18 décembre 2018
Durée *2h* - Bon courage !

— Les sorties de R sont dans l'annexe. Les sorties ne sont pas toutes utiles.
— Lorsque vous lisez une valeur dans les sorties de R données en annexe, précisez-moi bien dans quelle table vous la lisez.
— Apportez du soin aux conclusions des tests (phrase en français pour expliquer le résultat).

Exercice 1
Plusieurs variables ont été mesurées sur 50 états des Etats-Unis :

- Population : la taille estimée au 1er juillet 1975
- Income : revenu par population
- Illiteracy : illettrisme (en pourcentage de la population)
- Life.Exp espérance de vie moyenne
- Murder : taux d'homicide pour 100 000 individus
- HS Grad : pourcentage de diplômés niveau bac
- Frost : nombre de jours moyens avec des température négatives dans les grandes villes
- Area : surface de l'état en miles carrés

On souhaite expliquer la variable $Life.Exp$ à partir des autres variables.

```
> head(data)
           Population Income Illiteracy Life.Exp Murder HS.Grad Frost   Area
Alabama          3615   3624        2.1    69.05   15.1    41.3    20  50708
Alaska            365   6315        1.5    69.31   11.3    66.7   152 566432
Arizona          2212   4530        1.8    70.55    7.8    58.1    15 113417
Arkansas         2110   3378        1.9    70.66   10.1    39.9    65  51945
California      21198   5114        1.1    71.71   10.3    62.6    20 156361
Colorado         2541   4884        0.7    72.06    6.8    63.9   166 103766
```

Partie 1 : **ANNEXE 1**

1. Quel type de modèle proposez-vous ?
2. Quel est l'intérêt de la commande suivante, et quelle conclusion tirer ?

```
> res=lm( Life.Exp ~ ., data)
> library(car)
> vif(res)
Population     Income Illiteracy     Murder    HS.Grad      Frost       Area
  1.499915   1.992680   4.403151   2.616472   3.134887   2.358206   1.789764
```

3. Commenter l'analyse des résidus présentée dans l'annexe 1. Plusieurs étapes ont été faites à la suite des autres. Commenter ces étapes : pour chaque étape, si l'analyse des résidus vous satisfait, justifier pourquoi ; si non, expliquez ce qui ne vous convient pas et ce que l'on va faire pour y remédier. Peut-on s'arrêter dans l'analyse des résidus et garder le dernier modèle proposé ?

   Ce n'est pas la peine de m'écrire des lignes et des lignes... Mais je veux voir si vous avez compris la démarche de l'analyse des résidus : comment voit-on que le modèle est correct, ce que l'on peut faire si le modèle ne nous satisfait pas... Bref, ce que vous feriez en TP.

Partie 2 : **ANNEXE 2**
Dans cette partie, on supposera que le modèle sur lequel on travaille nous satisfait (cela peut être celui dont l'analyse des résidus est présentée en partie 1, ou un modèle amélioré si vous aviez jugé qu'il n'était pas satisfaisant). Les sorties sont dans l'annexe 2.

1. Tester s'il existe au moins une variable significative pour expliquer la variable $Life.Exp$.
2. On souhaite tester $H_0 : Life.Exp \sim x_2 + \ldots + x_7$ contre $H_1 : Life.Exp \sim x_1 + \ldots + x_7$. Donner l'expression de la statistique de test et sa loi sous H_0. Conclure grâce aux sorties. Puis effectuer les tests similaires pour les autres variables explicatives.
3. Effectuer des tests unilatéraux sur les variables que vous jugerez intéressantes, et conclure par une phrase en français chacun de ces tests.

4. On cherche un sous-modèle pour expliquer $Life.Exp$.
   Donner un sous-modèle qui pourrait convenir.
   On va effectuer un test pour décider si ce sous-modèle convient. Donner les hypothèses H_0 et H_1 associées à ce test.
   Donner les dimensions des modèles sous H_0 et sous H_1, et les expressions des vecteurs prédits par chacun des deux modèles.
   Donner l'expression de la statistique de test, sa loi sous H_0, et conclure en utilisant les sorties.
5. On a mis en oeuvre un algorithme de sélection de modèles.
   (a) Expliquer le principe en quelques lignes : à quoi ça sert ; quand est-ce-que c'est utile ; quels types d'algorithme existent ; pourquoi utiliser le critère AIC et notamment ce qu'est le sur-apprentissage et le sous-apprentissage.
   A nouveau, pas la peine de m'écrire des lignes et des lignes.
   (b) Le modèle obtenu avec l'algorithme stepwise est-il le même que celui que vous avez proposé dans la question précédente ? Est-ce- incohérent si on ne trouve pas les mêmes modèles ?
6. Donnez une prédiction de l'espérance de vie d'un état où
   Population=20000, Income=4000, Illiteracy=3 , Murder=10, HS.Grad=60, Frost=40, et Area=100000
   Vous pourrez donner l'expression de la valeur d'espérance de vie prédite puis faire les calculs.

