STID - IUT de Paris - Rives de Seine **Modèle linéaire**
2AFA
2020-2021

Feuille de TD 2 : Régression linéaire multiple

Exercice 1

Un fortifiant F est essayé sur 30 rats. 10 rats sont élevés sans fortifiant et on répartit les autres en 4 groupes de 5 rats chacun : 5 rats recoivent 1 mg de F, 5 autres 2 mg de F, 5 autres 3 mg de F, et les 5 derniers, 4 mg de F. On mesure le poids de chaque rat après 2 mois de traitement. On obtient les résultats suivants :

Dose de Fortifiant x_i	0		1	2	3	4
Poids y_i	84,9	82,9	114,3	128,8	125,5	129,1
	106,1	99,6	107,4	112,8	122,6	121,3
	114,8	98,2	124,9	114,0	114,1	116,6
	109,2	84,3	98,9	118,2	109,3	101,8
	112,0	118,0	124,3	119,5	102,2	130,3

On note x_i la dose de fortifiant donnée au $i$ème rat et y_i le poids final correspondant.

1. Les promoteurs du fortifiant F pensent que pour les doses utilisées, il y a une relation linéaire entre la quantité de fortifiant x et le poids en fin d'expérience Y. Introduire le modèle, noté M_2, et tester si la quantité de fortifiant influe de manière significative sur le poids en fin d'expérience. Indication : $\mathrm{SCR}(M_2) = 3584.54$. On donne également $\mathrm{SCT} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = 4911.24$.
2. Les expérimentateurs du fortifiant F se demandent néanmoins si, même pour les faibles doses utilisées, il n' y a pas "tassement" de l'effet, et envisagent pour répondre à cette question le modèle parabolique

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$

où les ε_i sont gaussiennes, indépendantes, centrées et de variance σ^2. On appelera M_3 ce modèle.

(a) Tester l'hypothèse $\beta_1 = 0$ et $\beta_2 = 0$ au risque $\delta = 5\%$ à l'aide de la table 1 d'analyse de la variance .

TABLE 1: Table d'analyse de la variance du passage du modèle M_1 au modèle M_3

Source	ddl	Somme des carrés	Carrés Moyens	statistique de test
passage de M_1 à M_3				
Résiduelle M_3		3225.18		
Résiduelle M_1				

(b) Tester l'hypothèse $\beta_2 = 0$ au risque $\delta = 5\%$ à l'aide de la table 2 d'analyse de la variance.

TABLE 2: Table d'analyse de la variance du passage du modèle M_2 au modèle M_3

Source	ddl	Somme des carrés	Carrés Moyens	statistique de test
passage de M_2 à M_3				
Résiduelle M_3				
Résiduelle M_2				

Exercice 2

On s'intéresse aux performances sportives d'enfants de 12 ans. Chaque enfant passe une dizaine d'epreuves (courses, lancers, sauts,...), et les résultats sont synthétisés dans un indice global noté Y. On cherche à mesurer l'incidence sur ces performances de deux variables (contrôlées) : la capacité thoracique X^1 et la force musculaire X^2. Ces trois quantités sont repérées par rapport à une valeur de référence, notée à chaque fois 0, les valeurs positives étant associées aux bonnes "performances". Elles sont mesurées sur un échantillon de 60 enfants. Pour chaque enfant, $i = 1, \cdots, 60$, on note x_i^1 la capacité thoracique, x_i^2 sa force musculaire et y_i sa performance sportive.

On suppose que les y_i sont les réalisations de n va Y_i de loi $\mathcal{N}(\beta_0 + \beta_1 x_i^1 + \beta_2 x_i^2, \sigma^2)$, c'est à dire qu'on adopte (dans un premier temps) le modèle linéaire suivant pour étudier le lien entre Y et les deux variables explicatives X^1 et X^2 : $Y_i = \beta_0 + \beta_1 x_i^1 + \beta_2 x_i^2 + \varepsilon_i$ où les ε_i sont des va gaussiennes indépendantes centrées et de variance σ^2.

1. Ecrire le modèle sous forme matricielle en précisant le contenu des vecteurs et matrices utilisés.

   Pour chacun des tests suivants on décrira les sous-modèles (du modèle complet ci-dessus) mis en jeu.
2. Tester au risque 5% l'hypothèse selon laquelle la performance sportive ne dépend ni de la capacité thoracique ni de la force musculaire.
3. Tester au risque 5% l'hypothèse selon laquelle la performance sportive ne dépend pas de la force musculaire en plus de la capacité thoracique.
4. Tester au risque 5% l'hypothèse selon laquelle la performance sportive ne dépend pas de la capacité thoracique en plus de la force musculaire.
5. Quel modèle proposeriez-vous d'adopter au vu des résultats des tests précédents.
6. Donner une prévision de la performance d'un enfant dont la force musculaire serait égale à 5 et la capacité thoracique égale à 3. Donner un intervalle de prévision de cette valeur.

On donne les informations suivantes

$$SCR(M_1) = 298704.6, SCR(M2) = 240997.4, SCR(M_2') = 286542.6, SCR(M3) = 229581.9$$

où M_1 est le modèle sans variable, M_2 est le modèle avec la seule variable x_1, M_2' est le modèle avec la seule variable x_2 et M_3 est le modèle avec les deux variables.

