Bordeaux INP

ENSC - 2ème année

CO7SFMA1 - Modélisation statistique - EXAMEN Janvier 2020 (durée : 1h30)

La qualité de la rédaction sera prise en compte dans la notation. Les deux exercices sont indépendants.

Vous devez traiter les 2 exercices proposés.

Pour les interprétations des différents tests, prendre toujours un risque de première espèce $\alpha = 5\%$.

Exercice 1 : régression linéaire multiple

Un cogniticien s'intéresse à une base de données (data frame `stateEU`) qui contient les mesures sur $n = 50$ états des Etats-Unis des 8 variables quantitatives suivantes :

- `Population` : population estimée,
- `Income` : revenu par individu,
- `Illiteracy` : illettrisme (pourcentage de la population),
- `Life.Exp` : espérance de vie moyenne (en années),
- `Murder` : taux d'homicide pour 100 000 individus,
- `HS Grad` : pourcentage de diplômés niveau baccalauréat (high-school graduates),
- `Frost` : nombre de jours moyens avec des températures inférieures à 0 degré Celcius (sur les 30 dernières années) dans les grandes villes,
- `Area` : surface de l'état en miles carrés.

Il désire étudier le lien qui peut exister entre la variable quantitative `Life.Exp` (que l'on pourra noter `y`) à expliquer et les $p = 7$ autres variables quantitatives (variables explicatives que l'on pourra noter `x1, x2, ..., x7`).

Le cogniticien chargé de l'étude utilise le logiciel R pour faire les traitements statistiques.

- Dans un premier temps, il décide d'utiliser un modèle de régression linéaire multiple avec l'ensemble des variables explicatives (modèle 1).
- Puis il fait de la sélection automatique de variables (en utilisant la fonction `step`) pour aboutir au modèle 2.

Les codes R et des extraits des sorties numériques et graphiques associées sont fournis ci-après.

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

> # Modele 1 (modele complet)
> #===========================
> res<-lm(Life.Exp~.,data=stateEU)
> summary(res)
Call:
lm(formula = Life.Exp ~ ., data = stateEU)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.094e+01  1.748e+00  40.586  < 2e-16 ***  # <--- *** ligne L4 ***
Population   5.180e-05  2.919e-05   1.775   0.0832 .
Income      -2.180e-05  2.444e-04  -0.089   0.9293
Illiteracy   3.382e-02  3.663e-01   0.092   0.9269
Murder      -3.011e-01  4.662e-02  -6.459 8.68e-08 ***   # <--- *** ligne L3 ***
HS.Grad      4.893e-02  2.332e-02   2.098   0.0420 *
Frost       -5.735e-03  3.143e-03  -1.825   0.0752 .  # <--- *** ligne L2 ***
Area        -7.383e-08  1.668e-06  -0.044   0.9649
---
Residual standard error: 0.7448 on 42 degrees of freedom
Multiple R-squared:  0.7362,Adjusted R-squared:  0.6922
F-statistic: 16.74 on 7 and 42 DF,  p-value: 2.534e-10   # <--- *** ligne L1 ***

> # Selection de variables
> #=======================
> step(res)
Start:  AIC=-22.18
Life.Exp ~ Population + Income + Illiteracy + Murder + HS.Grad +
    Frost + Area

             Df Sum of Sq    RSS     AIC
- Area        1    0.0011 23.298 -24.182
- Income      1    0.0044 23.302 -24.175
- Illiteracy  1    0.0047 23.302 -24.174
<none>                    23.297 -22.185
- Population  1    1.7472 25.044 -20.569
- Frost       1    1.8466 25.144 -20.371
- HS.Grad     1    2.4413 25.738 -19.202
- Murder      1   23.1411 46.438  10.305

Step:  AIC=-24.18
Life.Exp ~ Population + Income + Illiteracy + Murder + HS.Grad +
    Frost

             Df Sum of Sq    RSS     AIC
- Illiteracy  1    0.0038 23.302 -26.174
- Income      1    0.0059 23.304 -26.170
<none>                    23.298 -24.182
- Population  1    1.7599 25.058 -22.541
- Frost       1    2.0488 25.347 -21.968
- HS.Grad     1    2.9804 26.279 -20.163
- Murder      1   26.2721 49.570  11.569

Step:  AIC=-26.17
Life.Exp ~ Population + Income + Murder + HS.Grad + Frost

             Df Sum of Sq    RSS     AIC
- Income      1     0.006 23.308 -28.161
<none>                    23.302 -26.174
- Population  1     1.887 25.189 -24.280
- Frost       1     3.037 26.339 -22.048
- HS.Grad     1     3.495 26.797 -21.187
- Murder      1    34.739 58.041  17.456

