Ecole Centrale de Lyon
Statistique appliquée aux sciences de l'ingénieur

Examen - 20 décembre - 8h - 10h
Une feuille A4 recto verso autorisée
C.Helbert

1 Exercices théoriques

1.1 Exercice 1

Soit un modèle d'analyse de la variance à un facteur à I modalités, vérifiant les hypothèses suivantes : les v.a ϵ_{ij} sont indépendantes et de loi $\mathcal{N}(0, \sigma_i^2)$ pour tout $j = 1, .., n_i$. La variance des erreurs dépend donc de la modalité considérée.

1. Déterminer les estimateurs $\hat{\mu}_i$ et $\hat{\sigma}_i^2$ par maximum de vraisemblance des différents paramètres du modèle.
2. Si on suppose que $\mu_i = \mu$, que valent les $\hat{\mu}$ et $\hat{\sigma}_i^2$?

1.2 Exercice 2

On cherche à estimer le modèle suivant : $Y = \beta_0 + \beta_1 Z + \beta_2 Z^2 + \epsilon$ où ϵ est de loi normale de paramètre $(0, \sigma^2)$, σ^2 connu et où $Z \in [-1, 1]$. On choisit un plan d'expériences à six expériences : 2 expériences en -1, 2 autres en 1 et deux dernières en a.
Définition : un plan est D-optimal si le déterminant de la matrice de covariance de $\hat{\beta}$ est minimal, ou encore le déterminant de l'inverse de la matrice de covariance de $\hat{\beta}$ est maximal.

1. Ecrire la matrice X du modèle en fonction de a .
2. Donner la matrice de covariance de $\hat{\beta}$.
3. Comment choisir a de sorte que le plan soit D - optimal ?

D-optimalité: Choisir X de sorte que le volume de l'ellipsode de confiance soit le plus petit possible, i.e. min det((XtX)^1) = arg min 0 × ... × p

2 Exercice 3 : Poids des bébés

On s'intéresse à l'impact de plusieurs variables sur le poids du bébé à la naissance. Les données utilisées dans cette étude sont basées sur un échantillon de 5% de toutes les naissances survenues à Philadelphie en 1990. L'échantillon est composé de 1115 observations sur cinq variables :

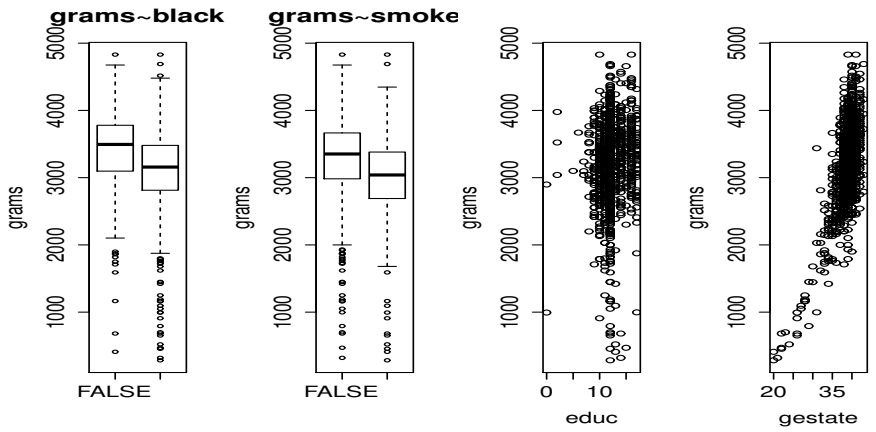
- black = 1 si la mère est noire, 0 sinon,
- educ = nombre d'années de scolarité de la mère, entre 0 et 17,
- smoke = 1 si la mère fume, 0 sinon,
- gestate = nombre total de semaines de gestation (grossesse),
- grams = poids de naissance en grammes.

1. On commence par faire une analyse descriptive des données. Commenter les sorties R ci-dessous. Donner le nom de l'analyse qui permettra l'étude de ces données. Justifier.

