DEA Analyse et Modélisation des Systèmes Biologiques

Introduction au logiciel R - 2000/2001

Contrôle sur machine (avec solution)

D. Chessel & J. Thioulouse

1. Questions

Récupérer les fichiers `wormshab.txt`, `human.txt` et `fiv.txt` à l'endroit habituel.

Lire le fichier wormshab.txt dans un dataframe que nous appellerons whab. Exécuter la commande attach(whab). Il s'agit des résultats d'une étude concernant l'influence des paramètres physico-chimiques du sol sur les vers de terre[1]. Le fichier a 62 lignes correspondant à 62 échantillons de sol prélevés dans des terrains différents. Les 18 colonnes correspondent à des variables caractérisant ces échantillons : Habitat = type de végétation couvrant le sol, Clay = pourcentage d'argile, Texture = texture du sol, MBC = carbone de la biomasse microbienne, worms = nombre de vers de terre, weight = poids total des vers de terre, pH = pH, P = phopsphore, K = potassium, S = souffre, Ca = calcium, Mg = magnesium, Silt = limon, Tsand = sable, OM = matière organique, Mn = manganèse, Cu = cuivre, Fe = fer.

	Habitat	Clay	Texture	MBC	worms	weight	pH	P	K	S	Ca	Mg	Silt	Tsand	OM	Mn	Cu	Fe
1	Burntcane	23	1	338	21	1.466	5.13	32	64	20	532	112	6	71	3.5	7	2	172
2	Burntcane	15	2	298	10	2.893	4.85	80	130	5	312	77	8	77	3.6	5	2	350
3	Burntcane	17	2	319	15	0.948	4.33	80	194	29	193	44	10	73	4.5	10	3	291
…																		
36	Gum	17	2	1411	30	12.130	5.38	24	68	8	1000	175	6	77	5.7	19	1	70
37	Gum	25	1	1459	1	0.440	4.43	58	142	45	113	82	11	64	6.2	70	7	242
38	Gum	21	1	1541	20	8.250	4.69	7	101	3	243	70	2	77	3.9	8	1	159
…																		
60	Trashedcane	40	3	481	73	16.630	4.92	14	169	12	608	167	8	52	2.6	75	13	25
61	Trashedcane	36	3	596	19	3.983	4.85	19	153	18	459	208	8	56	2.2	74	10	45
62	Trashedcane	38	3	692	28	3.238	4.76	21	101	18	539	266	8	54	2.5	62	8	38

1.1. Quel est le résultat de la commande suivante ?

```
whab[Clay<10, "worms"]
```

1.2. Comment calculer la moyenne des 17 variables quantitatives de whab ?

1.3. Comment calculer la matrice de covariance des 17 variables quantitatives de whab ?

1.4. Comment faire une représentation graphique pertinente du taux d'argile dans les différents types d'habitat des vers ?

1.5. Comment calculer le nombre moyen de vers dans les échantillons de sol où la teneur en argile est inférieure à la médiane (et supérieure à la médiane) ?

1.6. Comment calculer le poids moyen des vers dans les échantillons de sol où la teneur en argile est inférieure à la médiane (et supérieure à la médiane) ?

1.7. Comment calculer le pH moyen dans les échantillons de sol provenant des bananeraies ?

1.8. Comment trouver les numéros des échantillons de sol pour lesquels la teneur en fer est inférieure à 30, la teneur en cuivre est supérieure à 10, et le poids moyen des vers est inférieur à 0.10 ?

