Fiche TD avec le logiciel R : bem11

Des bébés en avance : Manipulation et analyse de données

Salomé Bourg

Le but de ce TP est de vous placer (un peu plus) dans la peau d'un vrai statisticien. Contrairement à ce qu'on vous a fait croire jusqu'à maintenant, la majorité des jeu de données que vous aurez à traiter plus tard ne sont jamais aussi bien « rangés ». La première partie de ce TP consiste donc à apprendre à nettoyer un jeu de données. Les données sont extraites d'une étude réalisée au Duke University Medical Center à Durham aux États-Unis qui recense les naissances de 671 grands prématurés entre 1981 et 1987, pour lesquels de nombreuses variables ont été relevées.

1 Présentation du jeu de données

Les variables sont décrites dans la table 1 page 8.

2 Trier des données et manipulation de tableaux

Importez les données[1] dans R et enregistrez-les dans un objet appelé `prema` à l'aide de la fonction `read.table()`.

```
dim(prema)
```

```
[1] 671  27
```

```
head(prema)
```

```
  identifiant  birth   exit hospstay    lowph pltct  race  bwt gest born_at_duke twn
1           1 81.511 81.604       34       NA   100 white 1250   35            1   0
2           2 81.514 81.539        9 7.250000   244 white 1370   32            1   0
3           3 81.552 81.552       -2 7.059998   114 black  620   23            1   0
4           4 81.558 81.667       40 7.250000   182 black 1480   32            1   0
5           5 81.593 81.599        2 6.969997    54 black  925   28            1   0
6           6 81.602 81.771       62 7.189999    NA white  940   28            1   0
  lol magsulf meth toc  delivery apg1 vent pneumo pda cld      pvh      ivh    ipe
1  NA      NA    0   0 abdominal    8    0      0   0   0     <NA>     <NA>   <NA>
2  NA      NA    1   0 abdominal    7    0      0   0   0     <NA>     <NA>   <NA>
3  NA      NA    0   1   vaginal    1    1      0   0  NA     <NA>     <NA>   <NA>
4  NA      NA    1   0   vaginal    8    0      0   0   0     <NA>     <NA>   <NA>
5  NA      NA    0   0 abdominal    5    1      1   0   0 definite definite   <NA>
6  NA      NA    1   0 abdominal    8    1      0   0   0   absent   absent absent
     year    sex dead
1 81.51196 female    0
```

1. https://pbil.univ-lyon1.fr/R/donnees/vlbw1.txt

```
2 81.51471 female    0
3 81.55304 female    1
4 81.55847   male    0
5 81.59406 female    1
6 81.60229 female    0
```

2.1 Variables indéfinies et/ou non politiquement correcte

Les variables **ivh**, **pvh** et **ipe** sont « indéfinies ». Autrement dit, on ne sait pas à quoi elles correspondent. Elles n'auront donc pas d'intérêt pour la suite de l'analyse. Retirer ces variables/colonnes de l'objet `prema`. Nous n'utiliserons pas la variable **race** au cours de cette analyse, vous pouvez donc la retirer du jeu de données. À ce stade, votre objet `prema` est censé contenir `671` lignes et `23` colonnes. Vérifier cette information à l'aide de la fonction `dim()`.

```
dim(prema)
```

```
[1] 671  23
```

2.2 Individus sans âge

Certains individus n'ont pas leur date de naissance et/ou leur date de sortie de l'hôpital. Retirer ces individus de votre objet `prema` à l'aide de la fonction `subset()`. Votre objet ne doit plus contenir que `640` lignes. À l'aide de la fonction `dim()` vous devez obtenir :

```
dim(prema)
```

```
[1] 640  23
```

2.3 Valeurs surprenantes

D'après les informations que nous avons à notre disposition, l'étude a été réalisée entre 1981 et 1987. Intéressez-vous à l'individu 598. Que remarquez-vous ?

```
    identifiant  birth   exit hospstay    lowph pltct bwt gest born_at_duke twn lol
598         598 86.828 96.871     3668 7.129997   296 730   25            1   0   8
    magsulf meth toc delivery apg1 vent pneumo pda cld     year    sex dead
598       0    0   0  vaginal    3    1      0   1  NA 86.82953 female    1
```

