

MAT7381 Solutions

Problème tiré des données du # 5.11

Une expérience a été menée afin de déterminer l'effet de trois méthodes d'enseignement de la lecture à des enfants. Les sujets ont composé des tests de compréhension avant et après la période d'enseignement. Nombre de cas: 66.

<i>Variables</i>	1. ID:	Identification du sujet
	2. Groupe:	Type de formation reçue (nommé Basal, DRTA, ou Strat)
	3. PRE1:	Score au premier pré-test de compréhension
	4. PRE2:	Score au second pré-test de compréhension
	5. POST1:	Score au premier post test de compréhension
	6. POST2:	Score au second post test de compréhension
	7. POST3:	Score au troisième post test de compréhension

Les données sont présentées dans le tableau 4 en annexe.

- a) Soit y la moyenne des trois post-tests. Effectuez une analyse de variance afin de déterminer si les différences entre les groupes (par rapport aux post-tests moyens) est significative.

```
> anova(lm(y~groupe))
Response: y
      Df Sum Sq Mean Sq F value    Pr(>F)
groupe  2 121.72  60.860   7.9272 0.0008597 ***
```

La valeur p de la statistique F globale est 0,0008597. On peut donc rejeter avec confiance l'hypothèse que les trois moyennes sont égales.

- b) Déterminer spécifiquement si le deuxième groupe et le troisième groupe diffèrent du premier

```
> summary(lm(y~groupe))
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.7576     0.5907  30.060 < 2e-16 ***
groupeDRTA    3.1515     0.8354   3.772 0.000364 ***
groupeStrat   2.5123     0.8453   2.972 0.004208 **

Residual standard error: 2.771 on 62 degrees of freedom
Multiple R-squared:  0.2036,    Adjusted R-squared:  0.178
F-statistic: 7.927 on 2 and 62 DF,  p-value: 0.0008597
```

Les différences entre les groupes 2 et 1 par rapport au groupe 1 sont toutes deux significatives.

- c) Les comparaisons faites en a) et b) ne tiennent pas compte des compétences initiales. Si on mesure les compétences initiales par x (la moyenne des deux pré-tests), déterminer une analyse de variance dans un modèle dans lequel la relation entre y et x est une droite différente dans chaque groupe.

Le modèle peut s'écrire comme ceci : $\mathcal{M} : E(y) = \beta_0 + \beta_1 x + \beta_2 t_2 + \beta_3 t_3 + \beta_4 x_2 + \beta_5 x_3$, où

t_2 est une variable dont la i^e valeur est égale à 1 si le sujet i subit le traitement DRTA, et 0 sinon

t_3 est une variable dont la i^e valeur est égale à 1 si le sujet i subit le traitement Strat, et 0 sinon

x_2 est une variable dont la i^e valeur est égale à x_i si le sujet i subit le traitement DRTA, et 0 sinon

x_3 est une variable dont la i^e valeur est égale à x_i si le sujet i subit le traitement Strat, et 0 sinon

Donc, les moyennes pour les trois traitements sont :

Traitement A : $\beta_0 + \beta_1 x$

Traitement B : $\beta_0 + \beta_1 x + \beta_2 + \beta_4 x = \beta_0 + \beta_2 + (\beta_1 + \beta_4)x$

Traitement C : $\beta_0 + \beta_1 x + \beta_3 + \beta_5 x = \beta_0 + \beta_3 + (\beta_1 + \beta_5)x$

Voici l'analyse de variance dans ce modèle :

```
> anova(a)
      Df Sum Sq Mean Sq F value    Pr(>F)
x       1  55.03  55.034   8.3564 0.0053714 **
drta    1  56.44  56.440   8.5698 0.0048487 **
strat   1  89.19  89.192  13.5429 0.0005067 ***
drtax   1   3.14   3.138   0.4765 0.4927019
```

```
stratx    1    5.34    5.342    0.8112    0.3714315
Residuals 59 388.57    6.586
```

Selon ces résultats, les coefficients de x ne varient pas selon le groupe

```
> summary(a)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.22106    1.95355    7.280 9.09e-10 ***
x            0.44843    0.23780    1.886  0.0643  .
drta        4.01296    2.88591    1.391  0.1696
strat       0.53187    2.81513    0.189  0.8508
drtax      -0.08738    0.36509   -0.239  0.8117
stratx      0.32652    0.36253    0.901  0.3714

Residual standard error: 2.566 on 59 degrees of freedom
Multiple R-squared:  0.3499,    Adjusted R-squared:  0.2948
F-statistic: 6.351 on 5 and 59 DF,  p-value: 8.826e-05
```

Nous avons donc l'équation suivante :