1.9. Quelle est la fonction de l'opérateur %/% ?

1.10. Si m est une matrice contenant un tableau individus (lignes) x variables (colonnes), que calcule la commande suivante ?

```
t(scale(m))%*%scale(m)/(nrow(m)-1)
```

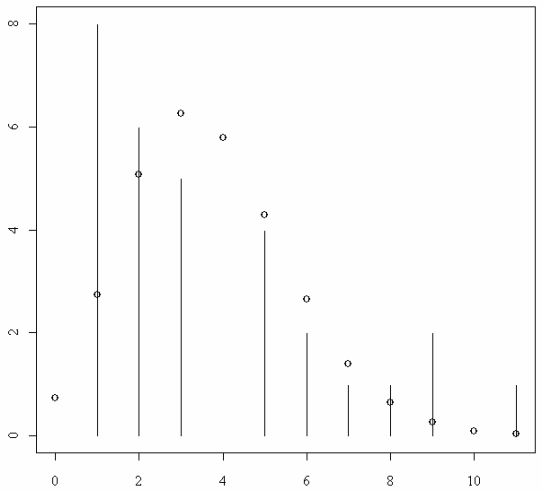
Dans les données originales de Mendel, on trouve [2] :

Expt. 2. Colour of albumen. 258 plants yielded 8023 seeds, 6022 yellow and 2001 green; their ratio, therefore is as 3.01 to 1 ...

1.11. Peut-on affirmer que les données sont trop proches du modèle (3/4 1/4) ? (utiliser la fonction chisq.test)

Par séquençage de l'ADN fécal, on a attribué 111 fèces de coyotes *Canis latrans* à 30 individus[3] . 8 individus sont représentés par une déjection, 6 individus sont représentés par 2 déjections, …, 1 individu est représenté par 11 déjections. La distribution complète est :

Crottes	1	2	3	5	6	7	8	9	11
Coyotes	8	6	5	4	2	1	1	2	1



1.12. Quel est la moyenne et la variance (*n* - 1) du nombre de fèces par coyotes ?

1.13. On veut ajuster une loi de Poisson. Comment est tracée la figure ?

1.14. Commenter le résultat.

Les données viennent d'une recherche sur une nouvelle méthode [4] de mesure de la composition du corps et donne le pourcentage de graisse (fat), l'âge (age) et le sexe (sex) de 18 personnes adultes en bonne santé agées de 23 à 61 ans. Lire le fichier human.txt.

```
> human
   age  fat sex
1   23  9.5   m
2   23 27.9   f
3   27  7.8   m
…
16  58 33.8   f
17  60 41.1   f
18  61 34.5   f
```

1.15. Comment l'âge et le pourcentage de graisse sont-ils reliés ?

1.16. Cette relation est elle différente chez les hommes (m) et chez les femmes (f) ?

Trois populations de chats *Felis catus* en milieu rural échantillonnées pendant plusieurs années ont permis [5] d'examiner 324 chats classés suivant le sexe (F-M), le génotype (Orange-Non Orange), l'âge (3 classes) et la présence d'anticorps spécifiques du virus FIV (*feline imunodeficiency virus*). Pour chaque combinaison sexe-génotype-âge on a le nombre d'individus séropositifs (fivposi) et le nombre d'individus négatifs (fivnega). Lire le fichier fiv.txt.

```
> fiv
   fivnega fivposi sex gen age
1       29       1   F  NO  a1
2       32       4   M  NO  a1
3       10       2   F   O  a1
4        7       2   M   O  a1
5       43       0   F  NO  a2
6       36       7   M  NO  a2
7       18       5   F   O  a2
8       16       7   M   O  a2
9       50       4   F  NO  a3
10      16       8   M  NO  a3
11      21       2   F   O  a3
12       2       2   M   O  a3
```

1.17. La prévalence dépend-elle du sexe ? La prévalence dépend-elle du génotype ?

1.18. La prévalence dépend-elle de l'âge ?

1.19. Quelles sont les estimations des paramètres du modèle "gen+sex" ?

1.20. Quelle est la fréquence observée de l'infection chez les mâles oranges ? Quelle est la probabilité prédite correspondante par le modèle "gen+sex" ?