Test de Fischer

3 variables quantitatives et 2 binaires

black	educ	smoke	gestate	grams
FALSE:453	Min. : 0.00	FALSE:846	Min. :20.00	Min. : 284
TRUE :662	1st Qu.:11.00	TRUE :269	1st Qu.:38.00	1st Qu.:2900
	Median :12.00		Median :39.00	Median :3267
	Mean :12.27		Mean :38.84	Mean :3220
	3rd Qu.:13.00		3rd Qu.:40.00	3rd Qu.:3630
	Max. :17.00		Max. :43.00	Max. :4830



Avant d'estimer un premier modèle, on commence par centrer les variables **educ** et **gestate**.

centraliser les données

```
> birth$gestate = birth$gestate - 39
> birth$educ = birth$educ - 12
> lm1 = lm(grams ~.,data = birth,contrast =list(black= contr.sum,smoke = contr.sum))
> Anova(lm1, type = "III")
```

```
Anova Table (Type III tests)

Response: grams
                Sum Sq   Df    F value    Pr(>F)
(Intercept) 8053896465    1 42346.6484 < 2.2e-16 ***
black          7306972    1    38.4194 8.031e-10 ***
educ            417786    1     2.1967    0.1386
smoke          5811565    1    30.5567 4.040e-08 ***
gestate      185347966    1   974.5426 < 2.2e-16 ***
Residuals    211110570 1110
---
```

la variable educ ne semble pas d'être significative

linéaire

2. Donner la forme du modèle étudié ci-dessus. Pour la variable *smoke*, donner l'hypothèse testée, la statistique ainsi que le résultat. Quelles informations doit-on retenir de cette sortie de R ?

gestate c'est le parametre le plus importante

que la variable n'a pas d'influence sur le modéle (coeff = 0) à partir du test de student

Un nouveau modèle est alors ajusté. Les commandes et sorties de R sont les suivantes.
> lm1 = lm(grams .-educ,data = birth,contrast =list(black= contr.sum,smoke = contr.sum))
> Anova(lm1, type = "III")

```
Anova Table (Type III tests)

Response: grams
                Sum Sq   Df   F value    Pr(>F)
(Intercept) 8058807127    1 42326.877 < 2.2e-16 ***
black          7927924    1    41.639 1.636e-10 ***
smoke          6833470    1    35.891 2.815e-09 ***
gestate      185497639    1   974.280 < 2.2e-16 ***
Residuals    211528356 1111
```

>summary(lm1)

```
Call:
lm(formula = grams ~ . - educ, data = birth, contrasts = list(black = contr.sum
    smoke = contr.sum))

Residuals:
     Min       1Q   Median       3Q      Max
-1464.13  -295.56     1.86   287.70  1611.83

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3212.865     15.617 205.735  < 2e-16 ***
black1        87.201     13.514   6.453 1.64e-10 ***
smoke1        92.508     15.441   5.991 2.82e-09 ***
gestate      156.570      5.016  31.213  < 2e-16 ***
```

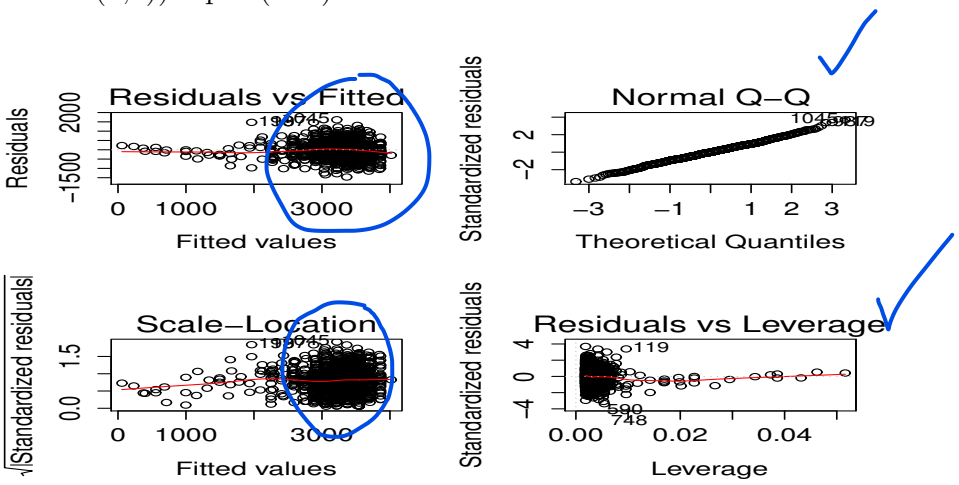
Degrees of Freedom: Number of observations minus the number of coefficients (including intercepts). The larger this number is the better and if it's close to 0, your model is seriously over fit

```
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1
```

