

Table des matières

Remerciements	vii
Avant-Propos	ix
I Introduction au modèle linéaire	1
1 La régression linéaire simple	3
1.1 Introduction	3
1.1.1 Un exemple : la pollution de l’air	3
1.1.2 Un second exemple : la hauteur des arbres	5
1.2 Modélisation mathématique	7
1.2.1 Choix du critère de qualité et distance à la droite	7
1.2.2 Choix des fonctions à utiliser	9
1.3 Modélisation statistique	10
1.4 Estimateurs des moindres carrés	11
1.4.1 Calcul des estimateurs de β_j , quelques propriétés	11
1.4.2 Résidus et variance résiduelle	15
1.4.3 Prédiction	15
1.5 Interprétations géométriques	16
1.5.1 Représentation des individus	16
1.5.2 Représentation des variables	17
1.6 Inférence statistique	19
1.7 Exemples	22
1.8 Exercices	29
2 La régression linéaire multiple	31
2.1 Introduction	31
2.2 Modélisation	32
2.3 Estimateurs des moindres carrés	34
2.3.1 Calcul de $\hat{\beta}$	35
2.3.2 Interprétation	37
2.3.3 Quelques propriétés statistiques	38
2.3.4 Résidus et variance résiduelle	40

2.3.5	Prévision	41
2.4	Interprétation géométrique	42
2.5	Exemples	43
2.6	Exercices	47
3	Validation du modèle	51
3.1	Analyse des résidus	52
3.1.1	Les différents résidus	52
3.1.2	Ajustement individuel au modèle, valeur aberrante	53
3.1.3	Analyse de la normalité	54
3.1.4	Analyse de l’homoscédasticité	55
3.1.5	Analyse de la structure des résidus	56
3.2	Analyse de la matrice de projection	59
3.3	Autres mesures diagnostiques	60
3.4	Effet d’une variable explicative	63
3.4.1	Ajustement au modèle	63
3.4.2	Régression partielle : impact d’une variable	64
3.4.3	Résidus partiels et résidus partiels augmentés	65
3.5	Exemple : la concentration en ozone	67
3.6	Exercices	70
4	Extensions : non-inversibilité et (ou) erreurs corrélées	73
4.1	Régression ridge	73
4.1.1	Une solution historique	74
4.1.2	Minimisation des MCO pénalisés	75
4.1.3	Equivalence avec une contrainte sur la norme des coefficients	75
4.1.4	Propriétés statistiques de l’estimateur ridge $\hat{\beta}_{\text{ridge}}$	76
4.2	Erreurs corrélées : moindres carrés généralisés	78
4.2.1	Erreurs hétéroscédastiques	79
4.2.2	Estimateur des moindres carrés généralisés	82
4.2.3	Matrice Ω inconnue	84
4.3	Exercices	85
II	Inférence	89
5	Inférence dans le modèle gaussien	91
5.1	Estimateurs du maximum de vraisemblance	91
5.2	Nouvelles propriétés statistiques	92
5.3	Intervalles et régions de confiance	94
5.4	Prévision	97
5.5	Les tests d’hypothèses	98
5.5.1	Introduction	98
5.5.2	Test entre modèles emboîtés	98
5.6	Applications	102

