(c) Les proportions des yeux bleus et des yeux verts sont-elles les mêmes ? Réaliser le test via une régression log-linéaire.

Solution:

Dans la question (a), nous avons ajusté le modèle `fit` au données. Le `summary` de `fit` est

```
summary(fit) |> coef()
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.871	0.144	26.82	0.000
Couleurmarron	0.933	0.170	5.47	0.000
Couleurnoir	0.683	0.177	3.85	0.000
Couleurvert	-0.316	0.222	-1.42	0.155

Son équation mathématique est (noter que "blue" est notre niveau de référence)

$$\log \mu(Couleur) = \beta_0 + \beta_1 I(Couleur = "marron") + \beta_2 I(Couleur = "noir") + \beta_3 I(Couleur = "vert")$$

Ainsi $\log \mu("blue") = \beta_0$ et $\log \mu("vert") = \beta_0 + \beta_3$. Pour répondre à la question, il suffit donc de tester si $\beta_3 = 0$. Pour cela, il suffit de lire la $p-$valeur ("Pr(>|z|)") qui figure à la ligne "Couleurvert" du `summary()`.

Nous ne rejetons donc pas l'hypothèse que les deux couleurs sont équiprobables.

Remarque: Si le niveau de référence n'est ni "bleu" ni "vert", il convient de modifier celui-ci (à l'aide de la fonction `relevel()`) afin de pouvoir répondre facilement à cette question.

Exercice 2

Dans cet exercice, nous utiliserons le jeu de données mdata.csv, que nous avions utilisé dans la Série 1. Vous devez charger/transformer les données comme expliqué dans le texte. Aussi, nous vous renvoyons à l' Exercice 3 de la Série 1 pour la description des variables.

(a) Ajustez un modèle log-linéaire *saturé* pour modéliser la table de contingence $tel \times repay$. Construisez des intervalles de confiance à 98% pour tous les paramètres du modèle, à l'exception de l'intercept.

Solution:

```
tab <- xtabs(~ tel + repay, data = mdata)
dt <- data.frame(tab) |> print()
```

```
##   tel        repay Freq
## 1  No      Default  142
## 2 Yes      Default   92
## 3  No NotInDefault  305
## 4 Yes NotInDefault  211
```

Nous pouvons à présent utiliser la fonction `glm()`.

```
fit <- glm(Freq ~ tel * repay, data = dt, family = poisson)
confint(fit, parrm = (coef(fit) |> names())[-1], level = 0.98)
```

	1 %	99 %
(Intercept)	4.754	5.145
telYes	-0.749	-0.126
repayNotInDefault	0.531	1.005
telYes:repayNotInDefault	-0.307	0.443

(b) Le tableau suivant donne le `summary()` du modèle saturé dont il est question en (a).

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	4.956	0.084	59.056	0.000
telYes	-0.434	0.134	-3.243	0.001
repayNotInDefault	0.764	0.102	7.525	0.000
telYes:repayNotInDefault	0.066	0.161	0.407	0.684

Utilisez cette sortie (*et uniquement celle-ci*) pour reconstruire le tableau de contingence *tel* × *repay*.

tel/repay	Default	NotInDefault
No	142	305
Yes	92	211

Que pouvez-vous dire des résidus du modèle étudié ? de sa déviance ? et du $pseudo - R^2$? Peut-on les obtenir sans faire de calculs ?

Solution:

Comme il s'agit d'un modèle saturé (4 paramètres pour 4 observations), il 'colle' parfaitement aux données: les valeurs prédites sont donc identiques aux valeurs observées. Dès lors, sur base de l'équation mathématique du modèle, il est facile de construire le tableau de contingence à partir des valeurs prédites, comme suit

tel/Y	Defaut	PasDefaut
Non	`exp(coef(fit)[1])`= 142	`exp(coef(fit)[1]+coef(fit)[3])`= 305
Oui	`exp(coef(fit)[1]+coef(fit)[2])`= 92	`exp(sum(coef(fit)))`= 211

Le plus simple (mais que nous ne pouvons pas faire ici) serait d'utiliser la fonction `fitted()`:

```
fitted(fit)
```

```
##   1   2   3   4
## 142  92 305 211
```

Un modèle saturé aboutit, par construction, à une déviance nulle (R calcule cette dernière à l'aide d'un algorithme numérique générique, c'est pourquoi on n'obtient pas exactement 0.)