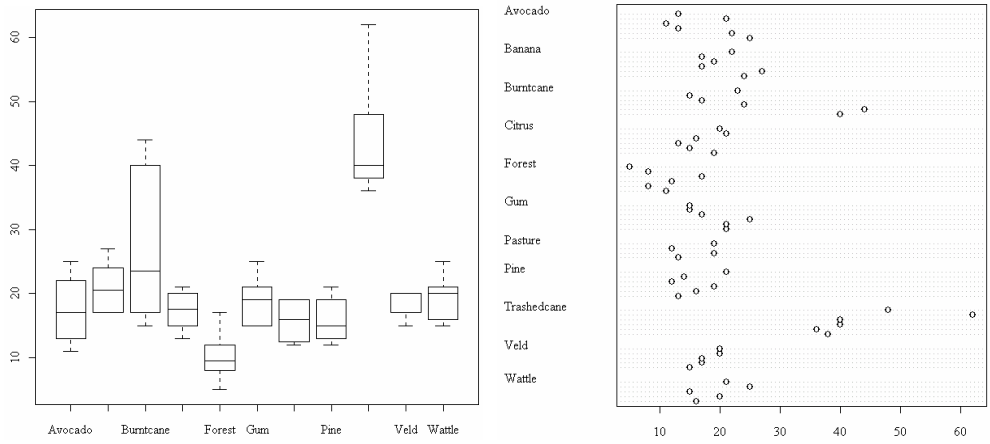
2. Solutions

```
1 > attach(whab)
> whab[Clay<10,"worms"]
[1]  89 162  96
```

```
2 > apply(whab[,2:18],2,mean)
    Clay  Texture      MBC    worms   weight       pH        P        K
  20.661    2.290 1565.242   41.855   11.020    5.322   32.597  185.355
 …
```

```
3 > cov(whab[,2:18])
             Clay   Texture        MBC    worms   weight       pH         P
Clay      106.0637   2.64093  -4214.835  -124.87  -44.845  -2.1696  -29.4995
…
```

```
4 > plot(Veg,Clay)
> dotplot(Clay,gr=Veg)
```




```
5 > tapply(worms,Clay<median(Clay),mean)
FALSE  TRUE
30.31 54.17
```

6 > tapply(weight/worms,Clay<median(Clay),mean)
```
 FALSE   TRUE
0.2585 0.2972
```

7 > tapply(pH,Veg=="Banana",mean)
```
FALSE  TRUE
5.271 5.797
> mean(pH[Veg=="Banana"])
[1] 5.797
```

8 > (1:62)[Fe<30 & Cu >10 & weight/worms <0.1]
```
[1]  6 58
```

9 > ?"%/%"
```
`%/%' indicates integer division . C'est la division entière
> 5%/%2
[1] 2
```

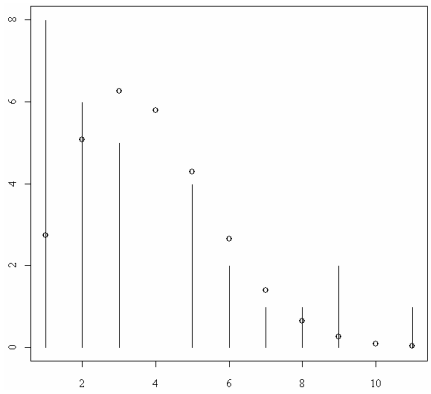
10 > ?scale
```
scale(x, center = TRUE, scale = TRUE)
> ?scale
> var(scale(worms))
     [,1]
[1,]    1
```
On obtient donc la matrice des corrélations

11 > chisq.test(c(6022,2001),p=c(3/4,1/4))

```
        Chi-square test for given probabilities

data:  c(6022, 2001)
X-squared = 0.015, df = 1, p-value = 0.9025
```
On ne peut pas rejeter, même au risque de 10%, l'hypothèse d'un tirage aléatoire sur (3/4,1/4). Le Khi2 n'est pas significativement faible.

