Approche par comparaison de modèles

Rémi Courset
remi.courset@univ-grenoble-alpes.fr

Données & exercices du TD :
webcom.upmf-grenoble.fr/LIP/Perso/DMuller/M2R/ACM/TD

Sous R

- **Télécharger les modules dans le dossier de travail**
  - *PRE.R*
  - *outliersFunction.R*

- ***Base Commune.r***

```
rm(list=ls()) # Nettoyer l'espace de travail
source("PRE.R") # Pour utiliser le module PRE
source("outliersFunction.R")  # Pour utiliser le module Outlier
ExoX<-read.table("Exox.txt",header=TRUE,sep="\t",dec=",") # Pour importer les données
```

Données & exercices du TD :
webcom.upmf-grenoble.fr/LIP/Perso/DMuller/M2R/ACM/TD

Exercice 1

CM1

Exercice 1

•Trente quatre étudiants réalisent un examen de psychologie cognitive. Calculez un PRE (Proportion de Réduction de l'Erreur) correspondant à l'hypothèse selon laquelle la moyenne à cet examen est supérieure à 10. Dans cet exercice, on admettra que les conditions d'applications ont déjà été vérifiées et que l'étude des observations déviantes a déjà été réalisée.

•Calculez un F permettant de savoir si la moyenne des étudiants est significativement différente de 10.

→ Faire sur Excel

Modèle simple : test de la moyenne (b_0) contre une valeur spécifique (B_0)

Sujet	$Pourc_i$	$\widehat{Pourc}$ MC (=B_0)	$\widehat{Pourc}$ MA (=b_0)	Erreurs (résidus) au carré $(Pourc_i - B_0)^2$ MC	Erreurs (résidus) au carré $(Pourc_i - b_0)^2$ MA
1	71.09	50	59.89	444.78	125.44
2	77.00	50	59.89	729.19	292.88
3	60.45	50	59.89	109.08	0.31
4	48.72	50	59.89	1.63	124.75
5	39.83	50	59.89	103.40	402.36
6	54.87	50	59.89	23.70	25.21
7	52.98	50	59.89	8.88	47.74
8	50.61	50	59.89	0.37	86.16
9	66.46	50	59.89	271.03	43.21
10	63.45	50	59.89	180.83	12.66
11	81.07	50	59.89	965.26	448.53
12	47.55	50	59.89	6.00	152.29
13	56.94	50	59.89	48.04	8.76
14	67.19	50	59.89	295.67	53.36
15	60.17	50	59.89	103.53	0.08
Somme	898.38			3291.43	1823.73
Moyenne	59.89				

- L'erreur du MC était 3291.43

- Avec le MA, elle devient 1823.73

En proportion, cette erreur est diminuée de

$$\frac{(3291.43 - 1823.73)}{3291.43}$$

soit 0.44 donc 44%

=> La Proportion de Réduction de l'Erreur (PRE) = 0.44 (taille de l'effet)

Formules de bases et test du modèle simple

$$SCR = SCE_C - SCE_A$$ ➡ Réduction SCE (équivaut à SC effet « dans l'anova »)

$$PRE = \frac{\left(SCE_C - SCE_A\right)}{SCE_C} = \frac{SCR}{SCE_C}$$ ➡ Proportion de Réduction de l'Erreur ou taille de l'effet (« équivaut » au η^2)

$$F = \frac{PRE\,/\,(PA - PC)}{(1 - PRE)\,/\,(N - PA)} = \frac{SCR\,/\,(PA - PC)}{SCE_A\,/\,(N - PA)} = \frac{SC_{EFFET}\,/\,(PA - PC)}{SC_{ERREUR}\,/\,(N - PA)}$$

$$p = LOI.F(F;DDLeffet;DDLerreur)$$

Exercice 2

CM1

Modèles à un facteur continu : régression simple

$$Prc_i = b_0 + b_1 BEPC_i + e_i$$

$$\hat{Prc}_i = 39.30 + 1.83 BEPC_i$$

Avec R : mc <- lm(Pourcentage~1,DF) # « 1 » veut juste dire modèle simple

Avec R : ma <- lm(Pourcentage~1+BEPC,DF) # (ou Pourcentage~BEPC)

anova(mc,ma) # Demande de comparer ces deux modèles

```
Model 1: Pourcentage ~ 1
Model 2: Pourcentage ~ BEPC
  Res.Df    RSS Df Sum of Sq     F  Pr(>F)
1     14 1823.7
2     13 1162.4  1    661.32 7.396 0.01753 *
```

Cette comparaison de modèles nous indique que la diminution de la SCE en passant du modèle C au modèle A est significative.