```
deviance(fit)
```

```
## [1] -8.66e-15
```

Les résidus sont nuls aussi

```
resid(fit)
```

```
## [1] 0 0 0 0
```

Et le $pseudo - R^2$ est de 100%

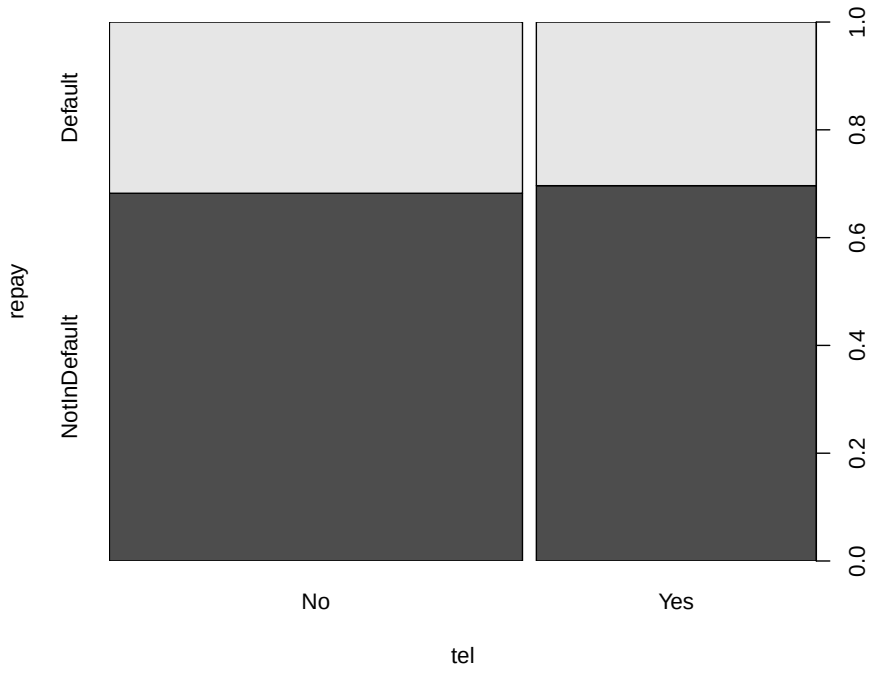
```
pr2(fit)
```

```
## [1] 100
```

(c) Visualisez graphiquement des proportions conditionnelles $P(repay|tel)$. Que suggère votre graphique quand à l'association entre `repay` et `tel` ? Expliquez.

Solution:

```
spineplot(tab)
```



Nous voyons un alignement presque parfait entre la partie gauche et droite de ce graphique. Cela illustrant le fait que la distribution de `repay` ne dépend pas de `tel`.

(d) Utilisez le modèle log-linéaire ajusté précédemment pour tester l'indépendance entre `repay` et `tel`. Proposer deux approches différentes (basées sur les GLM), dont une est le résultat direct des analyses précédentes (sans autres calculs).

Solution:

Nous remarquons que l'intervalle de confiance pour le terme d'interaction (que nous avons calculer en (a)) contient 0. Dès lors, nous ne pouvons pas rejeter le fait que ce terme d'interaction est égal à 0. Et donc l'indépendance n'est pas à rejeter. Nous pouvons également voir cela en inspectant le `summary(fit)` et en y observant la $p-$valeur d'interaction qui est de 0.684, beaucoup trop grande donc pour rejeter l'indépendance.

Une autre façon de répondre à cette question est d'effectuer le test de Pearson (par exemple):