Agissez en conséquence.

```
dim(prema)
```

```
[1] 639  23
```

La variable `hospstay` correspond à la durée que l'enfant a passé à l'hôpital. Intéressez-vous à l'individu 3 pour cette variable particulière. Que remarquez-vous ?

```
  identifiant  birth   exit hospstay    lowph pltct bwt gest born_at_duke twn lol
3           3 81.552 81.552       -2 7.059998   114 620   23            1   0  NA
  magsulf meth toc delivery apg1 vent pneumo pda cld     year    sex dead
3      NA    0   1  vaginal    1    1      0   0  NA 81.55304 female    1
```

Cet individu ne peut pas avoir passer un laps de temps « négatif » au sein du centre hospitalier. Les valeurs négatives pour cette variable n'ont pas de sens. Recalculer cette durée à partir des valeurs de `birth` et `exit`. Vous aurez besoin du nombre de jours dans une année qui correspond à 365.25 (utiliser cette valeur vous permet d'ignorer la présence d'une année bissextile, ici 1984). Il vous faudra maitriser la soustraction, la multiplication mais aussi la fonction `trunc()`. N'hésitez pas à consulter l'aide disponible dans R.

Malgré tous vos efforts, la variable `hospstay` contient encore des valeurs négatives probablement dues à une convention de codage des données (à quoi pourrat-elle correspondre ?). Retirez les individus pour lesquels on obtient ces valeurs négatives à l'aide de la fonction `subset()`.

```
dim(prema)
```
```
[1] 616  24
```

3 Manipulation des données et analyse préliminaire

3.1 Nombre de décès par an

Vous pouvez enfin passer à l'analyse de votre jeu de données. Une première hypothèse qui peut être faite est qu'entre 1981 et 1987, des avancées technologiques ont pu permettre une amélioration du nombre de bébés prématurés sauvés. On peut donc s'attendre à une diminution du pourcentage de décès en fonction de l'année dans ce centre hospitalier spécialisé. Cependant, dans l'état actuel de votre jeu de données, il vous est difficile de pouvoir directement répondre à cette question. C'est pourquoi vous allez devoir créer une nouvelle variable et un nouveau `data.frame`.

3.1.1 créer une nouvelle variable et un nouveau `data.frame`

Vous devez créer une variable qui correspond à la proportion de grands prématurés qui ont été sauvés (qui ne sont pas morts) parmi tous les grands prématurés accueillis au centre par an. Il existe différentes manières d'arriver à ce même résultat. Vous pouvez choisir de suivre votre propre méthode ou de suivre les lignes directrices proposées ci-après :

- fractionner votre jeu de données actuel par an (c'est à dire, créer différents `data.frame` contenant chacun l'intégralité des variables récoltées pour une année. Vous pouvez utiliser les fonctions `data.frame()` et `subset()`), par exemple :

```
dead81 <- subset(prema, trunc(prema$year) == 81)
nrow(dead81)
```
```
[1] 25
```

- calculer pour chacun de vos `data.frame` annuels, le taux de survie des grands prématurés.

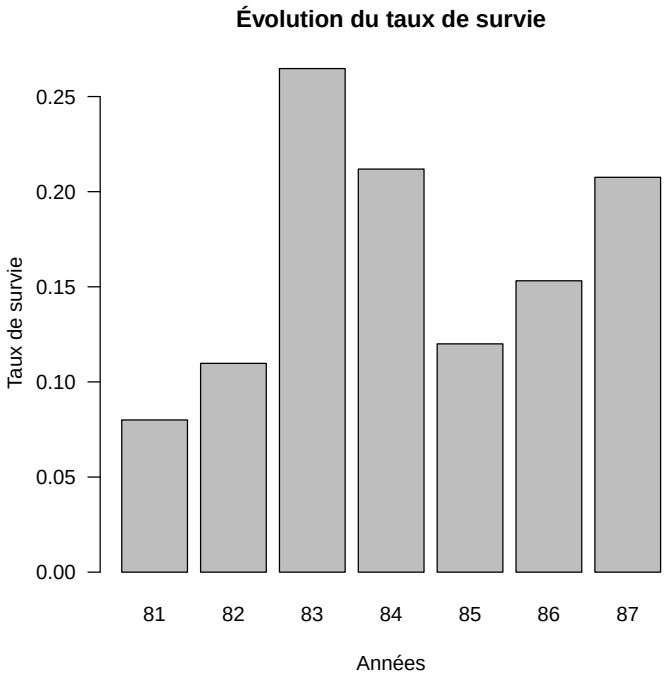
```
sum(dead81$dead)/nrow(dead81)
```
```
[1] 0.08
```