Autrement dit, l'effet de BEPC sur Pourcentage est significatif

Modèles à un facteur continu : régression simple

$$Prc_i = b_0 + b_1 BEPC_i + e_i$$

$$\hat{Prc}_i = 39.30 + 1.83 BEPC_i$$

Avec R : ma <- lm(Pourcentage~BEPC,DF)
summary(ma)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  39.3016     7.9551    4.94  0.00027 ***
BEPC          1.8328     0.6739    2.72  0.01753 *

Residual standard error: 9.456 on 13 degrees of freedom
Multiple R-squared: 0.3626, Adjusted R-squared: 0.3136
F-statistic: 7.396 on 1 and 13 DF,  p-value: 0.01753
```

Avec R : anova(ma)

```
Response: Pourcentage
           Df  Sum Sq Mean Sq F value  Pr(>F)
BEPC        1  661.32  661.32   7.396 0.01753 *
Residuals 13 1162.41   89.42
```

Elements utiles dans l'exo 2

•Pour tester l'hypothèse :

Avec R :

```
source(PRE.R)
fit1 <- lm(VD~VI,exo2)
summary(fit1)
anova(fit1) (éventuellement)
PRE(fit1)
```

•La corrélation entre taille et performance est égale à la racine carre du PRE (dans la régression simple !)

Pour le prouver :

```
cor(exo2$taille,exo2$perf)
```

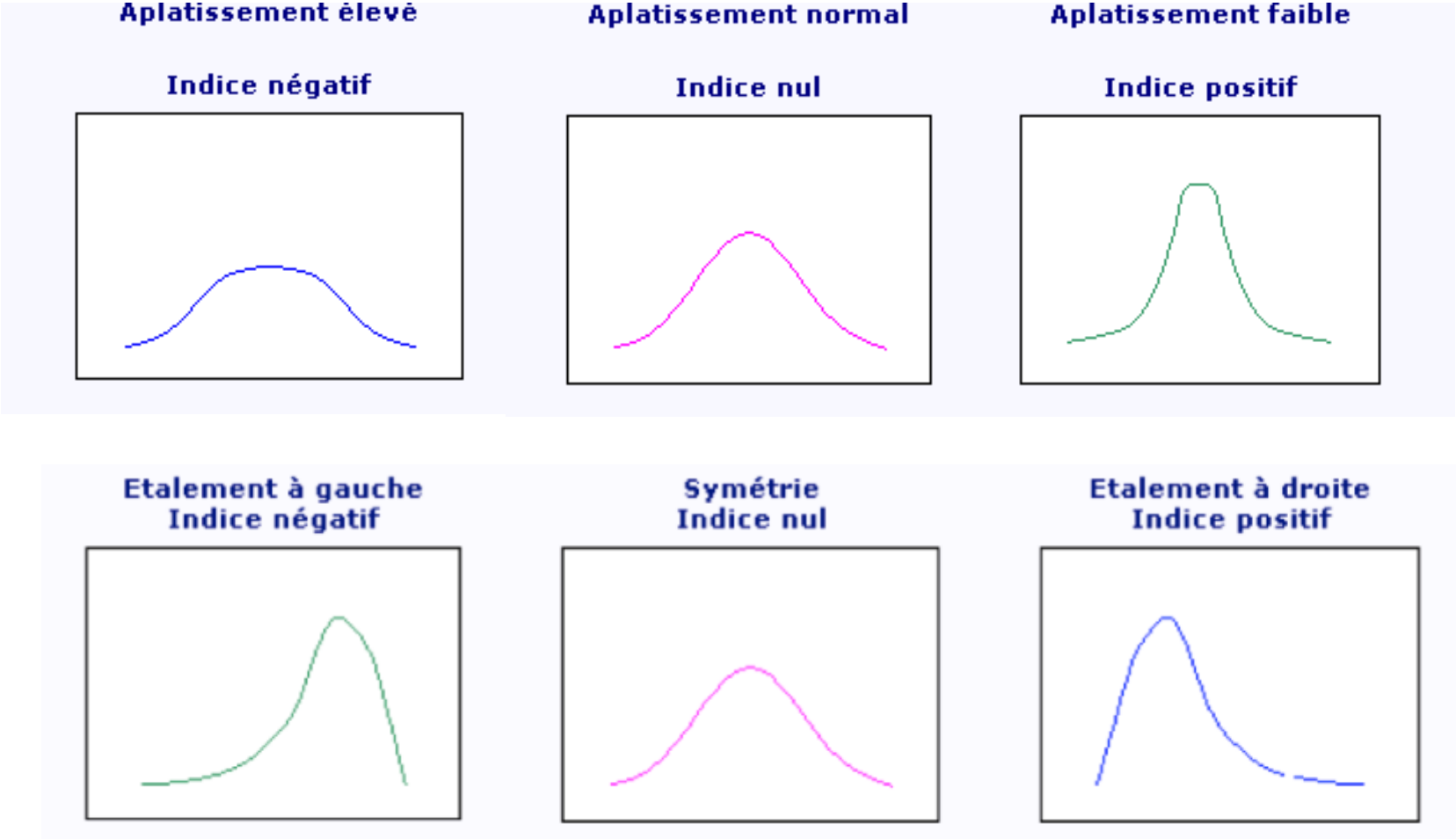
Exercice 3

CM2 & CM3

Points importants

- Déviants (Cook, RSS, Levier)
- Conditions d’application (normalité, homogénéité et indépendance des résidus)
- Codage de la VI catégorielle

Type de distribution des résidus



Types de problèmes et types de transformations

- Transformer les données pour atteindre la **normalité**
  - Pour les distributions « plates » : inverse (1/Y)
  - Pour les distributions « asymétriques + » : log (+10) ou 1/Y
  - Pour les distributions « asymétriques - » : racine carré
- Transformer les données pour atteindre une **variance constante** des résidus (homoscédasticité)
  - Cône des résidus ouvert à droite : inverse
  - Cône des résidus ouvert à gauche : racine carré

Points importants

- #Pour regarder les outliers