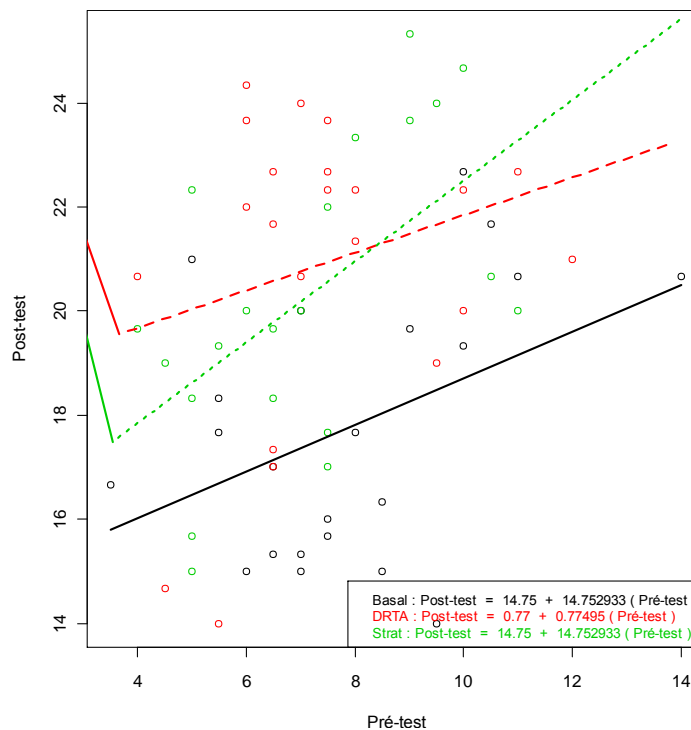
$$E(y) = 14.22106 + 0.44843x + 4.01296(\text{drta}) + 0.53187(\text{strat}) - 0.08738(\text{drtax}) + 0.32652(\text{stratx})$$

Donc, par groupe, les équations sont

Groupe Basal : $E(y) = 14.22106 + 0.44843x$

Groupe DRTA : $E(y) = 14.22106 + 0.44843x + 4.01296(\text{drta}) - 0.08738(\text{drtax})$

Groupe Strat : $E(y) = 14.22106 + 0.44843x + 0.53187(\text{strat}) + 0.32652(\text{stratx})$



- d) D'après l'analyse en c), est-ce que la pente de la droite dans le groupe DRTA est différente de celle dans le groupe Basal ? Celle dans le groupe Strat est-elle différente de celle dans le groupe Basal ?

La procédure `summary()` indique que les pentes ne sont pas différentes. Mais elle indique également que les constantes des droites (les ordonnées à l'origine) sont elles aussi égales. La procédure `anova()`, par contre, semble indiquer que les constantes sont, quant à elles, égales. Si on s'y fiait, on conclurait que le traitement n'a aucun effet, ni sur la pente, ni sur la constante : l'espérance de y est indépendante de x et indépendante du groupe. Une telle conclusion serait injustifiée : les très fortes corrélations entre les variables exogènes font qu'aucune d'elles, en présence des autres, n'est significative. La procédure `anova()` a l'avantage que les premiers facteurs sont testés en l'absence de celles qui viendront après. Elles révèlent donc des différences significatives cachées. Nous devrions simplifier le modèle en considérons la possibilité que les trois droites sont parallèles.

e) Tester l'hypothèse que les trois droites sont parallèles.

Nous testerons cette hypothèse à l'aide la statistique $F = [(SCR0-SCR)/v]/MCR$, où SCR et MCR sont la somme et la moyenne des carrés résiduelle dans le modèle, SCR0 est la somme des carrés résiduelle dans le modèle restreint (par l'hypothèse nulle) $\mathcal{M}_0 : E(y) = E(y) = \beta_0 + \beta_1x + \beta_2t_2$ et v est la différence entre les degrés de liberté de SCR0 et SCR.

```
> scr<-anova(a)$"Sum Sq"[6]
> scr
[1] 388.5689
> dl<-59
> mcr<-scr/dl
> mcr
[1] 6.585913
```

La moyenne des carrés résiduelle est $MCR = 6,585913$ à 59 degrés de liberté. On détermine maintenant le modèle restreint :

```
> a0<-lm(y~x+drta+strat)
> scr0<-anova(a0)$"Sum Sq"[4]
> scr0
[1] 397.0497
> dl0<-61
> dl0-dl
[1] 2
> n
[1] 65
> q
[1] 6
> scr0-scr
[1] 8.480833
> (scr0-scr)/(dl0-dl)
[1] 4.240416
> f<-(scr0-scr)/(dl0-dl)/mcr
> f
[1] 0.6438615
> 1-pf(f,dl0-dl,n-q)
[1] 0.5289107
```

On pourrait difficilement conclure que les pentes sont différentes. On modèle dans lequel les trois droites sont parallèles semble donc éminemment raisonnable.