```
drop1(fit, test = "Rao")
```

```
## Single term deletions
##
## Model:
## Freq ~ tel * repay
##           Df Deviance  AIC Rao score Pr(>Chi)
## <none>          0.000 35.9
## tel:repay  1    0.166 34.1     0.166     0.68
```

(e) Créez la nouvelle variable `Acc` qui regroupe les modalités "No acc" et "<0" de la variable `account` en une seule modalité nommée "Noacc or <0" et les modalités "[0-200)" et ">=200" en une seule modalité nommée ">=0". Réalisez la table de contingence `employm`×`Acc`.

Construisez un modèle log-linéaire saturé sur ces données en choisissant "Noacc or <0" comme référence. Au regard des $p-$valeurs des termes d'interactions, que pouvez-vous conclure quand à l'association entre `employm` et `Acc` (ici, il n'est pas demandé d'effectuer de calculs supplémentaires) ?

Que signifie la $p-$valeur qui figure dans la ligne `Acc>=0` du `summary()` de votre modèle.

Utilisez le modèle ajusté pour tester l'indépendance entre `employm` et `Acc`.

Solution:

Préparons d'abord les données et définissons les niveaux de références.

```
mdata$Acc <- mdata$account
levels(mdata$Acc) <- list("Noacc or <0" = c("No acc", "<0"), ">=0" = c("[0-200)", ">=200"))
mdata$Acc <- relevel(mdata$Acc, "Noacc or <0") #Ceci est facultatif
mdata$employm <- relevel(mdata$employm, "No or <1")  #Ceci aussi est facultatif
tab <- xtabs(~ Acc + employm, data = mdata)
tab
```

Acc/employm	No or <1	[1-4)	[4-7)	>=7
Noacc or <0	105	175	84	147
>=0	71	81	37	50

Ajustons le modèle log-linéaire saturé.

```
data <- data.frame(tab)
fit <- glm(Freq ~ Acc * employm, data = data, family = poisson)
summary(fit) |> coef()
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	4.654	0.098	47.69	0.000
Acc>=0	-0.391	0.154	-2.55	0.011
employm[1-4)	0.511	0.123	4.14	0.000
employm[4-7)	-0.223	0.146	-1.52	0.127
employm>=7	0.336	0.128	2.63	0.008
Acc>=0:employm[1-4)	-0.379	0.204	-1.86	0.063
Acc>=0:employm[4-7)	-0.429	0.250	-1.71	0.087
Acc>=0:employm>=7	-0.687	0.225	-3.06	0.002

Nous pouvons voir que deux termes d'interaction apparaissent comme non significativement différents de 0 et qu'un (le dernier) apparaît comme significatif. Cela ne nous permet pas de conclure sur l'association entre les deux variables étudiées, puisqu'il s'agit ici de tests individuels (isolés) et qu'il faudrait, à la place, un test global de la nullité simultanée de tous les termes d'interaction (tout en contrôlant l'erreur de type I).

Si on note par N la variable aléatoire qui représente les fréquences absolues ($Freq$). La $p-$valeur qui figure dans la ligne "Acc>=0" correspond alors au test suivant.

$$H_0 : E(N|Acc = ">=0", employm = "No or <1") = E(N|Acc = "Noacc or <0", employm = "No or <1")$$

ou, en termes de probabilité,

$$H_0 : Prob(Acc = ">=0"|employm = "No or <1") = Prob(Acc = "Noacc or <0"|employm = "No or <1")$$

Testons l'indépendance.

```
drop1(fit, test = "Rao")
```

```
## Single term deletions
##
## Model:
## Freq ~ Acc * employm
##             Df Deviance  AIC Rao score Pr(>Chi)
## <none>            0.00 66.1
## Acc:employm  3     9.67 69.8      9.74    0.021 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Au niveau 5%, nous rejetons l'indépendance entre les deux variables.

(f) Indépendamment du résultat du test précédent, ajustez un modèle log-linéaire considérant les deux variables `emploi` et `Acc` comme indépendantes. Que signifie la $p-$valeur dans la ligne `Acc>=0` du `summary()` de ce nouveau modèle ? Calculez le $pseudo - R^2$ et comparez-le à celui du modèle saturé.

Solution:

Voici le modèle d'indépendance

```
fit0 <- glm(Freq ~ Acc + employm, data = data, family = poisson)
summary(fit0) |> coef()
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	4.787	0.079	60.28	0.000
Acc>=0	-0.760	0.078	-9.70	0.000
employm[1-4)	0.375	0.098	3.83	0.000
employm[4-7)	-0.375	0.118	-3.17	0.002
employm>=7	0.113	0.104	1.09	0.277