racin de SSE — 1115 linhas

```
Residual standard error: 436.3 on 1111 degrees of freedom
Multiple R-squared: 0.5269,         Adjusted R-squared: 0.5256
F-statistic: 412.4 on 3 and 1111 DF,  p-value: < 2.2e-16
```

>par(mfrow=c(2,2)) >plot(lm1)



coeffs = 0 — statistique de Fischer

3. Analyser le test de non-régression (préciser l'hypothèse, la statistique et le résultat). on peut réjeter l'hypothèse nulle
4. Rappeler les hypothèses du modèle. Sont-elles vérifiées ?
5. Donner l'équation du modèle ajusté (indication : *black*1 et *smoke*1 correspondent aux modalités "FALSE"). Interpréter tous les éléments de la ligne *gestate* (notamment on rappellera l'hypothèse et la statistique du test réalisé).
6. Quelle est la part de variance expliquée par ce modèle ? 436.3²
7. D'après ce modèle, quel poids de naissance peut-on prévoir pour les deux individus suivants ? Sans faire de calcul et de facon approximative, donner un intervalle de prédiction à 95% pour ces deux individus.

hypothèses du modèle :
1. E(ei) = 0 et V(ei) = sigma²
2. i != j -> cov(i,j) = 0

z = 1,96

```
> newdata
  black educ smoke gestate
1 FALSE    0  TRUE      40
2  TRUE    0 FALSE      32
```

Y = 3212,865 + black*87,201 + smoke*92,508 + gestate*156,570

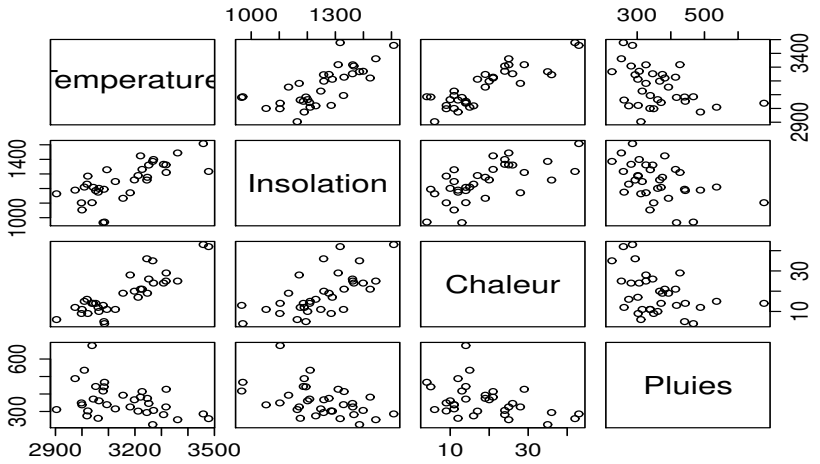
3 Exercice 4 : qualité du vin

On étudie la sensibilité de la qualité du vin aux conditions climatiques pour 34 années de 1924 à 1957. Les variables de l'étude sont les suivantes (elles correspondent à des sommes de quantités journalières sur toute la saison) :
- la **Temperature**
- l'**Insolation** (rayonnement solaire)
- la **Chaleur** (transfert d'énergie)
- les **Pluies** (hauteur d'eau)
- la **QUALITE** du vin