- créer un nouveau tableau contenant les taux de survie et les années correspondantes.

```
  annee tauxsurvie
1  1981  0.0800000
2  1982  0.1097561
3  1983  0.2647059
4  1984  0.2118644
5  1985  0.1200000
6  1986  0.1531532
7  1987  0.2075472
```

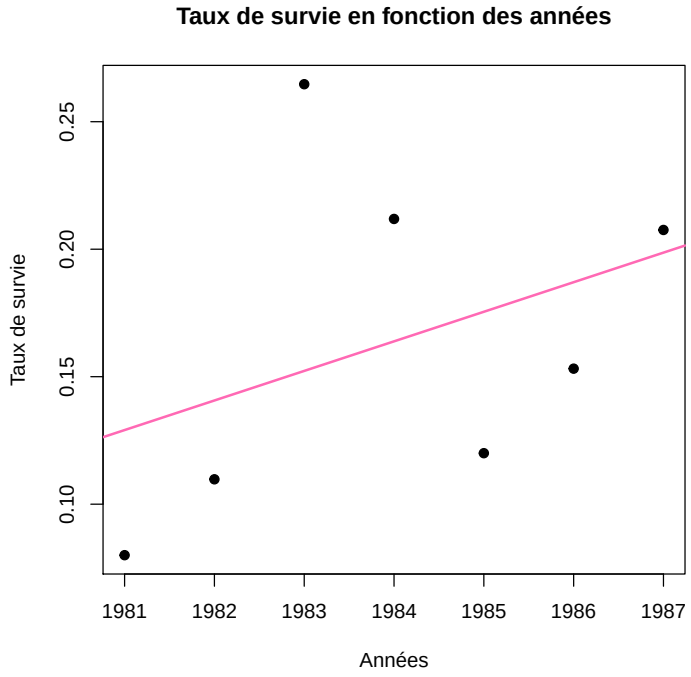
Une autre possibilité pour arriver au même résultat est d'utiliser la fonction `tapply()`, en utilisant l'aide de R essayez de comprendre comment elle fonctionne :

```
with(prema, tapply(dead, trunc(year), function(x) sum(x)/length(x))) -> res
barplot(res, main = "Évolution du taux de survie", xlab = "Années",
        ylab = "Taux de survie", las = 1)
```



3.1.2 Analyse

À partir de ces nouvelles données, tracer le graphique représentant le taux de survie en fonction de l'année en mettant des titres aux axes.