```
outliers(interference ~ condc,DF)
```

- # Residuals normality

```
hist(residuals(fit1)) # residuals histogram
qqnorm(residuals(fit1))
qqline(residuals(fit1)) # QQ plot
```

- # Homogeneity of variance

```
plot(fitted(fit1),residuals(fit1))
plot(fitted(fit1),abs(residuals(fit1)^.5))
abline(lm((abs(residuals(fit1)))^.5~fitted(fit1)))
```

Points importants

```
#Pour centrer la VI
DF$condc <- -0.5 * (DF$cond=="revstr") + 0.5 *
(DF$cond=="stroop”)

# 4) Testez la même hypothèse dans le module test t #
t.test (DF$interference ~ DF$condc,var.equal=TRUE)

# 5) Testez la même hypothèse dans le module ANOVA
ma1=aov(DF$interference~DF$condc)
summary(ma1)
```

Exercice 4

CM3

Famille de contrastes orthogonaux

	FBm	FBnm	NoFB	
C1 ($=\lambda_{1.k}$)	2 ($=\lambda_{1.1}$)	-1 ($=\lambda_{1.2}$)	-1 ($=\lambda_{1.3}$)	$\rightarrow \sum_k \lambda_{1.k} = 0 \Rightarrow (2-1-1=0)$
C2 ($=\lambda_{2.k}$)	0 ($=\lambda_{2.1}$)	1 ($=\lambda_{2.2}$)	-1 ($=\lambda_{2.3}$)	$\rightarrow \sum_k \lambda_{2.k} = 0 \Rightarrow (0+1-1=0)$
	2*0 = 0	-1*1 = -1	-1*-1 = 1	$\rightarrow \sum_k \lambda_{1.k}\lambda_{2.k} = 0 \Rightarrow (0-1+1=0)$

Ceci est une famille de contrastes orthogonaux car elle respecte deux règles :

Règle 1 : $\sum_k \lambda_k = 0$ et Règle 2 : $\sum_k \lambda_{1.k}\lambda_{2.k} = 0$ (les contrastes sont orthogonaux deux à deux)

Nous utiliserons toujours k-1 contrastes orthogonaux pour coder une variable catégorielle (où k = nb de modalités de la VI)

Tests des contrastes

	C1	C2	*M*
BS	**-1**	**-1**	**38,247**
SM	**0**	**2**	**56,766**
HS	**1**	**-1**	**64,903**

	Estimate	***Std. Error***	***t value***	***Pr(>|t|)***
Intercept	*53.305*	*3.153*	*16.908*	*< 2e-16 ****
c1	*13.328*	*3.817*	*3.492*	*0.000907 ****
c2	*1.730*	*2.254*	*0.768*	*0.445781*