5.7	Exercices	106
5.8	Notes	109
5.8.1	Intervalle de confiance : bootstrap	109
5.8.2	Test de Fisher pour une hypothèse linéaire quelconque	112
5.8.3	Propriétés asymptotiques	114
6	Variables qualitatives : ANCOVA et ANOVA	117
6.1	Introduction	117
6.2	Analyse de la covariance	119
6.2.1	Introduction : exemple des eucalyptus	119
6.2.2	Modélisation du problème	121
6.2.3	Hypothèse gaussienne	123
6.2.4	Exemple : la concentration en ozone	124
6.2.5	Exemple : la hauteur des eucalyptus	129
6.3	Analyse de la variance à 1 facteur	131
6.3.1	Introduction	131
6.3.2	Modélisation du problème	132
6.3.3	Interprétation des contraintes	134
6.3.4	Estimation des paramètres	134
6.3.5	Hypothèse gaussienne et test d’influence du facteur	136
6.3.6	Exemple : la concentration en ozone	137
6.3.7	Une décomposition directe de la variance	142
6.4	Analyse de la variance à 2 facteurs	142
6.4.1	Introduction	142
6.4.2	Modélisation du problème	143
6.4.3	Estimation des paramètres	145
6.4.4	Analyse graphique de l’interaction	146
6.4.5	Hypothèse gaussienne et test de l’interaction	148
6.4.6	Exemple : la concentration en ozone	150
6.5	Exercices	152
6.6	Note : identifiabilité et contrastes	155
III	Réduction de dimension	157
7	Choix de variables	159
7.1	Introduction	159
7.2	Choix incorrect de variables : conséquences	161
7.2.1	Biais des estimateurs	161
7.2.2	Variance des estimateurs	163
7.2.3	Erreur quadratique moyenne	163
7.2.4	Erreur quadratique moyenne de prévision	166
7.3	Critères classiques de choix de modèles	168
7.3.1	Tests entre modèles emboîtés	169
7.3.2	Le R^2	170

7.3.3	Le R^2 ajusté	171
7.3.4	Le C_p de Mallows	172
7.3.5	Vraisemblance et pénalisation	174
7.3.6	Liens entre les critères	176
7.4	Procédure de sélection	178
7.4.1	Recherche exhaustive	178
7.4.2	Recherche pas à pas	178
7.5	Exemple : la concentration en ozone	180
7.5.1	Variables explicatives quantitatives	180
7.5.2	Intégration de variables qualitatives	183
7.6	Exercices	184
7.7	Note : C_p et biais de sélection	187
8	Régularisation des moindres carrés : Ridge, Lasso et elastic-net	191
8.1	Introduction	191
8.2	Problème du centrage-réduction des variables	194
8.3	Ridge, lasso et elastic-net	195
8.3.1	Régressions avec la package glmnet	199
8.3.2	Interprétation géométrique	202
8.3.3	Simplification quand les X sont orthogonaux	203
8.3.4	Choix du paramètre de régularisation λ	206
8.4	Intégration de variables qualitatives	208
8.5	Exercices	210
8.6	Note : lars et lasso	213
9	Régression sur composantes : PCR et PLS	217
9.1	Régression sur composantes principales (PCR)	218
9.1.1	Changement de base	218
9.1.2	Estimateurs des MCO	219
9.1.3	Choix de composantes/variables	220
9.1.4	Retour aux données d'origine	222
9.1.5	La régression sur composantes en pratique	223
9.2	Régression aux moindres carrés partiels (PLS)	225
9.2.1	Algorithmes PLS	227
9.2.2	Choix de composantes/variables	228
9.2.3	Retour aux données d'origine	229
9.2.4	La régression PLS en pratique	230
9.3	Exercices	231
9.4	Notes	233
9.4.1	ACP et changement de base	233
9.4.2	Colinéarité parfaite : $ X'X = 0$	235

