Examen de Modélisation Statistique

7 janvier 2019 (durée 2h30)

La qualité de la rédaction sera prise en compte dans la notation.

Les données étudiées proviennent du site web `statlib`. Une description du contexte de l'étude et des variables statistiques mesurées est fournie ci-dessous (en anglais)

```
Description:  This datafile contains 315 observations on 14 variables.
Variable Names in order from left to right:
AGE: Age (years)
SEX: Sex (1=Male, 2=Female).
FUMEUR: Smoking status (1=Never, 2=Former, 3=Current Smoker)
QUETELET: Quetelet (weight/(height^2))
VITAMINE: Vitamin Use (1=Yes, fairly often, 2=Yes, not often, 3=No)
CALORIES: Number of calories consumed per day.
GRAS: Grams of fat consumed per day.
FIBRE: Grams of fiber consumed per day.
ALCOOL: Number of alcoholic drinks consumed per week.
CHOLESTEROL: Cholesterol consumed (mg per day).
BETADIET: Dietary beta-carotene consumed (mcg per day).
RETDIET: Dietary retinol consumed (mcg per day)
BETAPLASMA: Plasma beta-carotene (ng/ml)
RETPLASMA: Plasma Retinol (ng/ml)
```

Partie 1. Variation des calories consommées quotidiennement en fonction d'autres facteurs

1. Le statisticien effectue les analyses suivantes

```
res1 <- lm(CALORIE~(SEX + VITAMINE + FUMEUR) ,data=plasma)
res2 <- lm(CALORIE~(SEX + VITAMINE + FUMEUR)^2 ,data=plasma)
anova(res1,res2)

Analysis of Variance Table

Model 1: CALORIE ~ (SEX + VITAMINE + FUMEUR)
Model 2: CALORIE ~ (SEX + VITAMINE + FUMEUR)^2
  Res.Df       RSS Df Sum of Sq      F   Pr(>F)
1    309 137007932
2    301 126829945  8  10177987 3.0194 0.002817 **
```

Expliquer sa démarche (en particulier les modèles qui ont été estimés et les hypothèses sur ces modèles). Décrire le résultat obtenu avec la fonction `anova` (l'hypothèse nulle et l'hypothèse alternative, la statistique de test, sa loi sous H_0 et la conclusion du test).

2. Il effectue ensuite l'analyse suivante

```
res3 <- lm(CALORIE~SEX*FUMEUR,data=plasma)
anova(res3, res2)
Analysis of Variance Table

Model 1: CALORIE ~ SEX * FUMEUR
Model 2: CALORIE ~ (SEX + VITAMINE + FUMEUR)^2
  Res.Df       RSS Df Sum of Sq      F Pr(>F)
1    309 130048787
2    301 126829945  8   3218842 0.9549 0.4716
```

Quel modèle lui conseillez vous de garder au final ? Pourquoi la prise de vitamine n'est pas déterminante dans la variation des calories consommées ?

3. Le statisticien examine ensuite les cœfficients estimés pour le modèle qu'il a retenu.

```
summary(res3)
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         1903.77     179.93  10.581  < 2e-16 ***
SEXF                -207.57     187.88  -1.105 0.270095
FUMEURAncienF         92.65     226.95   0.408 0.683380
```

```
FUMEURFumeur        1220.92     304.14   4.014 7.49e-05 ***
SEXF:FUMEURAncienF    40.35     242.80   0.166 0.868136
SEXF:FUMEURFumeur  -1221.69     327.28  -3.733 0.000225 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 648.7 on 309 degrees of freedom
Multiple R-squared:  0.1052,Adjusted R-squared:  0.09074
F-statistic: 7.267 on 5 and 309 DF,  p-value: 1.876e-06
```

Quelle est la valeur de calorie ajustée par le modèle pour un homme, non fumeur et pour une femme qui est fumeuse.
Représentez sur un graphique les valeurs ajustées pour toutes les valeurs possibles des modalités des deux facteurs. Comment se traduit l'effet d'interaction entre les facteurs ?

4. Rappelez la définition du cœfficient R^2. Que mesure-t-il ? La valeur calculée pour le modèle `res3` vous semble-t-elle acceptable ?

Partie 2. Prise en compte d'autres variables

Le statisticien décide de prendre en compte des variables quantitatives pour modéliser les calories consommées quotidiennement. Il choisit quelques variables explicatives et ajuste le modèle suivant

```
res4 <- lm(CALORIE ~ AGE + VITAMINE + GRAS + FIBRE + ALCOOL + CHOLESTEROL, data = plasma)
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     381.3871    55.3284   6.893 3.11e-11 ***
AGE              -3.8428     0.7607  -5.052 7.52e-07 ***
VITAMINErarement -14.5450    27.3845  -0.531  0.59571
VITAMINEjamais   -73.5482    25.1526  -2.924  0.00371 **
GRAS              13.4892     0.4699  28.706  < 2e-16 ***
FIBRE             35.0331     2.1176  16.543  < 2e-16 ***
ALCOOL            18.1905     0.8957  20.309  < 2e-16 ***
CHOLESTEROL        0.3748     0.1154   3.249  0.00129 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 189.2 on 307 degrees of freedom
Multiple R-squared:  0.9244,Adjusted R-squared:  0.9227
F-statistic: 536.3 on 7 and 307 DF,  p-value: < 2.2e-16
```

1. Décrire le modèle `res4` ajusté sur ces données ?

2. Ce modèle vous semble-t-il préférable au modèle `res3` ?

3. Que représente le graphique de la Figure 1. Ce graphique vous permet-il de conforter l'idée que le modèle `res4` est bien adapté ?