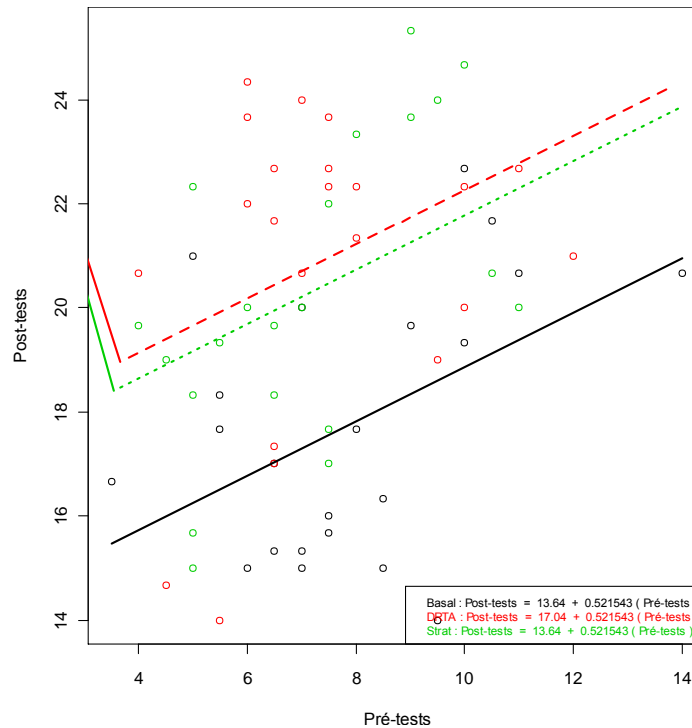
e) Déterminer un modèle dans lequel les trois droites sont parallèles. Dans ce modèle, déterminer si les différences entre les groupes sont significatives.

On considère, dans le cadre du modèle réduit \mathcal{M}_0 le modèle réduit davantage, soit $\mathcal{M}_{00} : E(y) = \beta_0 + \beta_1x$

```
> a00<-lm(y~x)
> dl00<-63
> scr00<-anova(a00)$"Sum Sq"[2]
> scr00
[1] 542.6818
> q0<-4
> scr0
[1] 397.0497
> mcr0
[1] 6.509012
> (scr00-scr0)
[1] 72.81604
> (scr00-scr0)/(dl00-dl)
[1] 36.40802
> (scr00-scr0)/(dl00-dl)/mcr0
[1] 11.18696
> f<-(scr00-scr0)/(dl00-dl)/mcr0
> f
[1] 11.18696
> 1-pf(f,dl00-dl,n-q0)
```

[1] 7.263379e-05

Là on peut rejeter avec confiance l'hypothèse que, pour une compétence initiale x , l'espérance de y est la même pour tous les groupes.



f) Dans le modèle établi en e), tester l'hypothèse que la moyenne du groupe DRTA est égale à celle du groupe Strat.

Dans le modèle $\mathcal{M}_0 : E(y) = E(y) = \beta_0 + \beta_1 x + \beta_2 t_2 + \beta_3 t_3$ on testera l'hypothèse que $\beta_2 = \beta_3 = 0$, soit $\ell' \beta = 0$, où $\ell' = [0 \ ; \ 0 \ ; \ 0 \ 1 \ ; \ -1]$. La

statistique est $T = \frac{\ell' \hat{\beta}}{\sqrt{\ell' \hat{V} \ell}}$, où \hat{V} est la matrice de covariance estimée de $\hat{\beta}$. Sous H_0 , T suit une loi de Student à $n-3$ degrés de liberté. Voici les calculs.

```
> V<-ls.diag(a0)$cov.scaled
> V
      (Intercept)      x      drta      strat
(Intercept)  1.6906794 -0.17686417 -0.38027662 -0.43157488
x             -0.1768642  0.02242658  0.01070360  0.01720827
drta         -0.3802766  0.01070360  0.59683688  0.30407721
strat        -0.4315749  0.01720827  0.30407721  0.61902130
> l
[1] 0 0 1 -1
> phi<-sum(l*beta)
> phi
[1] 0.4879795
> sphi<-sqrt(t(l)**solve(V)**l)
> sphi
      [,1]
[1,] 2.570258
> phi/sphi
      [,1]
[1,] 0.1898563
```

On ne peut pas rejeter l'hypothèse que les groupes deux et trois sont différents.

g) Déterminer un intervalle de confiance à 95 % pour la valeur moyenne de y pour un sujet ayant un résultat de 10 au pré-test dans chacun des groupes

```
> predict(a0,data.frame(x=x0,drta=0,strat=0),interval="confidence")
      fit      lwr      upr
1 18.85993 17.60151 20.11835
> predict(a0,data.frame(x=x0,drta=1,strat=0),interval="confidence")
      fit      lwr      upr
1 22.26036 20.92434 23.59639
> predict(a0,data.frame(x=x0,drta=0,strat=1),interval="confidence")
      fit      lwr      upr
1 21.77238 20.36397 23.18079
```