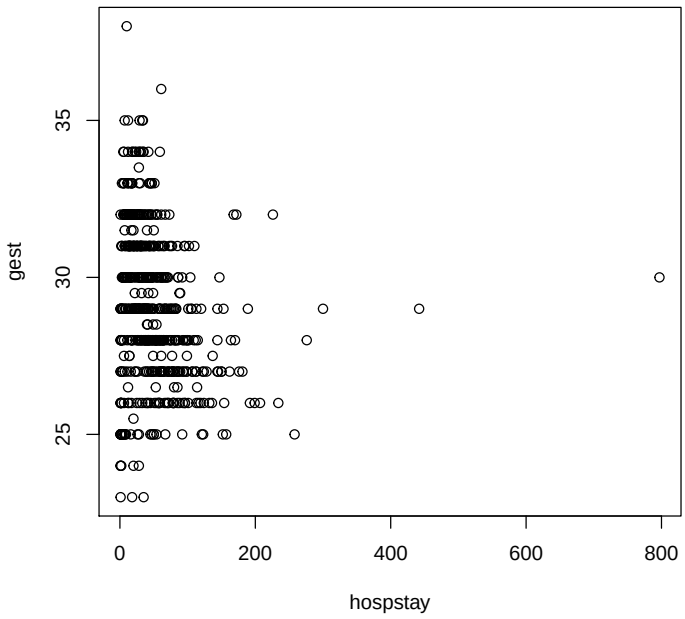


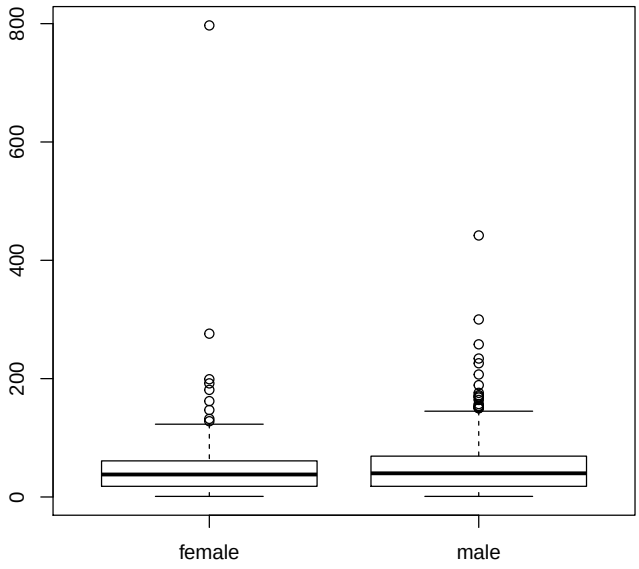
En vous basant sur ce que vous avez vu dans les TP précédents, que pouvez-vous conclure ?

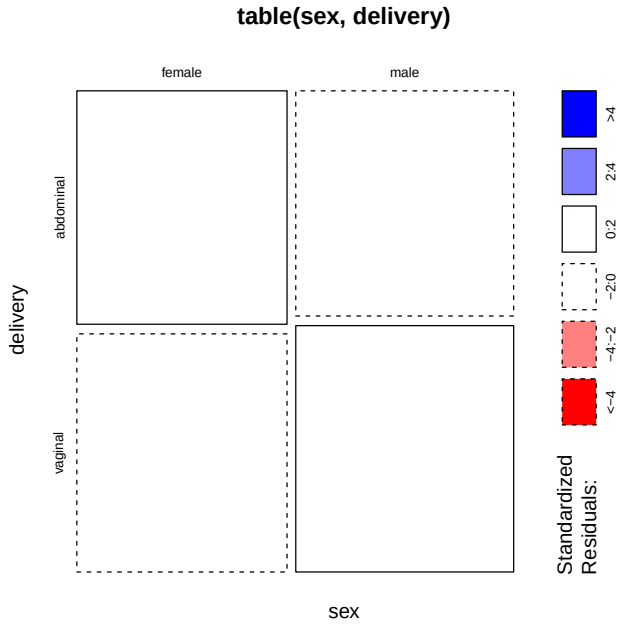
3.2 Pour aller plus loin

Explorez le jeu de données. D'autres questions peuvent trouver des réponses grâce aux nombreuses variables disponibles. Par exemple :

- Les variables durée de séjour à l'hôpital (`hospstay`) et durée de gestation (`gest`) sont-elles corrélées ?
- Le sexe a-t-il un effet sur la durée de séjour des grands prématurés ?
- La méthode d'accouchement et le sexe du bébé sont-ils corrélés ?


gest
25
30
35
0
200
400
600
800
hospstay


0
200
400
600
800
female
male



EST-CE que toutes les questions que vous pourriez poser ont un sens biologique ? Avez-vous tous les outils disponibles pour répondre aux questions qui vous intéressent ? Connaissez-vous d'autres tests qui pourraient permettre de répondre à plus de questions ?

birth	la date de naissance (en année depuis 1900)
exit	la date de sortie ou de décès
hospstay	la durée de séjour à l'hôpital en jours
lowph	la plus faible valeur de pH sanguin durant les 4 premiers jours de vie
pltct	le nombre de plaquettes (dans un volume non communiqué)
race	l'origine raciale [1]
bwt	masse à la naissance en grammes
gest	la durée de gestation en semaines
born at duke	prends la valeur 1 si né au centre médical Duke et 0 si transféré dans ce centre
twn	prend la valeur 0 en cas de gestation simple, 1 en cas de gestation multiple
lol	durée de travail en heures
magsulf	traitement de la mère au $MgSO_4$
meth	traitement de la mère au beta-methasone
toc	traitement de la mère avec des beta-adrénergiques
delivery	méthode d'accouchemnt
apg1	score d'Apgar après une minute
vent	utilisation de l'assistance ventilatoire
pneumo	occurence d'un pneumothorax
pda	persistence du canal artériel
cld	supplémentation en oxygène 30 jours après la naissance
ivh, pvh, ipe	indéfinis
year	année
sex	le sexe de l'enfant
dead	décès de l'enfant

TABLE 1 – Description des variables du jeu de données.

1. Notez que le relevé de cette variable n'est pas autorisé en France mais l'est aux Etats-Unis, et que les critères de constitution d'un groupe ethnique font débat.