- Interprétation de b_0 = *53.3* : prédiction pour C1 et C2 = 0, ces deux contrastes étant centrés, cela correspond à une condition moyenne. *53.3* est donc la moyenne

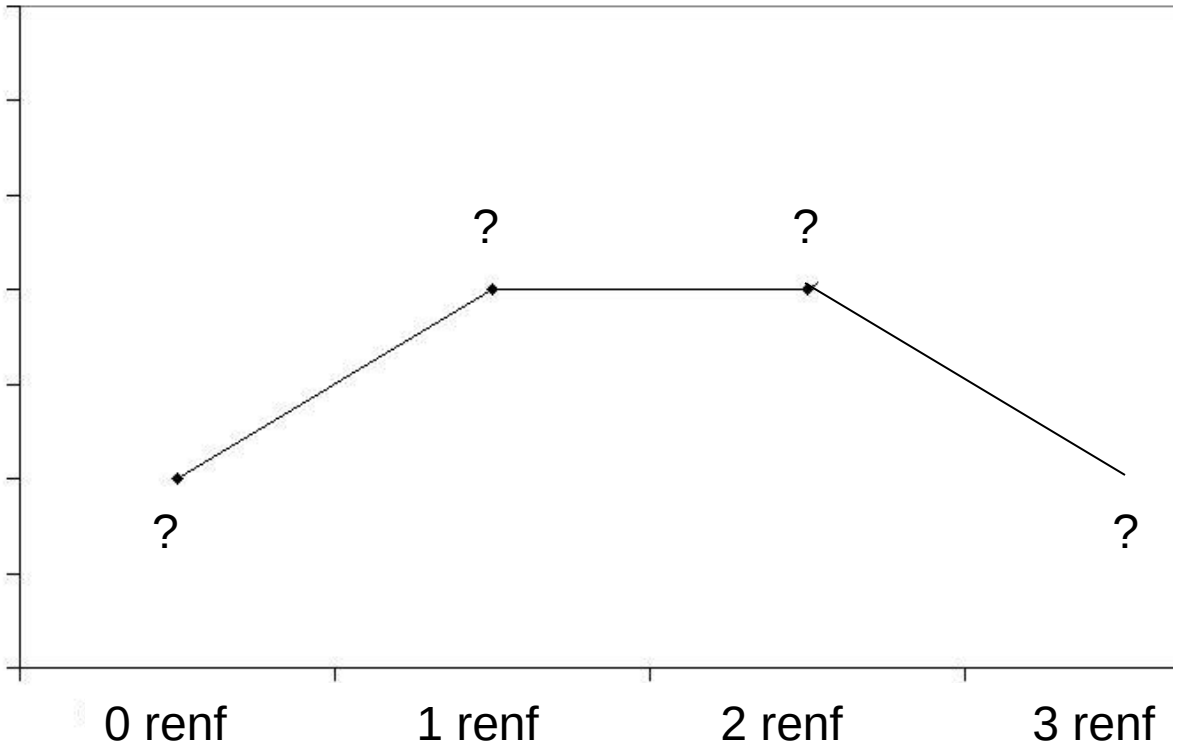
- Interprétation de b_1 = *13.328* : pour toute augmentation d'une unité, notre prédiction augmente de *13.328*. Il y a 2 unités de différence entre BS et HS, *13.328* correspond donc à 1/2 de la différence entre la moyenne de BS et HS.

- Interprétation de b_2 = *1.730* : pour toute augmentation d'une unité, notre prédiction augmente de *1.730*. Il y a 3 unités de différence entre BS/HS et SM, *1.730* correspond donc à 1/3 de la différence entre la moyenne de BS/HS et SM.

Exercice 5

CM4

Illustration



(adresse pour trouver des contrastes orthogonaux : http://www.bolderstats.com/orthogCodes/)

Codes de contraste

	O renf	1 renf	2 renf	3 renf
Mod	-1	1	1	-1
Res1	-1	0	0	1
Res2	0	-1	1	0

$$Mot_i = b_0 + b_1 Mod_i + b_2 Res1_i + b_3 Res2_i + e_i$$

Test du résidu

3b) Test du résidu => nous allons mettre ensemble tout ce qui n'est pas le modèle théorique :

MA : $Chgt_i = b_0 + b_1 Mod_i + b_2 Res1_i + b_3 Res2_i + e_i$ $\quad SCE_A = 1531.63$

MC : $Chgt_i = b_0 + b_1 Mod_i + e_i$ $\quad SCE_C = 1591.64$

Test du résidu => $SCR = SCE_C - SCE_A = 1591.64 - 1531.63 = 60$

$$F = \frac{SCR / (pa - pc)}{SCE_A / (N - pa)} = \frac{60 / (4 - 2)}{1531.63 / (20 - 4)} = 0.31$$

- Modèle significatif ET résidu non significatif => hypothèse vérifiée
- Nous pourrions également être encore plus durs avec nous-mêmes en testant le F du résidu avec un ddl de l'effet = 1

  => dans ce cas F(1,16) = 0.63

Exercice 6 & 7

CM5 & CM6

Interprétation en présence ou non d'un modèle interactif

$$\Rightarrow Y_i = b_0 + b_1 X_i + b_2 Z_i + e_i$$