On commence par faire une rapide étude descriptive des données.

données qualitatives

```
> summary(vin)
 Temperature      Insolation      Chaleur         Pluies          QUALITE
 Min.   :2904   Min.   : 966   Min.   : 4.00   Min.   :225.0   bon      :11
 1st Qu.:3045   1st Qu.:1178   1st Qu.:11.25   1st Qu.:302.2   mediocre:12
 Median :3140   Median :1253   Median :16.50   Median :342.5   moyen    :11
 Mean   :3158   Mean   :1247   Mean   :18.82   Mean   :360.4
 3rd Qu.:3251   3rd Qu.:1330   3rd Qu.:24.75   3rd Qu.:408.8
 Max.   :3478   Max.   :1508   Max.   :43.00   Max.   :677.0
```



1. Quel est le nom de l'analyse qui permet l'étude de ces données ?

3.1 Première étude

Dans cette étude on ne s'intéresse qu'à une partie des données : les vins de qualité **médiocre** et **moyen**. Sous R la modalité de référence (événement Y = 0) est celle dont la fréquence dans la population est la plus élevée.

2. Quel est l'événement modélisé dans cette étude ? Proposer un modèle en fonction de $x = (Temperature, Insolation, Chaleur, Pluies)$. Combien de paramètres a-t-on à estimer ?
3. Analyser la table d'analyse de la variance ci-dessous (on prendra un niveau de test à 5%). Rappeler l'hypothèse, la statistique et le résultat. Ce modèle convient-il ? Justifier.

```
> Anova(mod1,type = "III",test.statistic= "LR")
Analysis of Deviance Table (Type III tests)

Response: QUALITE
            LR Chisq Df Pr(>Chisq)
Temperature   5.5071  1   0.018939 *
Insolation    0.1278  1   0.720679
Chaleur       3.5037  1   0.061232 .
Pluies        7.3275  1   0.006791 **
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1
Message d'avis :
glm.fit: fitted probabilities numerically 0 or 1 occurred
```

4. Proposer deux procédures possibles pour sélectionner un meilleur modèle.

A l'issue de la sélection de variables, on obtient le modèle ci-dessous.

```
> Anova(mod2,type = "III",test.statistic= "LR")
Analysis of Deviance Table (Type III tests)

Response: QUALITE
           LR Chisq Df Pr(>Chisq)
Insolation  11.8536  1  0.0005755 ***
Pluies       4.6571  1  0.0309246 *
---
```

```
> summary(mod2)

Call:
glm(formula = QUALITE ~ Insolation + Pluies, family = binomial,
    data = vin1)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.23538  -0.36320  -0.01588   0.33553   2.30295

Coefficients:
```

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -29.66896   17.57359  -1.688   0.0914 .
Insolation    0.03056    0.01503   2.033   0.0421 *
Pluies       -0.01917    0.01138  -1.685   0.0920 .
---
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 31.841  on 22  degrees of freedom
Residual deviance: 12.973  on 20  degrees of freedom
AIC: 18.973

Number of Fisher Scoring iterations: 6
```

5. (a) Préciser l'hypothèse et la statistique pour chaque sortie de R (*Anova* et *summary*).
   (b) Les variables sélectionnées sont-elles en accord avec la sortie *Anova* sur le modèle précédent. Expliquer.
   (c) Quelle information obtient-on à partir de la colonne *estimate* de la commande *summary* ?
   (d) Donner l'équation du modèle.
6. Comment appelle-t-on la sortie R ci-dessous ? Quelle information apporte-t-elle dans le cas présent ? Qu'aurait-on du faire pour confirmer ce résultat ?