Exercice 3

On souhaite étudier la variation du taux d'hémoglobine dans le sang Y au cours d'une opération chirurgicale en fonction de la durée de l'opération X^1 et du volume de sang perdu pendant l'opération X^2. L'objet de l'étude est d'expliquer par un modèle linéaire la variable Y en fonction des deux variables explicatives X^1 et X^2. On dispose des résultats suivants où y_i représente la valeur observée en pourcentage de la variation du taux d'hémoglobine, x_i^1 est la durée de l'opération en heures décimales et x_i^2 est le volume en litres de sang perdu.

y_i	-1.70	-4.61	-5.82	-1.17	-4.23	-3.31	+0.42	-2.98
x_i^1	1.75	1.33	1.43	1.86	1.81	1.66	1.60	2.00
x_i^2	0.52	0.59	0.61	0.50	0.54	0.49	0.27	0.47

1. Combien de modèles peut-on envisager pour décrire les données ?
2. Utilisez les sorties R pour donner l'estimation des paramètres des modèles mis en jeu.
3. Utilisez les sorties R pour effectuer les tests de comparaison du modèle complet aux autres modèles. Précisez à chaque fois, les hypothèses à tester, la statistique de test et sa loi sous H_0 et interprétez la pvaleur.
4. Interprétez les résultats des autres tests de comparaison de modèles.
5. Quel modèle choisiriez-vous ?
6. On a calculé la valeur du critère AIC de tous ces modèles avec le logiciel R. Cela vous semble t-il cohérent avec le résultat des test précédents ?
7. Donner, pour un patient qui subit une opération d'une durée $x_0^1 = 1,25$ et dont le volume en litres de sang perdu est $x_0^2 = 0,52$ une prévision de la variation du taux d'hémoglobine dans le sang de ce patient.

```
> modele1=lm(Y~1,data=hemo)
> summary(modele1)

Call:
lm(formula = Y ~ 1, data = hemo)

Residuals:
   Min     1Q Median     3Q    Max
-2.895 -1.400 -0.220  1.357  3.345

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.9250     0.7177  -4.076  0.00472 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.03 on 7 degrees of freedom

> modele2=lm(Y~X1,data=hemo)
> summary(modele2)

Call:
```

```
lm(formula = Y ~ X1, data = hemo)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9851 -1.3593 -0.3617  1.0026  3.6362

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -9.039      5.764  -1.568    0.168
X1             3.639      3.405   1.069    0.326
--
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.01 on 6 degrees of freedom
Multiple R-squared:  0.16,      Adjusted R-squared: 0.01999
F-statistic: 1.143 on 1 and 6 DF,  p-value: 0.3262

> modele2bis=lm(Y~X2,data=hemo)
> summary(modele2bis)

Call:
lm(formula = Y ~ X2, data = hemo)

Residuals:
    Min      1Q  Median      3Q     Max
-1.0503 -0.5540 -0.4891  0.2654  1.7757

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.345      2.117   2.525  0.04501 *
X2           -16.582      4.166  -3.980  0.00728 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.149 on 6 degrees of freedom
Multiple R-squared: 0.7253,     Adjusted R-squared: 0.6795
F-statistic: 15.84 on 1 and 6 DF,  p-value: 0.007283

> modele3=lm(Y~X1+X2,data=hemo)
> summary(modele3)

Call:
lm(formula = Y ~ X1 + X2, data = hemo)

Residuals:
       1        2        3        4        5        6        7        8
 1.43822  0.33289 -0.73424  1.46954 -0.88114 -0.48775 -0.09141 -1.04612

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.017      4.629   0.436   0.6813
X1             1.694      2.079   0.815   0.4522
X2           -15.615      4.449  -3.510   0.0171 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.183 on 5 degrees of freedom
Multiple R-squared: 0.7575,     Adjusted R-squared: 0.6605
F-statistic: 7.809 on 2 and 5 DF,  p-value: 0.02896

> anova(modele1,modele3)
Analysis of Variance Table

Model 1: Y ~ 1
Model 2: Y ~ X1 + X2
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1      7 28.844
2      5  6.995  2    21.849 7.8089 0.02896 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> anova(modele1,modele2)
Analysis of Variance Table

Model 1: Y ~ 1
Model 2: Y ~ X1
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1      7 28.844
2      6 24.229  1    4.6149 1.1428 0.3262
> anova(modele1,modele2bis)
Analysis of Variance Table

Model 1: Y ~ 1
Model 2: Y ~ X2
  Res.Df     RSS Df Sum of Sq      F   Pr(>F)
1      7 28.8442
2      6  7.9241  1     20.92 15.841 0.007283 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> anova(modele2,modele3)
Analysis of Variance Table

Model 1: Y ~ X1
Model 2: Y ~ X1 + X2
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1      6 24.229
2      5  6.995  1    17.234 12.319 0.0171 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> anova(modele2bis,modele3)
Analysis of Variance Table

Model 1: Y ~ X2
Model 2: Y ~ X1 + X2
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1      6 7.9241
2      5 6.9950  1    0.9291 0.6641 0.4522

> AIC(modele1)
[1] 36.96276
> AIC(modele2)
[1] 37.56799
> AIC(modele2bis)
[1] 28.62672
> AIC(modele3)
[1] 29.62901
```