En suivant la même notation que celle utilisée ci-dessus, la $p-$valeur qui figure dans la ligne "Acc>=0" correspond alors au test suivant.

$$H_0 : E(N|Acc = ">=0") = E(N|Acc = "Noacc or <0")$$

ou, en terme de probabilité,

$$H_0 : Prob(Acc = ">=0") = Prob(Acc = "Noacc or <0")$$

Ce test est réalisé ici en supposant l'indépendance entre les variables `employm` et `Acc` et donc en utilisant toutes les observations, quelle que soit leur valeur `employm`.

Voici la $pseudo - R^2$ de `fit0`

```
pr2(fit0)
```

```
## [1] 94
```

alors que celle de `fit` (modèle saturé) est de 100%

(g) Faites un tableau de contingence et un graphique (approprié) mettant en jeu les variables `repay`, `tel` et `employm`. Une association homogène entre ces trois variables est-elle envisageable ? À ce stade, il n'est pas demandé de réaliser un quelconque test.

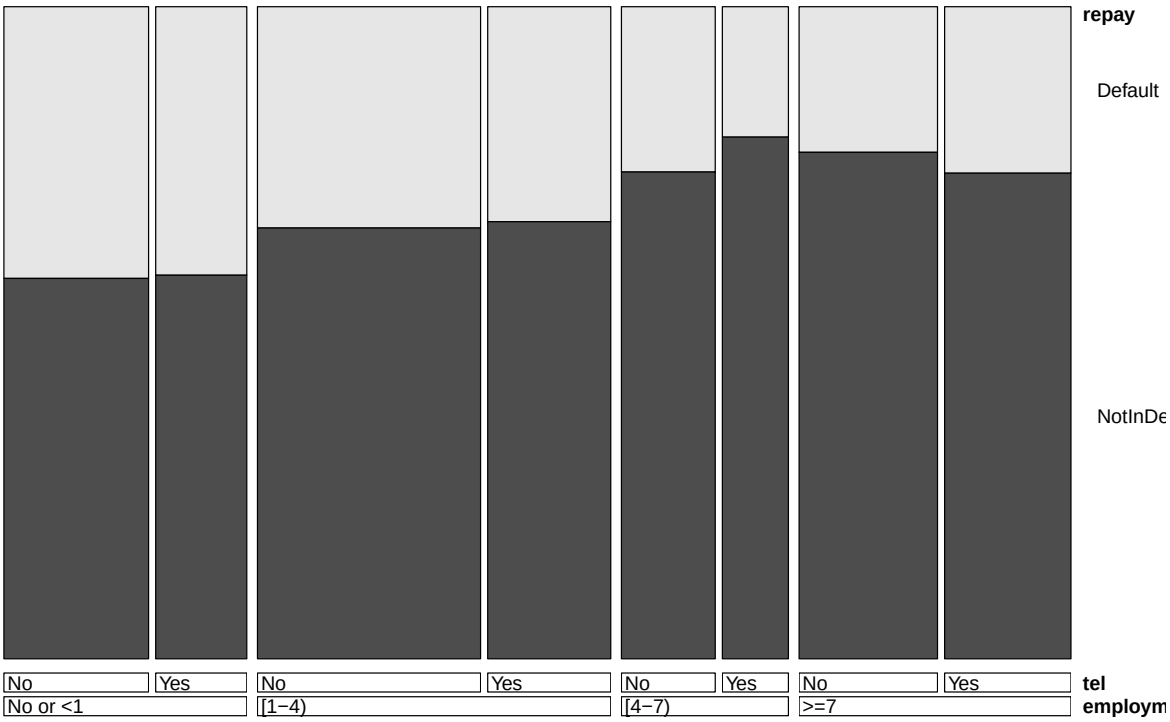
Confirmez ou infirmez votre réponse antérieure à l'aide d'un test approprié.