Exercice 2 ANNEXE 3

On travaille sur le fichier ozone (celui déjà vu en régression multiple) qui donne le maximum de la concentration d'ozone pendant 112 jours à Rennes, en fonction notamment du facteur pluie (temps pluvieux ou temps sec) et du facteur vent (orientation du vent : est, nord, ouest ou sud). Voici la table de contingence des données :

```
> table(vent, pluie)
       pluie
vent    Pluie Sec
  Est       2   8
  Nord     10  21
  Ouest    26  24
  Sud       5  16
```

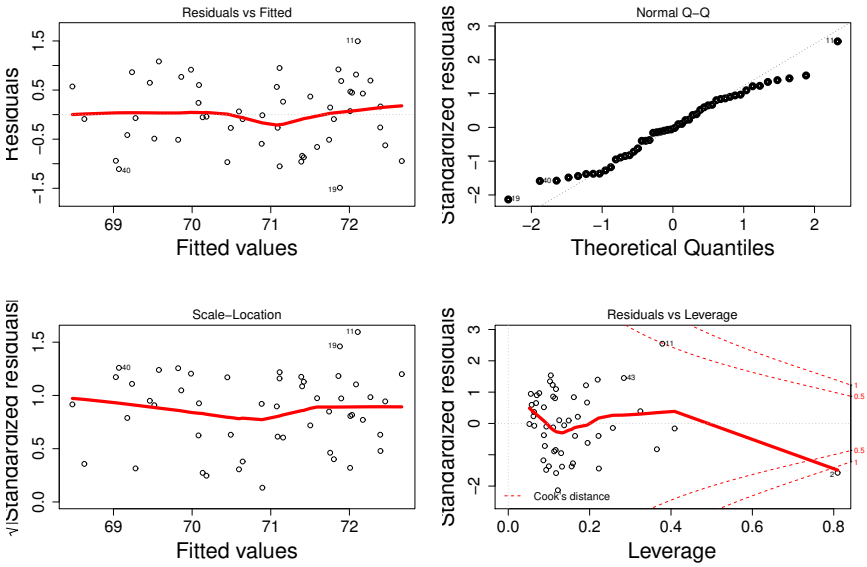
Il y a donc 2 jours où le temps était pluvieux avec un vent d'est, 8 jours de temps sec avec un vent d'est...

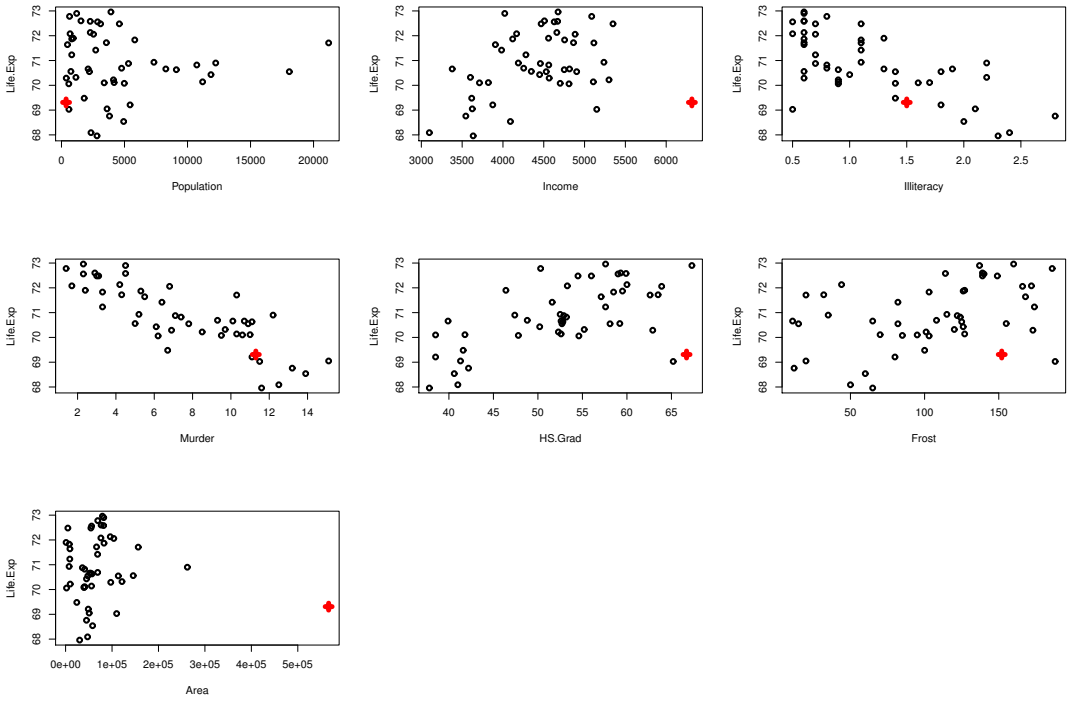
1. Commenter le plan d'expérience.
2. L'analyse descriptive des données, sous forme de boxplots est donnée dans l'annexe 3. La commenter.
3. Nous allons utiliser un modèle d'anova 2 pour étudier l'influence de la pluie et du vent sur la concentration en ozone. Justifier le choix d'un modèle d'anova 2.
4. A quelle condition peut-on modéliser un terme d'interaction entre les variables pluie et vent ? Peut-on modéliser un terme d'intéraction sur ces données ozone ?
5. Expliquer pourquoi en Anova2, il faut poser des contraintes d'identifiabilité sur les paramètres (ne pas expliciter les contraintes d'identifiabilité possibles).
6. Dans quel cas les tables anova(res) et Anova(res) sont-elles différentes et laquelle préférer alors ?

