L2 MIASHS
Année 2021-22

Statistiques Descriptives 2

TD 2 de Statistiques Descriptives 2

Exercice 1 Notation matricielle

On considère le modèle de régression linéaire simple du Chapitre 1 où l'on dispose de n observations (x_i, Y_i) vérifiant

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

où l'on suppose que les variables $\epsilon_i, i = 1 \dots n$ sont centrées, de variance σ^2 et non-corrélées. On veut retrouver les propriétés du Chapitre 1 à l'aide des notations matricielles du Chapitre 2.

1. Écrire le modèle sous la forme matricielle d'un modèle de régression linéaire multiple.
2. Calculer l'estimateur des moindres carrés $\hat{\beta}$ dans le modèle matriciel et retrouver les estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$ du modèle de régression simple.

Exercice 2 Retrouver la formule de $\hat{\beta}$ dans le modèle linéaire multivarié.

Exercice 3 Au vu de la représentation de la concentration d'ozone en fonction de la température à midi nous souhaitons modéliser l'ozone par la température via un modèle quadratique.

1. Afficher sous R les données à l'aide des commandes suivantes

```
ozone<-read.table("ozone.txt",header=T)
y = ozone$maxO3
x = ozone$T12
plot(x,y)
res = lm(y~ x)
abline(res$coefficients)
```

2. Ecrire le modèle et estimer les paramètres (les calculs se feront sous R).
3. Comparer ce modèle au modèle de régression linéaire à l'aide du critère BIC.

Exercice 4

1. Montrer que pour un modèle linéaire multiple

$$Y_i = \beta_0 + \beta^t X_i + \epsilon_i$$

on peut modifier les données X_i en $X_{i,*}$ de telle maniere que

$$Y_i = \beta_*^t X_{i,*} + \epsilon_i$$

pour un certain vecteur β_*. On pourra donc supposer que tous les modèles linéaires seront de ce type. On notera $\beta = \beta_*$ dans la suite.

2. On rappelle que la solution du problème d'estimation pour modèle linéaire Gaussien est donnée par

$$\hat{\beta} = (X^t X)^{-1} X^t Y \tag{1}$$

— Rappeler ce que valent $\mathbb{E}\left[\hat{\beta}\right]$ et

$$Var(\hat{\beta}) =: \mathbb{E}\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^t\right] \tag{2}$$

— Retrouver alors les formules de l'Exercice 2 du TD/TP précédant pour le cas de la régression linéaire simple.

Exercice 5 On considère les données suivantes :

y	x_1	x_2	x_3	x_4
125	13	18	25	11
158	39	18	59	30
207	52	50	62	53
182	29	43	50	29
196	50	37	65	56
175	64	19	79	49
145	11	27	17	14
144	22	23	31	17
160	30	18	34	22
175	51	11	58	40
151	27	15	29	31
161	41	22	53	39
200	51	52	75	36
173	37	36	44	27
175	23	48	27	20
162	43	15	65	36
155	38	19	62	37
230	62	56	75	50
162	28	30	36	20
153	30	25	41	33

où
— Y - est une mesure du succès à un test de langue.
— x_1 - est une mesure du score moyen du candidat à un test de compétance en éthnologie.
— x_2 - est une mesure de l'étique de travail du candidat.
— x_3 - est une mesure du score moyen à un test d'allemand
— x_4 - est une mesure du score moyen à un test de mathématiques.

1. Ecrire le modèle sous forme matricielle en supposant les erreurs d'observation Gaussiennes.

2. Estimer le vecteur β puis donner l'équation de la droite des moindres carrés. Pour cela, on donne (n'oubliez pas la colonne de 1 dans la matrice X !)

$$\left(X^t X\right)^{-1} = \begin{bmatrix} 0.5464 & 0.0050 & -0.0056 & -0.0075 & -0.0045 \\ 0.0050 & 0.0032 & -0.0000 & -0.0016 & -0.0014 \\ -0.0056 & -0.0000 & 0.0003 & -0.0000 & -0.0001 \\ -0.0075 & -0.0016 & -0.0000 & 0.0013 & 0.0000 \\ -0.0045 & -0.0014 & -0.0001 & 0.0000 & 0.0017 \end{bmatrix}$$

3. Calculer les estimations de σ^2 et $\mathrm{Var}(\hat{\beta})$ en prenant la formule (**??**) de l'exercice précédant.
4. Tester l'hypothese nulle H_0 : "$\beta_j = 0$" contre l'alternative H_1 :^{n} $\beta_j \neq 0^n$ pour $j = 0, 1, 2$. Pour cette question, il faut regarder la section 3.6 du polycopié sur ma page web.

Exercice 6 Jeux videos

Les données que nous étudions le score y de à un certain jeu vidéo en fonction de divers facteurs, ou covariables :

— x_1 est le taux d'une certaine hormone,
— x_2 est le score à un test de logique
— x_3 est le nombre d'amis avec lesquels on communique activement a propos de ce jeux.

x_1	x_2	x_3	y
124	33	8	81
49	31	6	55
181	38	8	80
4	17	2	24
152	39	6	52
75	30	7	88
54	29	7	45
43	35	6	50
41	31	5	69
17	23	4	66
22	21	3	45
16	8	3	24
10	23	3	43
63	37	6	38
170	40	8	72
15	38	6	41
15	25	4	38
221	39	7	52
171	33	7	52
97	38	6	66
254	39	8	89

(3)

1. Régresser Y sur x_1 et tester la signification de cette régression.

2. Trouver l'équation de la régression multiple de Y sur x_1 et x_2. On donne la matrice $(X^tX)^{-1}$ dans ce cas [1] :

$$(X^tX)^{-1} = \begin{bmatrix} 0.4919 & -0.0074 & -0.0058 \\ -0.0074 & 0.0003 & -0.0001 \\ -0.0058 & -0.0001 & 0.0003 \end{bmatrix}.$$

3. Construire la régression multiple de Y sur x_1, x_2 et x_3. On donne la matrice $(X^tX)^{-1}$ dans ce cas :

$$(X^tX)^{-1} = \begin{bmatrix} 0.5344 & 0.0014 & -0.0057 & -0.0075 \\ 0.0014 & 0.0021 & -0.0001 & -0.0015 \\ -0.0057 & -0.0001 & 0.0003 & -0.0000 \\ -0.0075 & -0.0015 & -0.0000 & 0.0013 \end{bmatrix}$$

4. Quel modèle est le plus pertinent au vu du critère AIC ? Il faut se référer à la section 3.8.2 du polycopié sur ma page web.
5. Même question au vu du critère BIC (même section du poly de cours).

Exercice 7 Refaire les deux derniers exercices en utilisant les fonctions R appropriées.

Exercice 8 Essayez la séance
https://web.stanford.edu/class/stats191/notebooks/Multiple*linear*regression.html

1. n'oubliez pas la colonne de "1" dans la matrice X !