Step:  AIC=-28.16
Life.Exp ~ Population + Murder + HS.Grad + Frost

             Df Sum of Sq    RSS     AIC
<none>                    23.308 -28.161
- Population  1     2.064 25.372 -25.920
- Frost       1     3.122 26.430 -23.877
- HS.Grad     1     5.112 28.420 -20.246
- Murder      1    34.816 58.124  15.528
```

```
Call:
lm(formula = Life.Exp ~ Population + Murder + HS.Grad + Frost,
    data = stateEU)

Coefficients:
(Intercept)   Population       Murder      HS.Grad        Frost
  7.103e+01    5.014e-05   -3.001e-01    4.658e-02   -5.943e-03

> # Modele 2 (modele simplifie)
> #============================
> res2 <- lm(formula = Life.Exp ~ Population + Murder + HS.Grad + Frost,  data = stateEU)
> summary(res2)
Call:
lm(formula = Life.Exp ~ Population + Murder + HS.Grad + Frost,
    data = stateEU)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.103e+01  9.529e-01  74.542  < 2e-16 ***
Population   5.014e-05  2.512e-05   1.996  0.05201 .
Murder      -3.001e-01  3.661e-02  -8.199 1.77e-10 ***
HS.Grad      4.658e-02  1.483e-02   3.142  0.00297 **
Frost       -5.943e-03  2.421e-03  -2.455  0.01802 *
---
Residual standard error: 0.7197 on 45 degrees of freedom
Multiple R-squared:  0.736,Adjusted R-squared:  0.7126
F-statistic: 31.37 on 4 and 45 DF,  p-value: 1.696e-12

> shapiro.test(res2$residuals)
Shapiro-Wilk normality test

data:  res2$residuals
W = 0.97935, p-value = 0.525

> plot(res2$fitted,res2$residuals)   # ---> *** Figure 1 ***
> abline(h=0)

> # Prediction
> #============
> predict(res2,data.frame(Murder=8, HS.Grad=75, Frost=80, Population=4250),interval="prediction",level=0.95)
       fit      lwr      upr
1 71.85724 70.23526 73.47921      # <--- *** ligne L5 ***
```

1. Quel modèle théorique est considéré à la partie `Modele 1` ?
2. a) Quelles sont les hypothèses H_0 et H_1 testées à la `* ligne L1 *` ? Préciser la conclusion de ce test.
   b) Quelles sont les hypothèses H_0 et H_1 testées à la `* ligne L2 *` ? Préciser la conclusion de ce test.
   c) Quelles sont les hypothèses H_0 et H_1 testées à la `* ligne L3 *` ? Préciser la conclusion de ce test.
   d) Quelles sont les hypothèses H_0 et H_1 testées à la `* ligne L4 *` ? Préciser la conclusion de ce test.
   e) Donner une estimation sans biais de l'écart-type du terme d'erreur du modèle.
3. Pour quelles raisons, le statisticien n'est-il pas satisfait du modèle 1 qui a été estimé ?
4. a) Rappeler en quelques lignes la philosophie du critère AIC et la procédure de sélection mise en œuvre ici avec la fonction "`step`" ?
   b) Quel est le modèle final retenu à l'issue de la procédure de sélection de variables ?
   c) Justifier pourquoi le modèle simplifié obtenu (`Modele 2`) convient au cogniticien en charge de l'étude.
   d) De votre côté, aurriez-vous poursuivi l'étude ? Justifier votre réponse.
5. Une prédiction de l'espérance de vie a été réalisée à la `* ligne L5 *`.
   a) Sur quelle base (modèle et valeurs considérées) a été réalisée cette prédiction ?
   b) Préciser à quoi correspondent les trois valeurs apparaissant sur la `* ligne L5 *`.

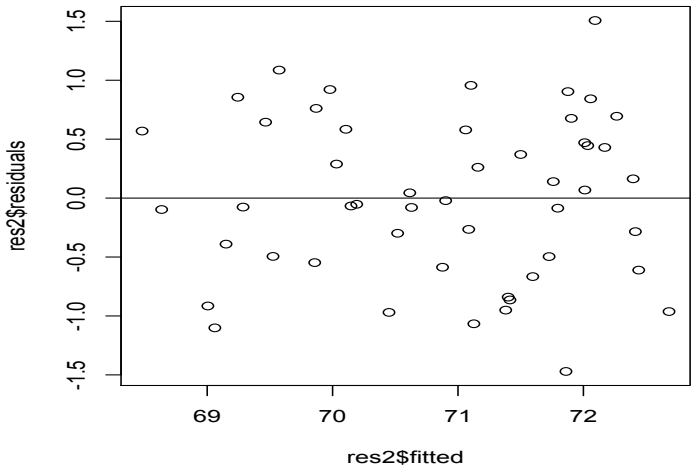

Figure 1: Graphique des résidus du modèle 2

Exercice 2 : Analyse de la variance

Une cogniticienne participe à une étude aéronautique de type "Facteur Humain" portant sur le stress des pilotes en situation extrême. Elle dispose des données expérimentales suivantes. $n = 120$ pilotes (6 groupes de 20 pilotes) ont passé (en simulateur de vols) une expérience de pilotage avec une phase d'atterrissage très complexe (mauvais temps) durant laquelle le niveau de stress du pilote a été mesuré : variable quantitative `Stress` qui correspond à un score compris entre 0 (pas de stress) et 10 (stress maximal). Les conditions du mauvais temps et ses effets sur le pilotage de l'avion sont contrôlés par deux facteurs :

- facteur `Vibration` à 3 niveaux : `V1` (fortes vibrations), `V2` (vibrations moyennes) et `V3` (peu ou pas de vibrations),
- facteur `Bruit` dans le cockpit à 2 niveaux : `B1` (présence de bruits anormaux) et `V2` (peu ou pas de bruits anormaux).

Dans cette étude, la cogniticienne doit évaluer si les facteurs `Vibration` et `Bruit` ont un impact sur le niveau de `Stress` du pilote. Pour cela, elle décide de faire de l'ANOVA (analyse de la variance) à deux facteurs. Les codes R et des extraits des sorties numériques et graphiques associées sont fournis ci-après.