10 Comparaison des différentes méthodes, étude de cas réels	237
10.1 Erreur de prévision et validation croisée	237
10.2 Analyse de l’ozone	241
10.2.1 Préliminaires	241
10.2.2 Méthodes et comparaison	241
10.2.3 Pour aller plus loin	245
10.2.4 Conclusion	248
 IV Le modèle linéaire généralisé	 249
11 Régression logistique	251
11.1 Présentation du modèle	251
11.1.1 Exemple introductif	251
11.1.2 Modélisation statistique	252
11.1.3 Variables explicatives qualitatives, interactions	255
11.2 Estimation	257
11.2.1 La vraisemblance	257
11.2.2 Calcul des estimateurs : l’algorithme IRLS	259
11.2.3 Propriétés asymptotiques de l’EMV	260
11.3 Intervalles de confiance et tests	261
11.3.1 IC et tests sur les paramètres du modèle	262
11.3.2 Test sur un sous-ensemble de paramètres	264
11.3.3 Prévision	267
11.4 Adéquation du modèle	269
11.4.1 Le modèle saturé	270
11.4.2 Tests d’adéquation de la déviance et de Pearson	272
11.4.3 Analyse des résidus	275
11.5 Choix de variables	279
11.5.1 Tests entre modèles emboîtés	279
11.5.2 Procédures automatiques	280
11.6 Exercices	282
 12 Régression de Poisson	 289
12.1 Le modèle linéaire généralisé (GLM)	289
12.2 Exemple : modélisation du nombre de visites	292
12.3 Régression Log-linéaire	295
12.3.1 Le modèle	295
12.3.2 Estimation	296
12.3.3 Tests et intervalles de confiance	297
12.3.4 Choix de variables	302
12.4 Exercices	303

13 Régularisation de la vraisemblance	309
13.1 Régressions ridge, lasso et elastic-net	309
13.2 Choix du paramètre de régularisation λ	313
13.3 Group-lasso	317
13.4 Exercices	319
14 Comparaison en classification supervisée	321
14.1 Préviation en classification supervisée	321
14.2 Performance d’une règle	323
14.2.1 Erreur de classification et <i>accuracy</i>	326
14.2.2 Sensibilité (<i>recall</i>) et taux de faux négatifs	327
14.2.3 Spécificité et taux de faux positifs	327
14.2.4 Mesure sur les tables de contingence	328
14.3 Performance d’un score	329
14.3.1 Courbe ROC	329
14.3.2 Courbe lift	331
14.4 Choix du seuil	332
14.4.1 Respect des proportions initiales	332
14.4.2 Maximisation d’indices ad hoc	332
14.4.3 Maximisation d’un coût moyen	333
14.5 Analyse des données chd	334
14.5.1 Les données	334
14.5.2 Comparaison des algorithmes	334
14.5.3 Pour aller plus loin	340
14.6 Application : détection d’images publicitaires sur internet	346
14.6.1 Les données	346
14.6.2 Ajustement des modèles	347
14.7 Exercices	351
15 Données déséquilibrées	353
15.1 Données déséquilibrées et modèle logistique	353
15.1.1 Un exemple	353
15.1.2 Rééquilibrage pour le modèle logistique	355
15.1.3 Exemples de schéma de rééquilibrage	356
15.2 Stratégies pour données déséquilibrées	361
15.2.1 Quelques méthodes de rééquilibrage	361
15.2.2 Critères pour données déséquilibrées	366
15.3 Choisir un algorithme de rééquilibrage	370
15.3.1 Rééquilibrage et validation croisée	370
15.3.2 Application aux données d’images publicitaires	372
15.4 Exercices	376

V	Introduction à la régression non paramétrique	379
16	Introduction à la régression spline	381
16.1	Introduction	381
16.2	Régression spline	385
16.2.1	Introduction	385
16.2.2	Spline de régression	386
16.3	Spline de lissage	390
16.4	Exercices	393
17	Estimateurs à noyau et k plus proches voisins	395
17.1	Introduction	395
17.2	Estimateurs par moyennes locales	398
17.2.1	Estimateurs à noyau	398
17.2.2	Les k plus proches voisins	402
17.3	Choix des paramètres de lissage	403
17.4	Ecriture multivariée et fléau de la dimension	406
17.4.1	Ecriture multivariée	406
17.4.2	Biais et variance	407
17.4.3	Fléau de la dimension	409
17.5	Exercices	411
A	Rappels	415
A.1	Rappels d'algèbre	415
A.2	Rappels de probabilités	418
	Bibliographie	419