12 > crottes
```
[1]  1  2  3  5  6  7  8  9 11
> coyotes
[1] 8 6 5 4 2 1 1 2 1
> a_rep(crottes,coyotes)
> a
 [1]  1  1  1  1  1  1  1  1  2  2  2  2  2  2  3  3  3  3  3  5  5  5  5  6  6
[26]  7  8  9  9 11
> mean(a)
[1] 3.7
> var(a)
[1] 8.08
```

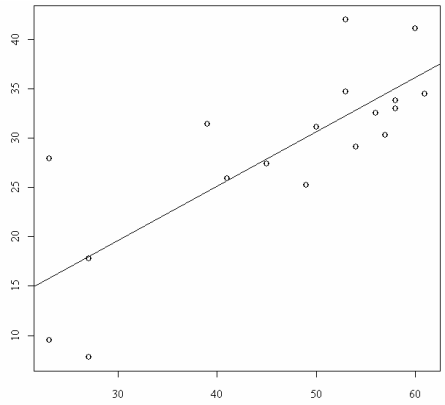
13 > plot(crottes,ncoy,ylim=c(0,8),type="h")
> points(0:11,dpois(0:11,mean(rep(crottes,ncoy)))*sum(ncoy))



14 > Un test est inutile pour rejeter clairement le modèle poissonien. On peut invoquer le marquage de leur territoire par les individus dominants.

15 > detach(whab)
> human

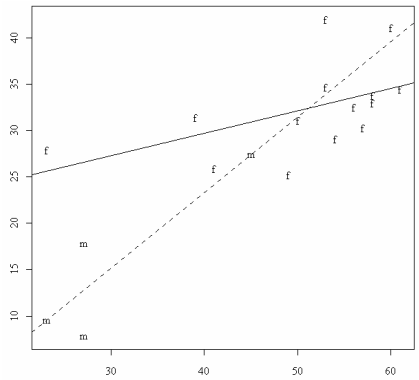
```
   age  fat sex
1   23  9.5   m
2   23 27.9   f
3   27  7.8   m
…
> attach(human)
> anova(lm0)
Analysis of Variance Table

Response: fat
          Df Sum Sq Mean Sq F value  Pr(>F)
age        1    892     892   44.27 1.1e-05 ***
sex        1    169     169    8.38   0.012 *
age:sex    1     79      79    3.91   0.068 .
Residuals 14    282      20
---
Signif. codes:  0  `***'  0.001  `**'  0.01  `*'  0.05  `.'  0.1  ` '  1

> summary(lm0)

Call:
lm(formula = fat ~ age * sex)

Residuals:
   Min     1Q Median     3Q    Max
-6.676 -2.886 -0.246  1.910  9.164

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   20.112      6.239    3.22   0.0061 **
age            0.240      0.120    1.99   0.0660 .
sexm         -29.269     10.410   -2.81   0.0139 *
age.sexm       0.572      0.289    1.98   0.0679 .
---

> plot(age,fat)
> abline(lm(fat~age))
```



Le pourcentage de graisse croit linéairement avec l'âge.

16 Il y a un effet sexe (p = 0.012) et on peut admettre une interaction (p = 0.068)

```
> plot(age,fat,pch=as.character(sex))
> abline(lm(fat[sex=="f"]~age[sex=="f"]))
> abline(lm(fat[sex=="m"]~age[sex=="m"]),lty=2)
```