   Dans la suite, on supposera que le modèle sur lequel on travaille nous satisfait (ie. l'analyse des résidus a été effectuées et nous convient). A partir des sorties de Rde l'annexe 3, répondre aux questions suivantes, en précisant bien la table que vous utilisez.
7. Est-ce-qu'au moins un des facteurs vent ou pluie est significatif pour expliquer le variable $maxO3$?
8. Tester l'existence d'un terme d'interaction entre les facteurs vent et pluie.
9. Tester l'existence d'un effet vent.
10. Tester l'existence d'un effet pluie.
11. Quel modèle allez-vous donc garder ?
12. Faire des tests de comparaison deux à deux (tests unilatéraux) pour comparer la concentration d'ozone, en fonction de si le temps est sec ou pluvieux, et en fonction de l'orientation du vent.
    Dans ce type de tests de comparaison deux à deux, qu'est ce qu'il ne faut surtout pas oublier de faire, et pourquoi ?

ANNEXE 1 (exercice 1, partie 1) ANALYSE DES RESIDUS

```
> res=lm(Life.Exp ~ ., data)
> par(mfrow=c(2,2))
> plot(res)
> abs(rstudent(res))[abs(rstudent(res))>2]
      11       19
2.735242 2.232206

> influence=influence.measures(res)
> leviers=influence$infmat[,"hat"]
> n=length(Population); p= 8;
> 2*p/n;
[1] 0.3333333
> leviers[leviers > 2*p/n]
        2         5        28        31
0.8174807 0.4095683 0.3661115 0.3457803

> shapiro.test(res$residuals)
W = 0.97122, p-value = 0.2946
```



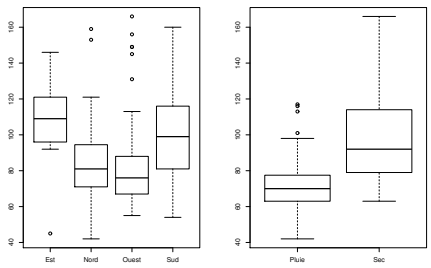


graphes de y en fonction de chaque variable x_i. La croix correspond à la donnée numéro 2.

