Feuille de TD 3

Modèle d'ANOVA à un et deux facteurs. Modèle d'Ancova

Exercice 1 (**Estimation du modèle singulier sous contraintes**)**.** On considère le modèle d'ANOVA à facteur dans sa version singulière : $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$. Calculer les estimateurs des paramètres pour les deux contraintes suivantes : $\sum_{i=1}^{I} \alpha_i = 0$ et $\sum_{i=1}^{I} n_i \alpha_i = 0$. On discutera de l'interprétation des α_i dans chaque cas, on comparera à la paramétrisation proposée dans R, i.e. $\alpha_1 = 0$.

Exercice 2 (**Estimation du modèle singulier sous contraintes**)**.** Considérons le modèle linéaire $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ dans le cas où $\boldsymbol{\beta} \in \mathbb{R}^p$ et X une matrice à n lignes et p colonnes *de rang* $r < p$. Soit $\hat{\boldsymbol{\beta}}$ l'estimateur des moindres carrés de $\boldsymbol{\beta}$ est tel que $X\hat{\boldsymbol{\beta}}$ est le projeté orthogonal de $\boldsymbol{Y}$ sur $[X]$: $X\hat{\boldsymbol{\beta}} = P_{[X]}\boldsymbol{Y}$. Alors $\hat{\boldsymbol{\beta}}$ vérifie les équations suivantes (dites normales) :

$$X'\boldsymbol{Y} = X'X\hat{\boldsymbol{\beta}}$$

1. A-t-on unicité de la solution des équations normales ?

Afin de trouver une expression pour $\hat{\boldsymbol{\beta}}$ on propose d'imposer des contraintes. Soit H est une matrice de dimension $p - r \times p$. On impose $p - r$ contraintes linéaires sur $\boldsymbol{\beta}$ et $\hat{\boldsymbol{\beta}}$ écrites sous la forme $H\boldsymbol{\beta} = H\hat{\boldsymbol{\beta}} = 0_{p-r}$. On cherche donc $\hat{\boldsymbol{\beta}}$ vérifiant :

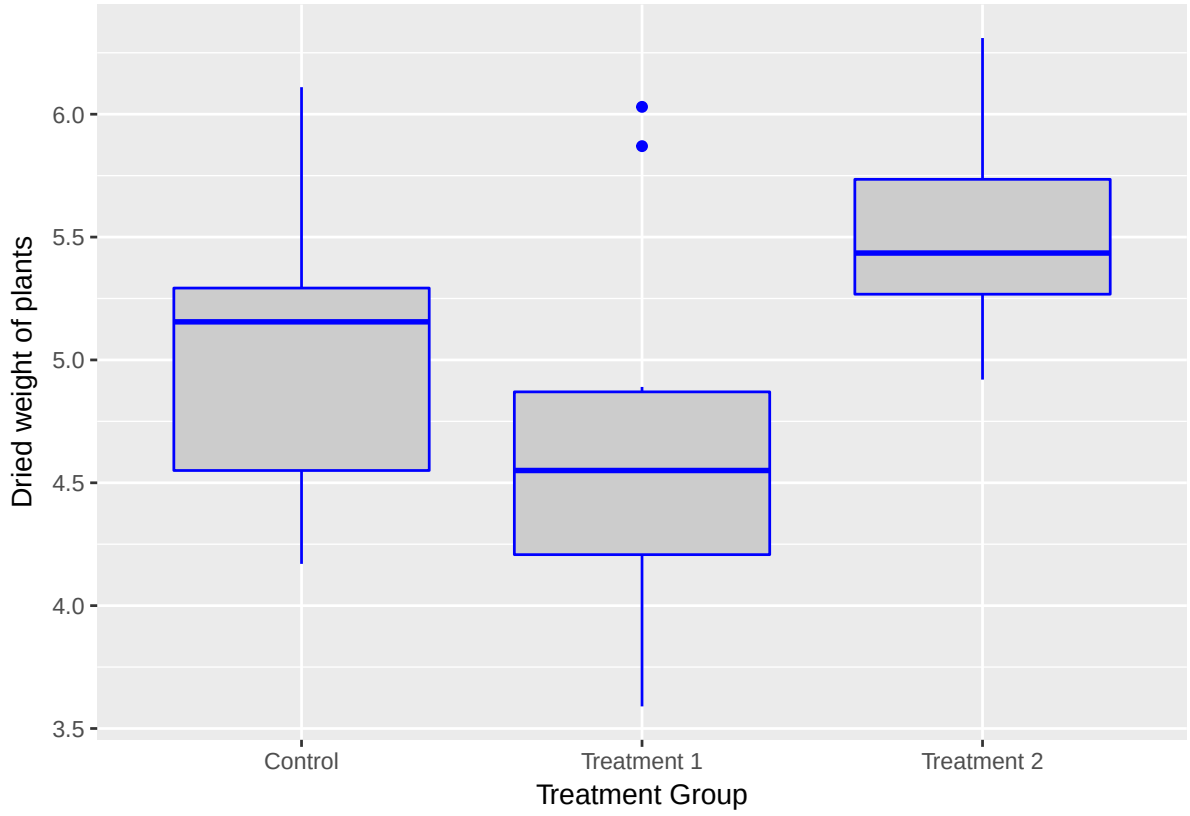
$$X'\boldsymbol{Y} = X'X\hat{\boldsymbol{\beta}} \quad \text{et} \quad H\hat{\boldsymbol{\beta}} = 0_{p-r}$$

