

Table des matières

1	La régression linéaire simple	1
1.1	Introduction	1
1.1.1	Un exemple : la pollution de l'air	1
1.1.2	Un deuxième exemple : la hauteur des arbres	3
1.2	Modélisation mathématique	5
1.2.1	Choix du critère de qualité et distance à la droite	5
1.2.2	Choix des fonctions à utiliser	7
1.3	Modélisation statistique	9
1.4	Estimateurs des moindres carrés	10
1.4.1	Calcul des estimateurs de β_j , quelques propriétés	10
1.4.2	Résidus et variance résiduelle	13
1.4.3	Prévision	14
1.5	Interprétations géométriques	15
1.5.1	Représentation des individus	15
1.5.2	Représentation des variables	15
1.5.3	Le coefficient de détermination R^2	16
1.6	Inférence statistique	17
1.7	Exemples	21
1.7.1	La concentration en ozone	21
1.7.2	La hauteur des eucalyptus	26
1.8	Exercices	29
1.9	Notes : estimateurs du maximum de vraisemblance	31
2	La régression linéaire multiple	33
2.1	Introduction	33
2.2	Modélisation	34
2.3	Estimateurs des moindres carrés	38
2.3.1	Calcul de $\hat{\beta}$	38
2.3.2	Interprétation	41
2.3.3	Quelques propriétés statistiques	41
2.3.4	Résidus et variance résiduelle	42
2.3.5	Prévision	44
2.4	Interprétation géométrique	44

2.5	Exemples	46
2.5.1	La concentration en ozone	46
2.5.2	La hauteur des eucalyptus	48
2.6	Exercices	50
3	Inférence dans le modèle Gaussien	53
3.1	Estimateurs du Maximum de Vraisemblance	53
3.2	Nouvelles propriétés statistiques	54
3.3	Intervalles et régions de confiance	56
3.4	Exemple	57
3.5	Prévision	59
3.6	Les tests d'hypothèses	60
3.6.1	Introduction	60
3.6.2	Test entre modèles emboîtés	61
3.7	Exemples	65
3.7.1	La concentration en ozone	65
3.7.2	La hauteur des eucalyptus	66
3.8	Exercices	69
3.9	Notes	71
3.9.1	Intervalle de confiance : bootstrap	71
3.9.2	Test de Fisher pour une hypothèse linéaire quelconque	74
3.9.3	Propriétés asymptotiques	76
4	Validation du modèle	81
4.1	Analyse des résidus	82
4.1.1	Les différents résidus	82
4.1.2	Ajustement individuel au modèle et valeur aberrante	84
4.1.3	Analyse de la normalité	85
4.1.4	Analyse de l'homoscédasticité	85
4.1.5	Analyse de la structure des résidus	86
4.1.6	Conclusion	89
4.2	Analyse de la matrice de projection	89
4.3	Autres mesures diagnostiques	91
4.4	Effet d'une variable explicative	94
4.4.1	Ajustement au modèle	94
4.4.2	Régression partielle : impact d'une variable	95
4.4.3	Résidus partiels et résidus partiels augmentés	96
4.5	Exemple : la concentration en ozone	97
4.6	Exercices	101
5	Régression sur variables qualitatives	103
5.1	Introduction	103
5.2	Analyse de la covariance	105
5.2.1	Introduction : exemple des eucalyptus	105
5.2.2	Modélisation du problème	106