ANNEXE 2 (exercice 1, partie 2)

```
res=lm( Life.Exp ~ .)
> summary(res)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.266e+01  1.670e+00  43.519  < 2e-16 ***
Population   5.626e-05  2.763e-05   2.036   0.0486 *
Income      -1.031e-04  2.610e-04  -0.395   0.6950
Illiteracy  -2.686e-01  3.346e-01  -0.803   0.4269
Murder      -3.002e-01  4.025e-02  -7.460 5.05e-09 ***
HS.Grad      1.785e-02  2.219e-02   0.804   0.4260
Frost       -2.800e-03  2.769e-03  -1.011   0.3182
Area         4.822e-06  2.433e-06   1.982   0.0546 .
Residual standard error: 0.6344 on 39 degrees of freedom
Multiple R-squared:  0.7993,Adjusted R-squared:  0.7632
F-statistic: 22.18 on 7 and 39 DF,  p-value: 9.286e-12

> anova(res)
           Df  Sum Sq Mean Sq F value    Pr(>F)
Population  1  0.3567  0.3567  0.8864   0.35226
Income      1 18.9669 18.9669 47.1302 3.284e-08 ***
Illiteracy  1 17.7559 17.7559 44.1210 6.662e-08 ***
Murder      1 21.1625 21.1625 52.5860 9.697e-09 ***
HS.Grad     1  2.2843  2.2843  5.6762   0.02217 *
Frost       1  0.3822  0.3822  0.9498   0.33577
Area        1  1.5809  1.5809  3.9283   0.05456 .
Residuals  39 15.6950  0.4024

res12 =lm( Life.Exp ~ Population + Murder)
> anova(res12, res)
  Res.Df    RSS Df Sum of Sq     F  Pr(>F)
1     44 20.681
2     39 15.695  5    4.9862 2.478 0.04834 *

res123 =lm( Life.Exp ~ Population + Murder + Area)
> anova(res123, res)
    Frost + Area
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     43 16.947
2     39 15.695  4    1.2517 0.7776 0.5465

res0=lm(Life.Exp ~1)
modele_Stepwise=step(res0,scope=formula(res),direction="both")
> summary(modele_Stepwise)
Call:
lm(formula = Life.Exp ~ Murder + Area + Population + HS.Grad,
    data = data4)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.138e+01  9.391e-01  76.014  < 2e-16 ***
Murder      -3.003e-01  3.450e-02  -8.704 5.96e-11 ***
Area         4.850e-06  2.231e-06   2.174   0.0354 *
Population   5.915e-05  2.220e-05   2.665   0.0109 *
HS.Grad      2.173e-02  1.565e-02   1.388   0.1725
Residual standard error: 0.6211 on 42 degrees of freedom
Multiple R-squared:  0.7928,Adjusted R-squared:  0.773
F-statistic: 40.16 on 4 and 42 DF,  p-value: 7.816e-14
```