Notons $G = \begin{pmatrix} X \\ H \end{pmatrix} \in \mathcal{M}_{n+p-r,p}$. La contrainte est dite *admissible* si $\text{Ker } G = \text{Ker}H \cap \text{Ker}X = \{0_p\}$.

2. Montrer que $G'G\hat{\boldsymbol{\beta}} = X'\boldsymbol{Y}$
3. Montrer que $(G'G)$ est inversible. En déduire une expression pour $\hat{\boldsymbol{\beta}}$.
4. $\hat{\boldsymbol{Y}} = P_{[X]}\boldsymbol{Y}$ dépend-il des contraintes ? En déduire que la matrice $X(G'G)^{-1}X'$ ne dépend pas des contraintes utilisées pour estimer $\boldsymbol{\beta}$.
5. Calculer $\mathbb{E}[\hat{\boldsymbol{\beta}}]$ et $\mathbb{V}(\hat{\boldsymbol{\beta}})$. Donner la loi de $\hat{\boldsymbol{\beta}}$ si le résidus sont gaussiens.
6. Donner un estimateur sans biais de σ^2. Dépend-il des contraintes ? Donner sa loi dans le cas gaussien.
7. Pour le modèle d'Anova à un facteur,
   (a) Montrer que la contrainte $\alpha_1 = 0$ est admissible.
   (b) Ecrire un test de $\alpha_i = 0$ versus $\alpha_i \neq 0$.

Exercice 3 (**Interprétation de sorties R**)**.** On s'intéresse aux résultats d'une expérience cherchant à expliquer un rendement agricole (mesuré en poids sec de plantes) en fonction du traitement (2 traitements différents et un groupe de contrôle). On dispose de 30 observations

réparties de façon équilibrées dans les trois modalités. Les données sont représentées sous la forme de boxplot sur la figure suivante :



Les instructions sont données ci-dessous.