Solution:

```
tab <- xtabs(~ tel + repay + employm, data = mdata)
tab
```

tel	repay	employm	Freq
No	Default	[1-4)	56
		[4-7)	18
		>=7	23
		No or <1	45
	NotInDefault	[1-4)	109
		[4-7)	53
		>=7	80
		No or <1	63
Yes	Default	[1-4)	30
		[4-7)	10
		>=7	24
		No or <1	28
	NotInDefault	[1-4)	61
		[4-7)	40
		>=7	70
		No or <1	40

```
doubledecker(repay ~ employm + tel, tab)
```



Dans le graphique ci-dessus, lorsque l'on compare les blocs "Non ou <1", "[1-4)", "[4-7)" et ">=7", on constate une certaine similitude dans la structure des trois premiers blocs, ce qui suggère des rapports de cotes du même ordre. Mais il est difficile de savoir s'il s'agit d'une association homogène ou non. Pour clarifier les choses, nous pouvons calculer les rapports de cotes conditionnels.

```
loddsratio(tab, log = FALSE)
```

```
##  odds ratios for tel and repay by employm
##
## No or <1    [1-4)    [4-7)      >=7
##    1.020    1.045    1.358    0.839
```

Ces chiffres confirment nos observations, mais ne nous permettent pas de tirer de conclusion non plus.

Et voici le test demandé.

```
data <- data.frame(tab)
MS <- glm(Freq ~ repay * employm * tel, data = data, family = poisson)
drop1(MS, test = "Rao")
```

```
## Single term deletions
##
## Model:
## Freq ~ repay * employm * tel
##                   Df Deviance AIC Rao score Pr(>Chi)
## <none>                  0.000 120
## repay:employm:tel  3    0.768 115     0.764     0.86
```

Avec une p-valeur de presque 0.9, l'association homogène est "acceptée", au niveau 5%.

(h) Simplifiez le modèle log-linéaire saturé `repay`×`employm`×`tel` autant que possible en utilisant des tests sur des modèles emboîtés. Répétez la même analyse en utilisant l'AIC et puis le BIC.

Solution:

Méthode basée sur les tests :

```
# étape 1
drop1(MS, test = "Rao")
```

```
## Single term deletions
##
## Model:
## Freq ~ repay * employm * tel
##                    Df Deviance AIC Rao score Pr(>Chi)
## <none>                  0.000 120
## repay:employm:tel  3    0.768 115     0.764     0.86
```

```
# étape 2 (Voir l'output  ci-dessus; colonne Pr(>Chi))
MH <- update(MS, . ~ . - repay:employm:tel)
drop1(MH, test = "Rao")
```

```
## Single term deletions
##
## Model:
## Freq ~ repay + employm + tel + repay:employm + repay:tel + employm:tel
##               Df Deviance AIC Rao score Pr(>Chi)
## <none>              0.77 115
## repay:employm  3    18.60 127     17.81  0.00048 ***
## repay:tel      1     0.78 113      0.02  0.89956
## employm:tel    3     7.74 116      7.00  0.07187 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# étape 3
MC1 <- update(MH, . ~ . - repay:tel)
drop1(MC1, test = "Rao")
```

```
## Single term deletions
##
## Model:
## Freq ~ repay + employm + tel + repay:employm + employm:tel
##               Df Deviance AIC Rao score Pr(>Chi)
## <none>              0.78 113
## repay:employm  3    18.77 125     17.95  0.00045 ***
## employm:tel    3     7.91 114      7.15  0.06718 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# étape 4
MP <- update(MC1, . ~ . - employm:tel)
drop1(MP, test = "Rao")
```

```
## Single term deletions
##
## Model:
## Freq ~ repay + employm + tel + repay:employm
##               Df Deviance AIC Rao score Pr(>Chi)
## <none>               7.9 114
## tel            1     35.7 140      27.6  1.5e-07 ***
## repay:employm  3     25.9 126      17.9  0.00045 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Il n'y a plus de simplification possible. Le modèle retenu est MP:

```
formula(MP)
```

```
## Freq ~ repay + employm + tel + repay:employm
```

càd le modèle d'indépendance partielle entre `tel` et `(repay, employm)`.

Méthode basée sur l'AIC :