4. Le statisticien décide de modéliser le logarithme de la variable `CALORIE`. Comment peut-on justifier ce choix à partir des statistiques descriptives suivantes :

```
summary(plasma$CALORIE)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  445.2  1338.0  1666.8  1796.7  2100.4  6662.2
```

Il execute les commandes suivantes et obtient la Figure 2.

```
res5.0 <- lm(log(CALORIE)~.,data=plasma)
res.5 <- step(res5.0)
plot(res.5$fitted,res.5$residuals)
abline(h=0)
```

Expliquez sa démarche. Commentez la Figure 2.

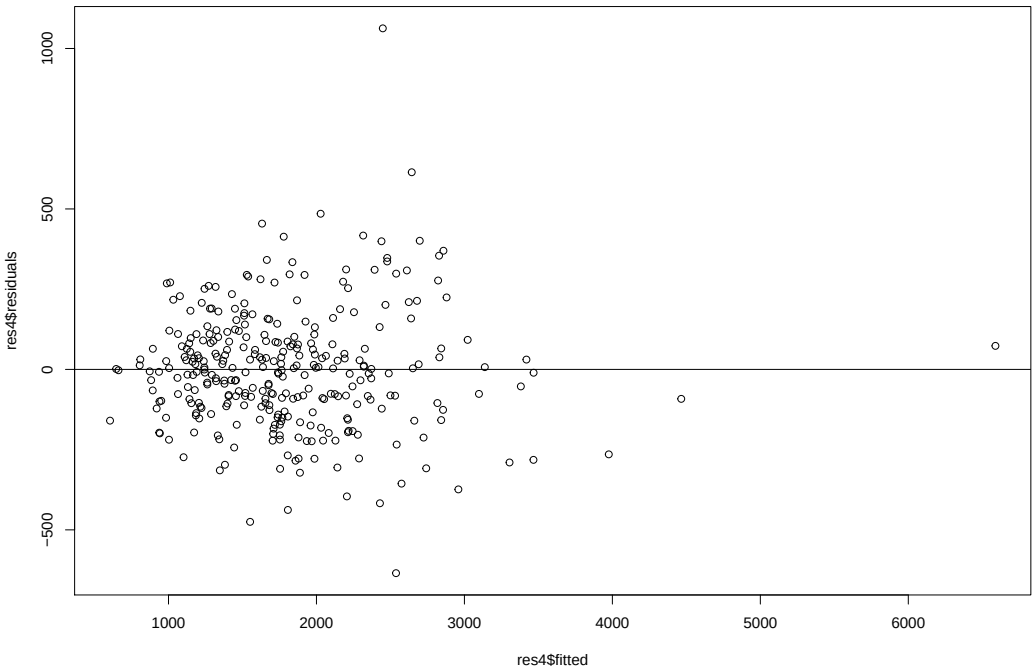

FIGURE 1 –

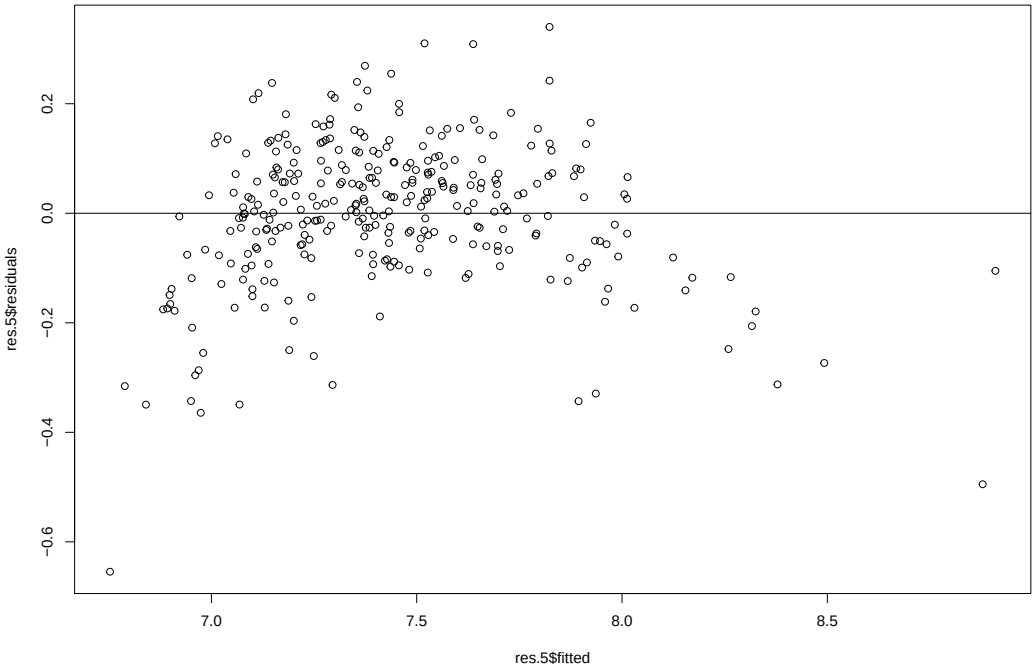

FIGURE 2 –

Partie 3. Un modèle linéaire généralisé

Une collègue statisticienne lui conseille de considérer plutôt un modèle linéaire généralisé où la variable `CALORIE` suivrait une loi exponentielle.

1. On suppose dans cette question que $Y_1, \ldots, Y_n$ sont issus d'une loi exponentielle de paramètre $\lambda > 0$, de densité $f(y, \lambda) = \frac{1}{\lambda} \exp(-y/\lambda)$, si $y \geq 0$ et $f(y, \lambda) = 0$ si $y < 0$. Vérifier qu'il s'agit bien d'une densité et que $\mathbb{E}(Y) = \lambda$. Ecrire la vraisemblance et calculer le score. En déduire l'expression de l'estimateur du maximum de vraisemblance de λ.

2. On note maintenant $Y_1, \ldots, Y_n$ les réalisations de la variable `CALORIE` et X_{ij} la valeur de la $j$ème variable observée sur l'individu i, pour $j = 1, \ldots, p$. On suppose que Y_i suit une loi exponentielle de paramètre λ_i où

$$\ln \lambda_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}$$

De quel type de modèle s'agit-il ? Quelle est la fonction lien ?
Ecrire la vraisemblance, le score et l'information de Fisher.

3. Sous quelle condition (classique) l'estimateur du maximum de vraisemblance est-il unique (justifiez rigoureusement votre réponse) ?

4. Le modèle suivant (appelé modèle `res.8`) est retenu par le statisticien.

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          6.211e+00  6.560e-02  94.679  < 2e-16 ***