- Instruction 1

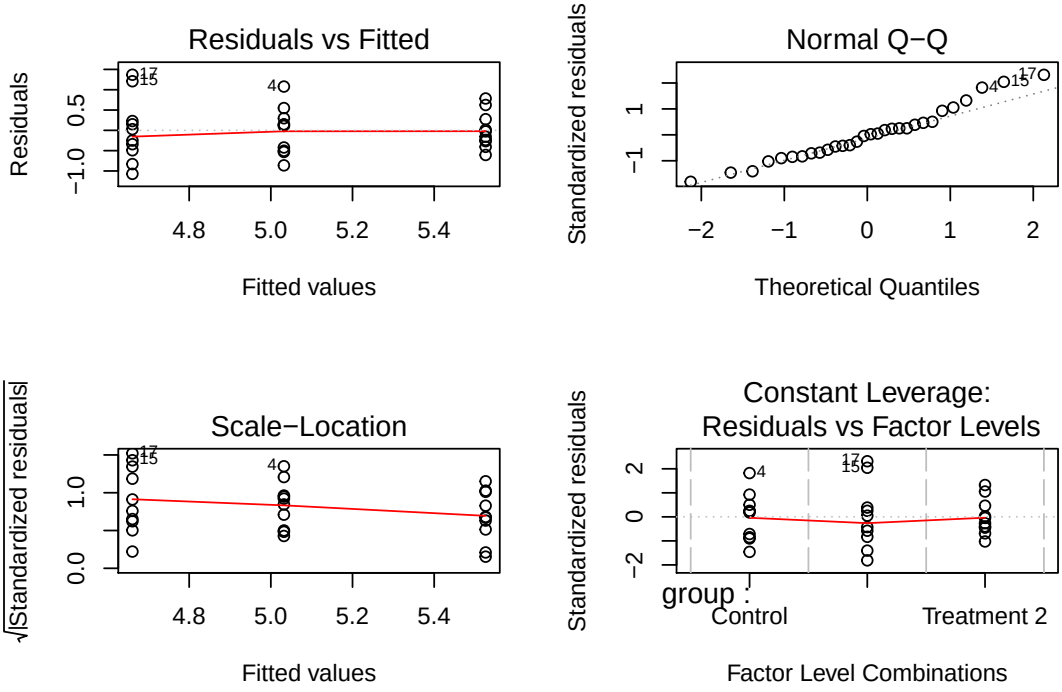
```
plant.mod1 = lm(weight ~ group, data = plant.df)
summary(plant.mod1)
```

```
##
## Call:
## lm(formula = weight ~ group, data = plant.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0710 -0.4180 -0.0060  0.2627  1.3690
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        5.0320     0.1971  25.527   <2e-16 ***
## groupTreatment 1  -0.3710     0.2788  -1.331   0.1944
## groupTreatment 2   0.4940     0.2788   1.772   0.0877 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6234 on 27 degrees of freedom
```

```
## Multiple R-squared:  0.2641, Adjusted R-squared:  0.2096
## F-statistic: 4.846 on 2 and 27 DF,  p-value: 0.01591
```

- Instruction 2

```
par(mfrow=c(2,2))
plot(plant.mod1)
```



- Instruction 3

```
contrasts(plant.df$group) <- contr.sum
plant.mod2 = lm(weight ~ group, data = plant.df)
summary(plant.mod2)
```

```
##
## Call:
## lm(formula = weight ~ group, data = plant.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0710 -0.4180 -0.0060  0.2627  1.3690
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.0730     0.1138  44.573   <2e-16 ***
## group1       -0.0410     0.1610  -0.255   0.8009
## group2       -0.4120     0.1610  -2.560   0.0164 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6234 on 27 degrees of freedom
```

```
## Multiple R-squared:  0.2641, Adjusted R-squared:  0.2096
## F-statistic: 4.846 on 2 and 27 DF,  p-value: 0.01591
```

- Instruction 4

```
anova(plant.mod1)
```

```
## Analysis of Variance Table
##
## Response: weight
##           Df  Sum Sq Mean Sq F value  Pr(>F)
## group      2  3.7663  1.8832  4.8461 0.01591 *
## Residuals 27 10.4921  0.3886
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Instruction 5