```
maic <- step(MS, direction = "both")
```

```
## Start:  AIC=120
## Freq ~ repay * employm * tel
##
##                      Df Deviance AIC
## - repay:employm:tel  3    0.768 115
## <none>                     0.000 120
##
## Step:  AIC=115
## Freq ~ repay + employm + tel + repay:employm + repay:tel + employm:tel
##
##                      Df Deviance AIC
## - repay:tel           1     0.78 113
## <none>                      0.77 115
## - employm:tel         3     7.74 116
## + repay:employm:tel   3     0.00 120
## - repay:employm       3    18.60 127
##
## Step:  AIC=113
## Freq ~ repay + employm + tel + repay:employm + employm:tel
##
##                  Df Deviance AIC
## <none>                 0.78 113
## - employm:tel     3     7.91 114
## + repay:tel       1     0.77 115
## - repay:employm   3    18.77 125
```

```
## formula(maic)
```

```
## Freq ~ repay + employm + tel + repay:employm + employm:tel
```

le modèle choisi par l'AIC est celui de l'indépendance conditionnelle entre `repay` et `tel` étant donné `employm`.

Méthode basée sur le BIC :

```
n <- nrow(data)
mbic <- step(MS, k = log(n), direction = "both")
```

```
## Start:  AIC=133
## Freq ~ repay * employm * tel
##
##                      Df Deviance AIC
## - repay:employm:tel  3    0.768 125
## <none>                     0.000 133
##
## Step:  AIC=125
```

```
## Freq ~ repay + employm + tel + repay:employm + repay:tel + employm:tel
##
##                     Df Deviance AIC
## - repay:tel          1     0.78 122
## - employm:tel        3     7.74 124
## <none>                     0.77 125
## + repay:employm:tel  3     0.00 133
## - repay:employm      3    18.60 135
##
## Step:  AIC=123
## Freq ~ repay + employm + tel + repay:employm + employm:tel
##
##                 Df Deviance AIC
## - employm:tel    3     7.91 121
## <none>                 0.78 122
## + repay:tel      1     0.77 125
## - repay:employm  3    18.77 132
##
## Step:  AIC=121
## Freq ~ repay + employm + tel + repay:employm
##
##                 Df Deviance AIC
## <none>                  7.9 121
## + employm:tel    3      0.8 122
## + repay:tel      1      7.7 124
## - repay:employm  3     25.9 131
## - tel            1     35.7 146
```