- ✓ b_0 sera notre prédiction quant à la valeur de Y lorsque X = 0 et Z = 0
- ✓ b_1 sera la pente de X **LORSQUE** Z est tenu constant
- ✓ b_2 sera la pente de Z **LORSQUE** X est tenu constant

$$\Rightarrow Y_i = b_0 + b_1 X_i + b_2 Z_i + b_3 X_i * Z_i + e_i$$

- ✓ b_0 sera notre prédiction quant à la valeur de Y lorsque X = 0 et Z = 0

Ensuite, parce qu'un produit de X et Z est présent dans l'équation :

- ✓ b_1 sera la pente de X **LORSQUE** Z = 0 (EFFET SIMPLE de X pour valeur 0 de Z)
- ✓ b_2 sera la pente de Z **LORSQUE** X = 0 (EFFET SIMPLE de Z pour valeur 0 de X)
- ✓ b_3 correspondra au changement de pente pour un changement d'une unité sur l'autre variable. L'effet de X dépend-il des valeurs de Z ? L'effet de Z dépend-il des valeurs de X ?

Effets simples du type d'item pour Non bruit

Encore une fois utilisation d'un codage = 0 pour la condition d'intérêt, d'où :

✓ Nonbruit : non bruit = 0 et bruit = 1

✓ Typec : Inc = - 0.5 et Cont = 0.5 et nonbtype = Nonbruit*Typec

$$nbcorr_i = b_0 + b_1 typec_i + b_2 nonbruit_i + b_3 nonbruit_i * typec_i + e_i$$

$$nbcorr_i = (b_0 + b_2 nonbruit_i) + (b_1 + b_3 nonbruit_i) typec_i + e_i$$

Le test de b_1 sera l'effet du type d'item lorsque nonbruit = 0 donc lorsqu'il n'y a pas de bruit

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance
Intercept	1	23.12500	0.79895	28.94	<.0001	.
nonbruit	1	2.00000	1.12989	1.77	0.0807	1.00000
typec	1	15.35000	1.59790	9.61	<.0001	0.50000
nonbtype	1	-4.90000	2.25977	-2.17	0.0333	0.50000

$$nbcorr_i = (23.13 + 2 nonbruit_i) + (15.35 - 4.90 nonbruit_i) typec_i$$

- Pour b_1 on retrouve 15.35, c'est-à-dire l'effet simple que nous voulions
- Ce test indique que lorsqu'il n'y a pas de bruit la performance est significativement meilleure pour les items contrôles, t(76) = 9.61, p < .001

Exercice 8

CM6