```
test.treatment = pairwise.t.test(plant.df$weight,plant.df$group,

                 p.adjust.method="bonferroni")
test.treatment
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  plant.df$weight and plant.df$group
##
##             Control Treatment 1
## Treatment 1 0.583   -
## Treatment 2 0.263   0.013
##
## P value adjustment method: bonferroni
```

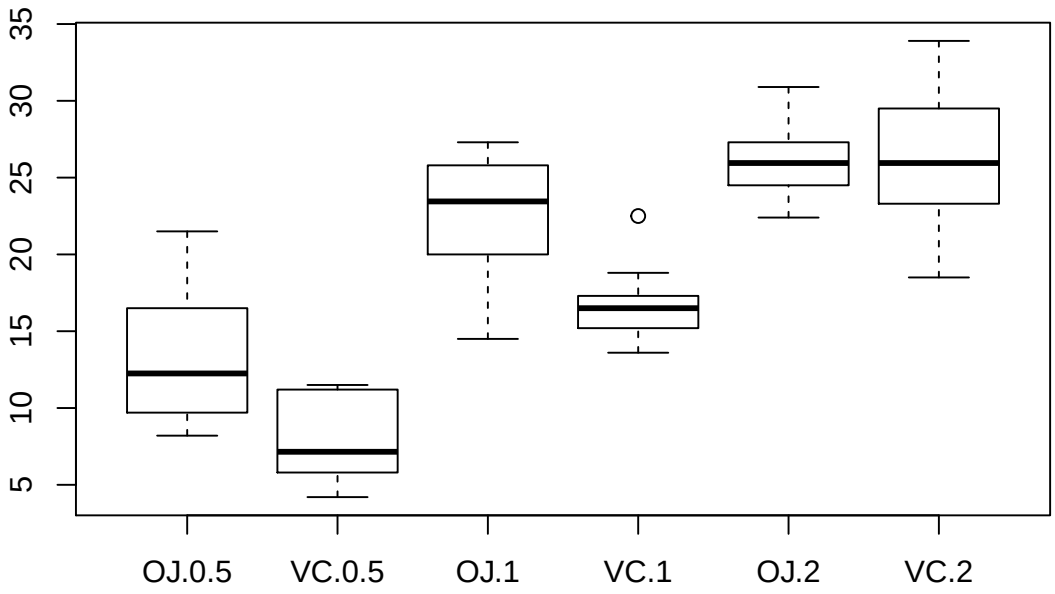
1. Quelle contrainte est utilisée dans Instruction 1 ?
2. Les hypothèses du modèle linéaire sont-elles vérifiées ?
3. Que fait-on dans Instruction 3 ?
4. Rappelez l'expression de la statistique du test du modèle. Rejette-t-on $\mathcal{H}_0$ ici ?
5. Que pensez-vous du R^2.
6. Interpréter les sorties Coefficients.
7. Que fait-on dans l'instruction 5. Rappelez la statistique de test. A quoi correspond l'instruction Bonferroni ? Interprétez les résultats.

Exercice 4 (Analyse d'un jeu de données par un modèle d'ANOVA à 2 facteurs). Nous considérons le jeu de données `ToothGrowth`. La réponse Y est la longueur des odontoblates (cellules intervenant dans la croissance des dents) chez $n = 60$ cochons de Guinée. Chaque animal a reçu une des trois doses possibles de vitamince C (0.5, 1, and 2 mg/day) par le biais d'une des deux méthodes d'administration (jus d'orange ou acide ascorbique). On s'intéresse à l'influence de ces facteurs sur la croissance dentaire. Le jeu de données est représenté par le box-plot suivant. (Pour l'exercice, nous avons transformé la variable quantitative `dose` en un facteur `doselevel`).

```
ToothGrowth$doselevel = as.factor(ToothGrowth$dose)
summary(ToothGrowth)
```

```
##       len        supp         dose       doselevel
##  Min.   : 4.20   OJ:30   Min.   :0.500   0.5:20
##  1st Qu.:13.07   VC:30   1st Qu.:0.500   1  :20
##  Median :19.25           Median :1.000   2  :20
##  Mean   :18.81           Mean   :1.167
##  3rd Qu.:25.27           3rd Qu.:2.000
##  Max.   :33.90           Max.   :2.000
```

```
names(ToothGrowth)=c('len','suppfactor','dose','doselevel')
boxplot(len~suppfactor*doselevel,data=ToothGrowth)
```