```
> table(Vibration,Bruit)
         Bruit
Vibration B1 B2
       V1 20 20
       V2 20 20
       V3 20 20
> boxplot(Stress~Vibration*Bruit,ylab="Stress")  # --->  Voir Figure 2 (gauche)
> interaction.plot(Vibration,Bruit,Stress)  # --->  Voir Figure 2 (droite)

> # Modele avec interaction
> #=========================
> resA1 <- lm(Stress~Vibration*Bruit)    # ---> *** Ligne A1 ***
> anova(resA1)
Analysis of Variance Table

Response: Stress
                Df  Sum Sq Mean Sq F value    Pr(>F)
Vibration        2  47.580  23.790 23.8627 2.204e-09 ***
Bruit            1  90.403  90.403 90.6794 3.607e-16 ***
Vibration:Bruit  2   2.915   1.457  1.4617    0.2361    # ---> *** Ligne A2 ***
Residuals      114 113.652   0.997
```

```
> # Modele sans interaction
> #=========================
> resA2 <- lm(Stress~Vibration+Bruit)
> anova(resA2)
Analysis of Variance Table

Response: Stress
           Df  Sum Sq Mean Sq F value    Pr(>F)
Vibration   2  47.580  23.790  23.674 2.387e-09 *** # ---> *** Ligne A3 ***
Bruit       1  90.403  90.403  89.963 3.826e-16 *** # ---> *** Ligne A4 ***
Residuals 116 116.567   1.005

> summary(resA2)
Call:
lm(formula = Stress ~ Vibration + Bruit)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.14488    0.18302  44.503  < 2e-16 ***
VibrationV2 -1.32085    0.22415  -5.893 3.80e-08 ***
VibrationV3  0.02932    0.22415   0.131    0.896    # ---> *** Ligne A6 ***
BruitB2     -1.73592    0.18302  -9.485 3.83e-16 ***   # ---> *** Ligne A7 ***
---
Residual standard error: 1.002 on 116 degrees of freedom
Multiple R-squared:  0.5421,Adjusted R-squared:  0.5302
F-statistic: 45.77 on 3 and 116 DF,  p-value: < 2.2e-16    # ---> *** Ligne A5 ***