ANNEXE 3 (exercice 2)




```
################ Modèle AVEC interaction ###############
res=lm(log(maxO3) ~ vent*pluie)
> summary(res)
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        4.18551    0.14408  29.051  < 2e-16 ***
ventNord           0.02239    0.15783   0.142  0.88747
ventOuest          0.05774    0.14973   0.386  0.70066
ventSud            0.29073    0.17047   1.705  0.09142 .
pluieSec           0.54457    0.16108   3.381  0.00105 **
ventNord:pluieSec -0.28460    0.18001  -1.581  0.11724
ventOuest:pluieSec -0.37388    0.17261  -2.166  0.03283 *
ventSud:pluieSec  -0.42698    0.19240  -2.219  0.02888 *
Residual standard error: 0.2038 on 94 degrees of freedom
Multiple R-squared:  0.4011,Adjusted R-squared:  0.3565
F-statistic: 8.992 on 7 and 94 DF,  p-value: 1.893e-08

> anova(res)
           Df Sum Sq Mean Sq F value    Pr(>F)
vent        3 1.3379 0.44597 10.7422 3.900e-06 ***
pluie       1 1.0318 1.03176 24.8524 2.818e-06 ***
vent:pluie  3 0.2435 0.08118  1.9554     0.126
Residuals  94 3.9025 0.04152

> Anova(res)
           Sum Sq Df F value    Pr(>F)
vent       0.7325  3  5.8816  0.001005 **
pluie      1.0318  1 24.8524 2.818e-06 ***
vent:pluie 0.2435  3  1.9554  0.126035
Residuals  3.9025 94

> comp.Tukey = TukeyHSD(aov(log(maxO3) ~ vent * pluie))
$vent
                 diff         lwr         upr     p adj
Nord-Est   -0.20080991 -0.43772730 0.036107478 0.1264407
Ouest-Est  -0.22476974 -0.45044112 0.000901635 0.0513304
Sud-Est    -0.03241550 -0.28271331 0.217882312 0.9866126
Ouest-Nord -0.02395983 -0.17288308 0.124963422 0.9749331
Sud-Nord    0.16839441 -0.01572392 0.352512741 0.0857064
Sud-Ouest   0.19235424  0.02295152 0.361756960 0.0193634

$pluie
              diff       lwr       upr p adj
Sec-Pluie 0.266355 0.1702269 0.3624831 3e-07

$`vent:pluie`
                              diff         lwr       upr     p adj
Nord:Pluie-Est:Pluie    0.02239438 -0.57546810 0.6202569 1.0000000
Ouest:Pluie-Est:Pluie   0.07635713 -0.49001649 0.6427308 0.9998927
Sud:Pluie-Est:Pluie     0.29072528 -0.35504001 0.9364906 0.8586336
Est:Sec-Est:Pluie       0.54456610 -0.06562474 1.1547569 0.1165936
Nord:Sec-Est:Pluie      0.33600897 -0.23516046 0.9071784 0.6081748
```

```
Ouest:Sec-Est:Pluie     0.35661965 -0.21143711 0.9246764 0.5256738
Sud:Sec-Est:Pluie       0.43839741 -0.14048045 1.0172753 0.2810247
Ouest:Pluie-Nord:Pluie  0.05396275 -0.23324110 0.3411666 0.9990332
Sud:Pluie-Nord:Pluie    0.26833090 -0.15442172 0.6910835 0.5112614
Est:Sec-Nord:Pluie      0.52217172  0.15605722 0.8882862 0.0006429
Nord:Sec-Nord:Pluie     0.31361459  0.01706531 0.6101639 0.0303059
Ouest:Sec-Nord:Pluie    0.33422527  0.04371632 0.6247342 0.0126703
Sud:Sec-Nord:Pluie      0.41600303  0.10486553 0.7271405 0.0017867
Sud:Pluie-Ouest:Pluie   0.21436814 -0.16253942 0.5912757 0.6484643
Est:Sec-Ouest:Pluie     0.46820897  0.15615230 0.7802656 0.0002650
Nord:Sec-Ouest:Pluie    0.25965184  0.03319873 0.4861049 0.0131595
Ouest:Sec-Ouest:Pluie   0.28026252  0.06177915 0.4987459 0.0032432
Sud:Sec-Ouest:Pluie     0.36204028  0.11679331 0.6072873 0.0003550
Est:Sec-Sud:Pluie       0.25384083 -0.18617405 0.6938557 0.6317633
Nord:Sec-Sud:Pluie      0.04528369 -0.33879278 0.4293602 0.9999567
Ouest:Sec-Sud:Pluie     0.06589437 -0.31353771 0.4453265 0.9994216
Sud:Sec-Sud:Pluie       0.14767214 -0.24777673 0.5431210 0.9424877
Nord:Sec-Est:Sec       -0.20855713 -0.52923578 0.1121215 0.4790335
Ouest:Sec-Est:Sec      -0.18794645 -0.50304765 0.1271547 0.5911286
Sud:Sec-Est:Sec        -0.10616869 -0.44038398 0.2280466 0.9759523
Ouest:Sec-Nord:Sec      0.02061068 -0.21001978 0.2512411 0.9999935
Sud:Sec-Nord:Sec        0.10238845 -0.15373946 0.3585164 0.9189762
Sud:Sec-Ouest:Sec       0.08177776 -0.16733160 0.3308871 0.9711174

################ Modèle SANS interaction ###############
res=lm(log(maxO3) ~ vent+pluie)
> summary(res)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.45013    0.07407  60.082  < 2e-16 ***
ventNord    -0.21254    0.07647  -2.779  0.00654 **
ventOuest   -0.22550    0.07418  -3.040  0.00304 **
ventSud     -0.04605    0.08010  -0.575  0.56671
pluieSec     0.21379    0.04351   4.913 3.64e-06 ***
Residual standard error: 0.2067 on 97 degrees of freedom
Multiple R-squared:  0.3637,Adjusted R-squared:  0.3374
F-statistic: 13.86 on 4 and 97 DF,  p-value: 5.603e-09

> anova(res)
          Df Sum Sq Mean Sq F value    Pr(>F)
vent       3 1.3379 0.44597  10.434 5.182e-06 ***
pluie      1 1.0318 1.03176  24.139 3.637e-06 ***
Residuals 97 4.1460 0.04274

> Anova(res)
          Sum Sq Df F value    Pr(>F)
vent      0.7325  3  5.7128   0.00121 **
pluie     1.0318  1 24.1391 3.637e-06 ***
Residuals 4.1460 97

> comp.Tukey = TukeyHSD(aov(log(maxO3) ~ vent + pluie))
$vent
                  diff         lwr          upr     p adj
Nord-Est   -0.20080991 -0.43759587 0.0359760461 0.1261924
Ouest-Est  -0.22476974 -0.45031592 0.0007764419 0.0511457
Sud-Est    -0.03241550 -0.28257446 0.2177434567 0.9866110
Ouest-Nord -0.02395983 -0.17280047 0.1248808055 0.9749298
Sud-Nord    0.16839441 -0.01562178 0.3524105999 0.0854821
Sud-Ouest   0.19235424  0.02304550 0.3616629827 0.0192506

$pluie
              diff      lwr       upr p adj
Sec-Pluie 0.266355 0.170268 0.3624419 3e-07
```