```
table(ToothGrowth$doselevel, ToothGrowth$suppfactor)
```

```
##
##       OJ VC
##   0.5 10 10
##   1   10 10
##   2   10 10
```

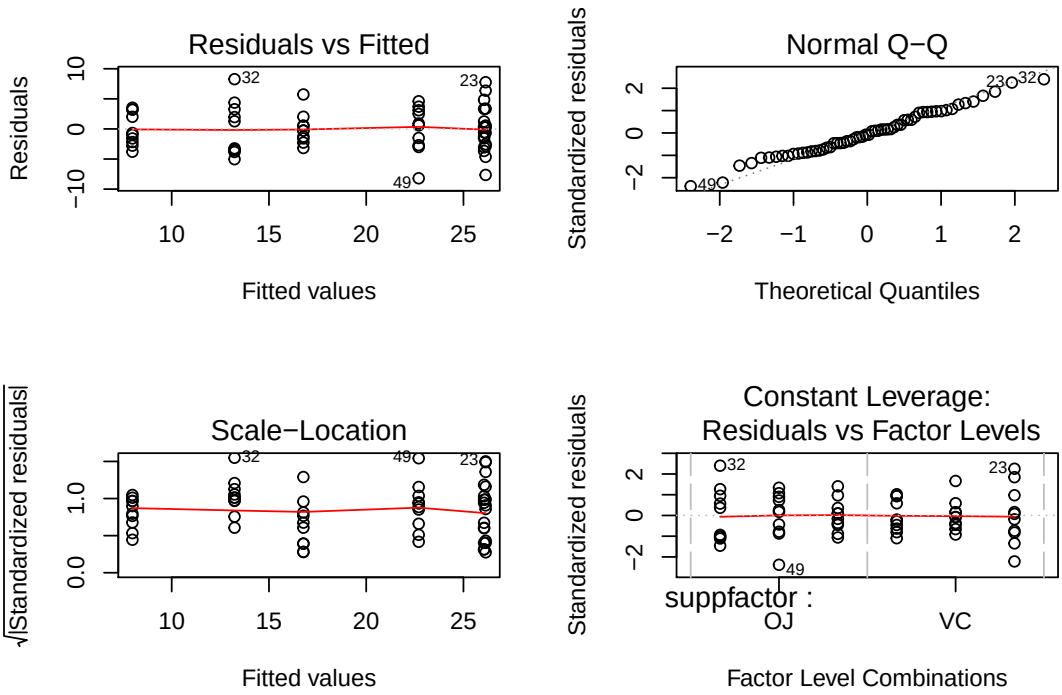
- Instruction 1

```
mod2=lm(len~suppfactor*doselevel,data=ToothGrowth)
summary(mod2)
```

```
##
```

```
## Call:
## lm(formula = len ~ suppfactor * doselevel, data = ToothGrowth)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -8.20  -2.72  -0.27   2.65   8.27
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 13.230      1.148  11.521 3.60e-16 ***
## suppfactorVC                -5.250      1.624  -3.233  0.00209 **
## doselevel1                   9.470      1.624   5.831 3.18e-07 ***
## doselevel2                  12.830      1.624   7.900 1.43e-10 ***
## suppfactorVC:doselevel1     -0.680      2.297  -0.296  0.76831
## suppfactorVC:doselevel2      5.330      2.297   2.321  0.02411 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.631 on 54 degrees of freedom
## Multiple R-squared:  0.7937, Adjusted R-squared:  0.7746
## F-statistic: 41.56 on 5 and 54 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(mod2)
```



- Instruction 2

```
anova(mod2)
```

```
## Analysis of Variance Table
##
```

```
## Response: len
##                      Df  Sum Sq Mean Sq F value    Pr(>F)
## suppfactor            1  205.35  205.35  15.572 0.0002312 ***
## doselevel             2 2426.43 1213.22  92.000 < 2.2e-16 ***
## suppfactor:doselevel  2  108.32   54.16   4.107 0.0218603 *
## Residuals            54  712.11   13.19
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Instruction 3

```
Anova(mod2)
```

```
## Anova Table (Type II tests)
##
## Response: len
##                       Sum Sq Df F value    Pr(>F)
## suppfactor            205.35  1  15.572 0.0002312 ***
## doselevel            2426.43  2  92.000 < 2.2e-16 ***
## suppfactor:doselevel  108.32  2   4.107 0.0218603 *
## Residuals             712.11 54
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1. Écrire le modèle `mod2` correspondant (sans oublier les hypothèses et les gammes de variation des indices).
2. Quelles sont les contraintes utilisées par le logiciel R ?
3. Les hypotèses du modèle linéaire sont elles vérifiées ?
4. A-t-on un effet significatif des facteurs sur la croissance ? Quid de l'interaction ?
5. Pourquoi les résultats des tests de type I et II sont-ils tous les mêmes ?

Exercice 5 (**Modèle d'Ancova**)**.** On cherche maintenant à expliquer une variable Y en fonction d'un facteur et d'une variable quantitative. On note Y_{ij} la j-ième observation dans la modalité i du facteur d'intérêt. On note x_{ij} la valeur de la covariable pour l'individu ij. Le modèle s'écrit

$$Y_{ij} = \mu + \alpha_i + bx_{ij} + c_i x_{ij} + E_{ij}$$

1. Donner une interprétation des paramètres du modèle. On pourra faire un dessin.
2. Écrire le modèle dans sa version régulière. Combien de paramètres doivent être estimés ?. Estimer les paramètres de ce modèle par moindres carrés.
3. Proposer un estimateur sans biais de σ^2.
4. Proposer une contrainte dans le modèle singulier pour le rendre estimable. Proposer alors un estimateur pour les paramètres du modèle singulier.
5. Ecrire un test de l'hypothèse $\mathcal{H}_0$ suivante

$$\mathcal{H}_0 : Y_{ij} = \mu + E_{ij}$$

(on précisera la statistique de test, sa loi et la région de rejet de $\mathcal{H}_0$).

6. On souhaite tester les effets du facteur et de la covariable. Après avoir listé les modèles possibles, proposer des tests de type I et II en spécifiant la statistique de test et sa loi.
7. A quelle condition les deux types de tests sont-ils équivalents ?

Exercice 6 (Ancova en pratique). Nous nous intéressons à la consommation (exprimée en en miles par galon) de voitures en fonction de leur puissance et du type de transmission (automatique ou manuelle).