> shapiro.test(resA2$residuals)    # ---> *** Ligne A8 ***
Shapiro-Wilk normality test

data:  resA2$residuals
W = 0.99115, p-value = 0.6409

> bartlett.test(Stress,Vibration,Bruit)    # ---> *** Ligne A9 ***
Bartlett test of homogeneity of variances

data:  Stress and Vibration
Bartlett's K-squared = 1.1226, df = 2, p-value = 0.5705
```

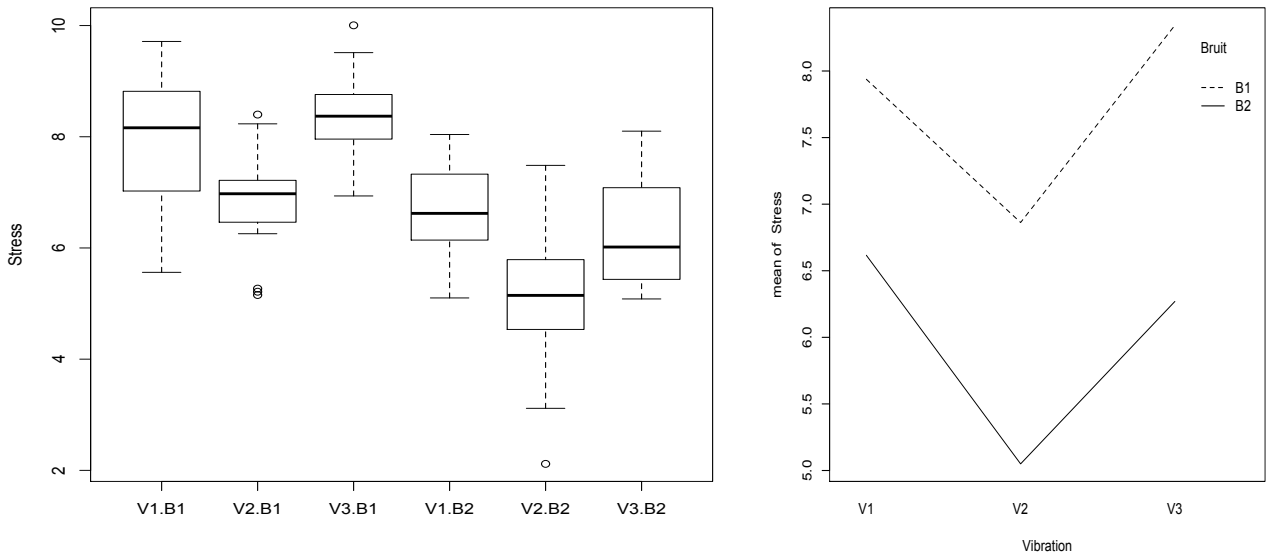

Figure 2: A gauche : boxplots de la variable `Stress` en fonction des facteurs `Vibration` et `Bruit`. A droite : graphique des interactions entre les facteurs `Vibration` et `Bruit`.

1. Ecrire le modèle théorique d'analyse de la variance qui est considéré à la `* ligne A1 *`.
2. Quelles informations intéressantes peut-on retirer des deux graphiques de la Figure 2 ?
3. Quelles sont les hypothèses H_0 et H_1 testées à la `* ligne A2 *` ? Préciser la conclusion de ce test et justifier le passage au second modèle d'ANOVA considéré.
4. Quelles sont les hypothèses H_0 et H_1 testées à la `* ligne A3 *` ? Préciser la conclusion de ce test.
5. Quelles sont les hypothèses H_0 et H_1 testées à la `* ligne A4 *` ? Préciser la conclusion de ce test.
6. Donner une estimation sans biais de la variance du terme d'erreur du modèle d'ANOVA sous-jacent.
7. Quelles sont les hypothèses H_0 et H_1 testées à la `* ligne A5 *` ? Préciser la conclusion de ce test.
8. Quelles sont les hypothèses H_0 et H_1 testées à la `* ligne A6 *` ? Préciser la conclusion de ce test.
9. Quelles sont les hypothèses H_0 et H_1 testées à la `* ligne A7 *` ? Préciser la conclusion de ce test.
10. Justifier pourquoi il est important de faire le test de normalité des résidus (à la `* ligne A8 *`).
11. Quelles sont les hypothèses H_0 et H_1 testées à la `* ligne A9 *` ? Préciser la conclusion de ce test.
12. La cogniticienne décide de retenir ce modèle d'ANOVA. Que pensez-vous de ce modèle ? Justifier votre réponse.
13. a) Donner le modèle final estimé (sous forme d'un tableau synthétique par exemple).

    b) En déduire une estimation du score moyen de stress pour les niveaux `V2` et `B1` des facteurs `Vibration` et `Bruit`.

    c) Même question pour les niveaux `V1` et `B2` des facteurs `Vibration` et `Bruit`.

    d) Ce dernier score moyen (conditions `V1` et `B2`) diffère-t-il significativement du score moyen de stress estimé dans les conditions `V3` et `B2` des facteurs `Vibration` et `Bruit`. Justifier votre réponse.