5.2.3	Hypothèse Gaussienne	108
5.2.4	Exemple : la concentration en ozone	109
5.2.5	Exemple : la hauteur des eucalyptus	114
5.3	Analyse de la variance à 1 facteur	116
5.3.1	Introduction	116
5.3.2	Modélisation du problème	117
5.3.3	Estimation des paramètres	119
5.3.4	Interprétation des contraintes	120
5.3.5	Hypothèse Gaussienne et test d'influence du facteur	120
5.3.6	Exemple : la concentration en ozone	122
5.3.7	Une décomposition directe de la variance	127
5.4	Analyse de la variance à 2 facteurs	127
5.4.1	Introduction	127
5.4.2	Modélisation du problème	128
5.4.3	Estimation des paramètres	130
5.4.4	Analyse graphique de l'interaction	131
5.4.5	Hypothèse gaussienne et test de l'interaction	133
5.4.6	Tableau d'analyse de la variance	135
5.4.7	Conclusion	136
5.4.8	Exemple : la concentration en ozone	136
5.5	Exercices	138
5.6	Notes : identifiabilité et contrastes	139
6	Choix de variables	143
6.1	Introduction	143
6.2	Choix incorrect de variables : conséquences	145
6.2.1	Analyse du biais des estimateurs	145
6.2.2	Analyse de la variance des estimateurs	147
6.2.3	Erreur quadratique moyenne	148
6.2.4	Erreur quadratique moyenne de prévision	151
6.3	La sélection de variable en pratique	153
6.3.1	Deux jeux de données ou beaucoup d'observations	153
6.3.2	Un seul jeu de données et peu d'observations	154
6.4	Critères classiques de choix de modèles	155
6.4.1	Tests entre modèles emboîtés	155
6.4.2	Le R^2	156
6.4.3	Le R^2 ajusté	158
6.4.4	Le C_p de Mallows	159
6.4.5	Vraisemblance et pénalisation	162
6.4.6	Lien entre les critères	163
6.5	Procédure de sélection	165
6.5.1	Recherche exhaustive	165
6.5.2	Recherche pas à pas	166
6.6	Exemple : la concentration en ozone	168
6.7	Sélection et shrinkage	170

6.8	Exercices	173
6.9	Notes : extension du C_p	174
7	Moindres carrés généralisés	179
7.1	Introduction	179
7.2	Moindres carrés pondérés	180
7.3	Estimateur des Moindres Carrés Généralisés	183
7.3.1	Estimateur des MCG et optimalité	184
7.3.2	Résidus et estimateur de σ^2	185
7.3.3	Hypothèse Gaussienne	186
7.3.4	Matrice Ω inconnue	186
7.4	Extension des moindres carrés pondérés : la régression locale	191
7.5	Exercices	194
8	Régression biaisée	197
8.1	Régression Ridge	198
8.1.1	Equivalence avec une contrainte sur la norme des coefficients	199
8.1.2	Propriétés statistiques de l'estimateur ridge $\hat{\beta}_{\text{ridge}}$	200
8.1.3	La régression ridge en pratique	202
8.1.4	Exemple des biscuits	205
8.2	Lasso	209
8.2.1	La méthode	209
8.2.2	La régression lasso en pratique	210
8.2.3	Exemple des biscuits	211
8.3	Régression sur composantes principales	213
8.3.1	Hypothèse \mathcal{H}_1 satisfaite : $ X'X \neq 0$	214
8.3.2	Colinéarité parfaite : $ X'X = 0$	215
8.3.3	Pratique de la régression sur composantes principales	217
8.3.4	Exemple des biscuits	221
8.4	Régression aux moindres carrés partiels (PLS)	223
8.4.1	Algorithmes PLS et recherche des composantes	225
8.4.2	Recherche de la taille k	226
8.4.3	Analyse de la qualité du modèle	228
8.4.4	Exemple des biscuits	230
8.5	Exercices	231
A	Corrections des exercices	239
A.1	Régression linéaire simple	239
A.2	Régression linéaire multiple	243
A.3	Inférence dans le modèle gaussien	248
A.4	Validation du modèle	253
A.5	Régression sur variables qualitatives	256
A.6	Choix de variables	262
A.7	Moindres carrés généralisés	264

A.8	Régression biaisée	265
B	Rappels	281
B.1	Rappels d'algèbre	281
B.2	Rappels de Probabilités	285
B.2.1	Généralités	285
B.2.2	Vecteurs aléatoires gaussiens	286
B.3	Tables des lois usuelles	287
B.3.1	Loi Normale $X \sim \mathcal{N}(0, 1)$	287
B.3.2	Loi de Student $X \sim \mathcal{T}_\nu$	288
B.3.3	Loi du Khi-deux à ν ddl $X \sim \chi_\nu^2$	289
B.3.4	Loi de Fisher à ν_1, ν_2 ddl $X \sim \mathcal{F}_{(\nu_1, \nu_2)}$	290
	Bibliographie	293
	Index	295
	Notations	301