```
data(mtcars)
attach(mtcars)
don <- mtcars[,c("am","mpg","hp")]
don$am <- as.factor(don$am)
```

Un extrait des données est fourni ci-dessous.

```
head(don)
```

```
##                   am  mpg  hp
## Mazda RX4          1 21.0 110
## Mazda RX4 Wag      1 21.0 110
## Datsun 710         1 22.8  93
## Hornet 4 Drive     0 21.4 110
## Hornet Sportabout  0 18.7 175
## Valiant            0 18.1 105
```

- Instruction 1

```
mod2 = lm(mpg~am*hp,data=don)
summary(mod2)
```

```
##
## Call:
## lm(formula = mpg ~ am * hp, data = don)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3818 -2.2696  0.1344  1.7058  5.8752
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 26.6248479  2.1829432  12.197 1.01e-12 ***
## am1          5.2176534  2.6650931   1.958   0.0603 .
## hp          -0.0591370  0.0129449  -4.568 9.02e-05 ***
## am1:hp       0.0004029  0.0164602   0.024   0.9806
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.961 on 28 degrees of freedom
## Multiple R-squared:  0.782,  Adjusted R-squared:  0.7587
## F-statistic: 33.49 on 3 and 28 DF,  p-value: 2.112e-09
```

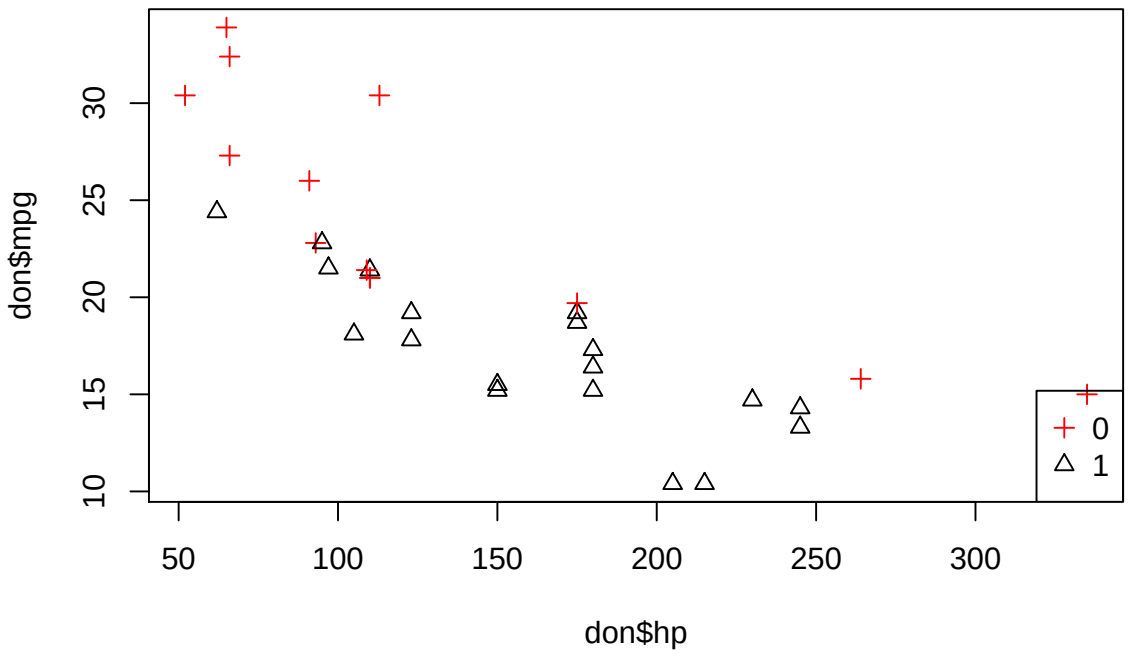


FIGURE 1 – Consommation (en miles par galon) en fonction de la puissance des véhicules

- Instruction 2