```
> pred <- predict(mod2,vin1[,1:5])
> table(pred>0,vin1$QUALITE=="moyen")

       FALSE TRUE
FALSE     11    1
TRUE       1   10
```

3.2 Deuxième étude

On recommence une étude similaire avec les vins de QUALITE **moyen** et **bon**. La modalité de référence (événement Y = 0) est le vin de qualité **bon**. Les résultats obtenus sont les suivants.

```
> mod3 <- glm(QUALITE~Temperature+ Insolation+Chaleur+Pluies,data = vin2,
                family = binomial)
> anova(mod3,test= "Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: QUALITE

Terms added sequentially (first to last)
```

```
           Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                         21     30.498
Temperature  1  13.2167        20     17.282 0.0002775 ***
Insolation   1   0.4965        19     16.785 0.4810396
Chaleur      1   0.4413        18     16.344 0.5064922
Pluies       1   1.6362        17     14.708 0.2008457
---

> Anova(mod3,type = "III",test.statistic= "LR")
Analysis of Deviance Table (Type III tests)

Response: QUALITE
            LR Chisq Df Pr(>Chisq)
Temperature  1.37780  1     0.2405
Insolation   0.37876  1     0.5383
Chaleur      0.03698  1     0.8475
Pluies       1.63621  1     0.2008

> summary(mod3)

Call:
glm(formula = QUALITE ~ Temperature + Insolation + Chaleur +
    Pluies, family = binomial, data = vin2)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.77523  -0.40536   0.06206   0.39708   1.55628

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 60.181756  44.629550   1.348    0.178
Temperature -0.017454   0.016371  -1.066    0.286
Insolation  -0.007039   0.011623  -0.606    0.545
Chaleur     -0.028920   0.151776  -0.191    0.849
Pluies       0.017491   0.014816   1.181    0.238

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 30.498  on 21  degrees of freedom
Residual deviance: 14.708  on 17  degrees of freedom
AIC: 24.708

Number of Fisher Scoring iterations: 6

> mod3 = glm(QUALITE~1,family=binomial,data=vin2)
```

```
> mod3 <- step(mod3, direction="forward",
        scope=list(lower=~1,upper=~Temperature+ Insolation+Chaleur+Pluies),
                   k = log(nrow(vin2)))

> Anova(mod3,type = "III",test.statistic= "LR")
Analysis of Deviance Table (Type III tests)

Response: QUALITE
            LR Chisq Df Pr(>Chisq)
Temperature   13.217  1  0.0002775 ***
---

> summary(mod3)

Call:
glm(formula = QUALITE ~ Temperature, family = binomial, data = vin2)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.44909  -0.74071   0.02799   0.42473   2.07174

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 70.99880   32.53633   2.182   0.0291 *
Temperature -0.02201    0.01006  -2.188   0.0287 *
---
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 30.498  on 21  degrees of freedom
Residual deviance: 17.282  on 20  degrees of freedom
AIC: 21.282

Number of Fisher Scoring iterations: 6
```

7. Analyser et interpréter l'ensemble des résultats présentés ci-dessus.
8. En analysant le résultat ci-dessous, est-il plus facile de distinguer un vin **mediocre** d'un vin **moyen** ou un vin **bon** d'un vin **moyen** ?

```
> pred <- predict(mod3,vin2[,1:5])
> table(pred>0,vin2$QUALITE=="moyen")

       FALSE TRUE
  FALSE     9    3
  TRUE      2    8
```

9. La prévision de ces deux modèles sur deux nouvelles années est la suivante. Quelle qualité de vin peut-on espérer pour chacune de ces années ?

```
> vin_new
  Temperature Insolation Chaleur Pluies
1        3150       1340      17    340
2        3400       1400      22    280
> pred1 <- predict(mod2,vin_new,type = "response")
> pred2 <- predict(mod4,vin_new, type = "response")
> pred1
        1         2
0.9915390 0.9995686
> pred2
          1          2
0.84231511 0.02132563
```

10. Les études réalisées ci-dessus ont le bon gout d'être simples à interpréter mais une analyse (modélisation) plus pertinente aurait du être faite. Laquelle ?