AGE                 -2.468e-03  1.012e-03  -2.439 0.015327 *
GRAS                 1.341e-02  9.115e-04  14.707  < 2e-16 ***
FIBRE                3.986e-02  3.495e-03  11.406  < 2e-16 ***
ALCOOL               2.299e-02  4.018e-03   5.722 2.55e-08 ***
CHOLESTEROL         -5.044e-04  3.233e-04  -1.560 0.119845
RETDIET              3.511e-04  7.304e-05   4.808 2.42e-06 ***
AGE:CHOLESTEROL      7.559e-06  3.921e-06   1.928 0.054808 .
AGE:RETDIET         -1.914e-06  1.104e-06  -1.733 0.084126 .
GRAS:FIBRE          -3.299e-04  6.379e-05  -5.173 4.23e-07 ***
GRAS:ALCOOL         -5.207e-05  1.721e-05  -3.025 0.002705 **
GRAS:RETDIET        -1.789e-06  4.804e-07  -3.725 0.000234 ***
FIBRE:ALCOOL        -9.788e-04  2.805e-04  -3.489 0.000557 ***
FIBRE:CHOLESTEROL    3.486e-05  1.682e-05   2.073 0.039040 *
CHOLESTEROL:RETDIET -1.482e-07  7.547e-08  -1.963 0.050546 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4.a. A votre avis selon quelle démarche le statisticien a pu aboutir à ce modèle ?

4.b. Expliquez comment est calculée la valeur ajustée pour la variable `CALORIE` pour l'individu 1. Quelle est *a priori* l'effet de l'age sur le nombre de calories consommées ?

	AGE	SEX	FUMEUR	QUETELET	VITAMINE	CALORIE	GRAS	FIBRE	ALCOOL	CHOLESTEROL	BETADIET	RETDIET	BETAPLASMA	RETPLASMA
1	64	F	AncienF	21.48380	souvent	1298.8	57.0	6.3	0.0	170.3	1945	890	200	915
2	76	F	NonFumeur	23.87631	souvent	1032.5	50.1	15.8	0.0	75.8	2653	451	124	727

4.c. Commentez les résultats ci dessous

```
> summary(abs(plasma$CALORIE - res.8$fitted.values))
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
  0.0009  53.2551 108.0930 143.6093 198.3662 993.4037
> summary(abs(exp(res.5$fitted.values)- plasma$CALORIE))
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
   0.5867   61.1656  133.7596  181.1431  217.8701 2800.5489
```

4.d. Quelles analyses supplémentaires du modèle `res.8` pourrait-on effectuer pour vérifier qu'il s'agit d'un modèle bien adapté à ces données ?