```
mod0 = lm(mpg~1,data=don)
anova(mod0,mod2)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ 1
## Model 2: mpg ~ am * hp
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     31 1126.05
## 2     28  245.43  3    880.61 33.488 2.112e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Instruction 3

```
anova(mod2)
```

```
## Analysis of Variance Table
##
## Response: mpg
##           Df Sum Sq Mean Sq F value    Pr(>F)
## am         1 405.15  405.15 46.2210 2.197e-07 ***
## hp         1 475.46  475.46 54.2419 5.088e-08 ***
## am:hp      1   0.01    0.01  0.0006    0.9806
## Residuals 28 245.43    8.77
```

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

- Instruction 4

```
mod2b=lm(mpg~am + hp,data = don)
summary(mod2b)
```

```
##
## Call:
## lm(formula = mpg ~ am + hp, data = don)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3843 -2.2642  0.1366  1.6968  5.8657
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 26.584914   1.425094  18.655  < 2e-16 ***
## am1          5.277085   1.079541   4.888 3.46e-05 ***
## hp          -0.058888   0.007857  -7.495 2.92e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.909 on 29 degrees of freedom
## Multiple R-squared:  0.782,  Adjusted R-squared:  0.767
## F-statistic: 52.02 on 2 and 29 DF,  p-value: 2.55e-10
```

- Instruction 5

```
anova(mod2b)
```

```
## Analysis of Variance Table
##
## Response: mpg
##           Df Sum Sq Mean Sq F value    Pr(>F)
## am         1 405.15  405.15  47.871 1.327e-07 ***
## hp         1 475.46  475.46  56.178 2.920e-08 ***
## Residuals 29 245.44    8.46
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1. Écrire le modèle `mod2` correspondant (sans oublier les hypothèses et les gammes de variation des indices).
2. Quelles sont les contraintes utilisées par le logiciel R ?
3. Sur la ligne `am1:hp`, interpréter la valeur 0.0004029 dans la sortie `summary` dans l'instruction 1. .
4. Justifier le passage au modèle `mod2b`.
5. Le type de transmission a-t-il une influence sur la consommation (à justifier soigneusement) ? Détailler les hypothèses comparées par le test que vous utilisez.
6. Donner une estimation de la consommation moyenne (miles / galon) pour un véhicule manuel de puissance 150 ? Même question pour un véhicule automatique ?

Exercice 7 (**Orthogonalité d'un design expérimental. Exercice facultatif**). On considère un modèle d'ANOVA à deux facteurs :

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + E_{ijk}$$

avec $i = 1, \dots, I$, $j = 1, \dots, J$, et $k = 1, \dots, n_{ij}$. On cherche à déterminer les conditions sur les (n_{ij}) sous lequels les tests de type I sont égaux aux tests de type II. On reprend les notations du cours.

1. Montrer que
$$R(\alpha|\mu) = \|P_{[A]+[I]}\boldsymbol{Y} - P_{[I]}\boldsymbol{Y}\|^2$$
et
$$R(\alpha|\mu,\beta) = \|P_{[I]+[A]+[B]}\boldsymbol{Y} - P_{[I]+[B]}\boldsymbol{Y}\|^2$$
2. Notons $A\backslash I$ le supplémentaire orthogonal de $[I]$ dans $[A]$. Notons $A+B\backslash B$ le supplémentaire orthogonal de $[B]$ dans $[A]+[B]$. Montrer que $R(\alpha|\mu) = R(\alpha|\mu,\beta)$ si et seulement si
$$\|P_{A\backslash I}\boldsymbol{Y}\|^2 = \|P_{A+B\backslash B}\boldsymbol{Y}\|^2$$
3. Montrer que l'égalité des projecteurs équivaut à $A\backslash I = A + B\backslash B$.
4. En déduire que si $\forall (i,j), n_{ij} = \frac{n_{i+}n_{+j}}{n}$, alors les tests de type I et II sont équivalents.