```
## formula(mbic)
```

```
## Freq ~ repay + employm + tel + repay:employm
```

Dans cet exemple, le BIC choisit le même modèle que la méthode 1 basée sur les tests.

Exercice 3

Le but de cet exercice est de montrer que l'indépendance deux à deux n'implique pas l'indépendance mutuelle.

Pour cela, considérant l'expérience qui consistent à lancer deux pièces (indépendamment l'une de l'autre). On définit les variables suivantes:

- X la variable aléatoire binaire prenant la valeur 1 si la première pièce montre face et 0 sinon,
- Y la variable aléatoire binaire prenant la valeur 1 si la deuxième pièce montre face et 0 sinon.
- Z la variable aléatoire prenant la valeur 1 si les deux pièces montrent le même coté et 0 sinon; càd $Z = I(X = Y)$.

(a) Montrer que X, Y et Z sont indépendantes deux à deux.

Solution:

- Par construction, les deux variables X et Y sont indépendantes.
- Montrons que X et Z sont indépendantes. Nous avons que

$$P(X = 1, Z = 1) = P(X = 1, Y = 1) = P(X = 1)\,P(Y = 1) = 0.25$$

  D'autre part,

$$P(X = 1) \times P(Z = 1) = 0.5 \times (P(X = 1, Y = 1) + P(X = 0, Y = 0)) = 0.5 \times (0.25 + 0.25) = 0.25.$$

Donc $P(X=1,Z=1) = P(X=1) \times P(Z=1)$. De la même façon, nous pouvons facilement montrer tous les autres cas, à savoir que $P(X=i,Z=j) = P(X=i,Z=j)$, $i,j=0,1$. On en déduit que X et Z sont indépendantes.

- Par symétrie, on peut aussi affirmer que Y et Z sont indépendantes.

Nous en concluons que ces trois variables sont indépendantes les unes des autres.

(b) Montrer que X, Y et Z ne sont pas mutuellement indépendants.

Solution:

Il nous suffit de trouver un exemple qui contredit l'affirmation comme quoi ces variables sont mutuellement indépendantes. Prenons par exemple la probabilité suivante

$$P(X=1,Y=0,Z=0) = P(X=1,Y=0) = 0.5 \times 0.5 = 0.25.$$

Alors que

$$\begin{aligned} P(X=1) \times P(Y=0) \times P(Z=0) &= 0.25 \times P(Z=0) \\ &= 0.25 \times (P(X=1)P(Y=0) + P(X=0)P(Y=1)) = 0.125. \end{aligned}$$

Dès lors, $P(X=1,Y=0,Z=0) \neq P(X=1) \times P(Y=0) \times P(Z=0)$ et nous n'avons pas l'indépendance mutuelle entre X, Y et Z.

Exercice 4

(a) Montrer que

$$\text{(a) } X \perp\!\!\!\perp Y|Z \text{ et } \text{ (b) } X \perp\!\!\!\perp Z|Y \iff \text{ (c) } X \perp\!\!\!\perp (Y,Z)$$

Solution:

(i) Montrons que (c) $\implies$ (a).

Nous supposons donc (c) $p(x|y,z) = p(x)$ et nous devons prouver (a) $p(x|y,z) = p(x|z)$. Pour cela, il suffit de montrer que $p(x|z) = p(x)$. Ce qui est vrai. En effet,

$$\text{(c)} \iff p(x,y,z) = p(x)p(y,z) \implies \sum_y p(x,y,z) = \sum_y p(x)p(y,z) \implies p(x,z) = p(x)p(z).$$

(ii) Montrons que (c) $\implies$ (b).

Il suffit pour cela de suivre exactement la même démarche que celle utilisée pour prouver (i).

(iii) Montrons que (a) et (b) $\implies$ (c).

Nous supposons donc (a) $p(x|y,z) = p(x|z)$ et (b) $p(x|y,z) = p(x|y)$ et nous devons prouver (c) $p(x|y,z) = p(x)$. Pour cela, il suffit encore une fois de montrer que $p(x|z) = p(x)$. Ce qui est vrai. En effet,

$$\begin{aligned} \text{(a) et (b)} &\implies p(x|z) = p(x|y) \implies p(x,z)p(y) = p(x,y)p(z) \\ &\implies \sum_y p(x,z)p(y) = \sum_y p(x,y)p(z) \implies p(x,z) = p(x)p(z). \end{aligned}$$

(b) Montrer que

$$\text{(a) } X \perp\!\!\!\perp (Y,Z) \text{ et } \text{ (b) } Y \perp\!\!\!\perp (X,Z) \iff \text{ (c) } X \perp\!\!\!\perp Y \perp\!\!\!\perp Z$$

Solution:

(i) Montrons que (c) $\implies$ (a).

Nous supposons donc (c) $p(x,y,z) = p(x)p(y)p(z)$ et nous devons prouver (a) $p(x,y,z) = p(x)p(y,z)$. Pour cela, il suffit de montrer que $p(y,z) = p(y)p(z)$. Ce qui est vrai. En effet,

$$\text{(c)} \implies \sum_x p(x,y,z) = \sum_x p(x)p(y)p(z) \implies p(y,z) = p(y)p(z).$$

(ii) Montrons que (c) $\implies$ (b).

Il suffit pour cela de suivre exactement la même démarche que celle utilisée pour prouver (i).

(iii) Montrons que (a) et (b) $\implies$ (c).

Nous supposons donc (a) $p(x,y,z) = p(x)p(y,z)$ et (b) $p(x,y,z) = p(y)p(x,z)$ et nous devons prouver (c) $p(x,y,z) = p(x)p(y)p(z)$. Pour cela, il suffit encore une fois de montrer que $p(y,z) = p(y)p(z)$. Ce qui est vrai. En effet,

$$\text{(a) et (b)} \implies p(x)p(y,z) = p(y)p(x,z) \implies \sum_x p(x)p(y,z) = \sum_x p(y)p(x,z) \implies p(y,z) = p(y)p(z).$$