On peut aussi se méfier du point pivot et dire NON.

```
17 > detach(human)
> fiv
   fivnega fivposi sex gen age
1       29       1   F  NO  a1
2       32       4   M  NO  a1
3       10       2   F   O  a1
…
Pour se faire une idée :

> glm0_glm(cbind(fivposi,fivnega)~sex+gen+age,family=binomial,data=fiv)
> glm0

Call:  glm(formula = cbind(fivposi, fivnega) ~ sex + gen + age, family = binomial,
data = fiv)

Coefficients:
(Intercept)         sexM         genO        agea2        agea3
     -3.444        1.500        1.031        0.280        0.874

Degrees of Freedom: 11 Total (i.e. Null);  7 Residual
Null Deviance:      34.4
Residual Deviance: 9.55         AIC: 51
> anova(glm0,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: cbind(fivposi, fivnega)

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                    11       34.4
sex   1     13.2        10       21.2   0.00027
gen   1      7.8         9       13.4   0.00512
age   2      3.8         7        9.5       0.1
> summary(glm0)

Call:
glm(formula = cbind(fivposi, fivnega) ~ sex + gen + age, family = binomial,
    data = fiv)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8866  -0.4328   0.0513   0.2748   1.5484

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.444      0.494   -6.97  3.2e-12 ***
sexM           1.500      0.376    3.98  6.8e-05 ***
genO           1.031      0.354    2.92   0.0036 **
agea2          0.280      0.449    0.62   0.5332
agea3          0.874      0.476    1.84   0.0660 .
---
Signif. codes:  0  `***'  0.001  `**'  0.01  `*'  0.05  `.'  0.1  ` '  1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 34.4332  on 11  degrees of freedom
Residual deviance:  9.5474  on  7  degrees of freedom
AIC: 50.99

Number of Fisher Scoring iterations: 4

La prévalence dépend fortement du sexe (augmentation pour les mâles)
La prévalence dépend fortement du génotype (augmentation pour le gène Orange).

18 > En première approche la réponse est NON (p = 0.1). Si on est aussi raffiné que
les auteurs, on peut invoquer une hypothèse alternative a priori d'une augmentation
avec l'âge et donc faire un test unilatéral sur le coefficient agea3. On obtient
alors un test significatif avec p = 0.0660/2 = 3%.

19 > glm(cbind(fivposi,fivnega)~sex+gen,family=binomial,data=fiv)

Call:   glm(formula = cbind(fivposi, fivnega) ~ sex + gen, family = binomial,
data = fiv)

Coefficients:
(Intercept)         sexM         genO
     -2.902        1.307        0.977

Degrees of Freedom: 11 Total (i.e. Null);  9 Residual
Null Deviance:      34.4
Residual Deviance: 13.4         AIC: 50.8

20 > sum(fiv$fivposi[(fiv$gen=="O" & fiv$sex=="M")])
[1] 11
> sum(fiv$fivnega[(fiv$gen=="O" & fiv$sex=="M")])
[1] 25
> 11/36
[1] 0.3056
>
predict(glm(cbind(fivposi,fivnega)~sex+gen,family=binomial,data=fiv),type="response"
)
 [1] 0.05208 0.16880 0.12734 0.35039 0.05208 0.16880 0.12734 0.35039 0.05208
[10] 0.16880 0.12734 0.35039

ou
> 1/(1+exp(2.9015 -0.9769 - 1.3073))
[1] 0.3504
```

3. Références

[1] Annual General Meeting of the South African Sugar Industry Agronomists' Association - 23 /11/2000. Mount Edgecombe.

[2] Mendel, G. (1956) Mathematics of heredity. In : The word of mathematics. Part V. Newman, J.R. (Ed.) Tempus Books of Microsoft Press. 923-934.

[3] Kohn, M.H., York, E.C., Kamradt, D.A., Haught, G., Sauvajot, R.M. & Wayne, R.K. (1999) Estimating population size by genotyping faeces. *Proceedings of the Royal Society of London B* : 266, 657-663.

[4] Mazess R.B., Peppler W.W. & Gibbons M. (1984) Total body composition by dual-photon (^{153}Gd) absorptiometry. *American Journal of Clinical Nutrition* : 40, 834-839. *In* Hand, D.J., Daly, F., Lunn, A.D., McConway, K.J. & Ostrowski, E. (1994*) A handbook of small data sets*. Chapman & Hall, London. 1-458.

[5] Pontier, D., Fromont, E., Courchamp, F., Artois, M. & Yoccoz, N.G. (1998) Retroviruses and sexual size dimorphism in domestic cats (Felis catus L.). *Proceedings of the Royal Society of London* B : 265, 167-173.