

Table des matières

Remerciements	vii
Avant-Propos	ix
I Les bases de la régression	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 Prévision	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	Note : intervalle de confiance par bootstrap	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	135
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	143
6.4.1	Introduction	143
6.4.2	Modélisation du problème	144
6.4.3	Estimation des paramètres	146
6.4.4	Analyse graphique de l'interaction	147
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
7.6	Exercices	183
7.7	Note : C_p et biais de sélection	185
8	Ridge, Lasso et elastic-net	189
8.1	Introduction	189
8.2	Problème du centrage-réduction des variables	192
8.3	Ridge et lasso	193
8.3.1	Régressions elastic net avec glmnet	197
8.3.2	Interprétation géométrique	200
8.3.3	Simplification quand les X sont orthogonaux	201
8.3.4	Choix du paramètre de régularisation λ	204
8.4	Intégration de variables qualitatives	206
8.5	Exercices	208
8.6	Note : lars et lasso	211
9	Régression sur composantes : PCR et PLS	215
9.1	Régression sur composantes principales (PCR)	216
9.1.1	Changement de base	216
9.1.2	Estimateurs des MCO	217
9.1.3	Choix de composantes/variables	218
9.1.4	Retour aux données d'origine	220
9.2	Régression aux moindres carrés partiels (PLS)	221
9.2.1	Algorithmes PLS	222
9.2.2	Choix de composantes/variables	223
9.2.3	Retour aux données d'origine	224
9.3	Exemple de l'ozone	225
9.4	Exercices	229
9.5	Notes	231
9.5.1	ACP et changement de base	231
9.5.2	Colinéarité parfaite : $ X'X = 0$	232
10	Comparaison des différentes méthodes, étude de cas réels	235
10.1	Erreur de prévision et validation croisée	235
10.2	Analyse de l'ozone	239
10.2.1	Préliminaires	239
10.2.2	Méthodes et comparaison	239
10.2.3	Pour aller plus loin	243

10.2.4 Conclusion	246
IV Le modèle linéaire généralisé	247
11 Régression logistique	249
11.1 Présentation du modèle	249
11.1.1 Exemple introductif	249
11.1.2 Modélisation statistique	250
11.1.3 Variables explicatives qualitatives, interactions	253
11.2 Estimation	255
11.2.1 La vraisemblance	255
11.2.2 Calcul des estimateurs : l'algorithme IRLS	257
11.2.3 Propriétés asymptotiques de l'EMV	258
11.3 Intervalles de confiance et tests	259
11.3.1 IC et tests sur les paramètres du modèle	260
11.3.2 Test sur un sous-ensemble de paramètres	262
11.3.3 Prévision	265
11.4 Adéquation du modèle	267
11.4.1 Le modèle saturé	268
11.4.2 Tests d'adéquation de la déviance et de Pearson	270
11.4.3 Analyse des résidus	272
11.5 Choix de variables	275
11.5.1 Tests entre modèles emboîtés	276
11.5.2 Procédures automatiques	277
11.6 Prévision - scoring	279
11.6.1 Règles de prévision	279
11.6.2 Scoring	282
11.7 Exercices	288
12 Régression de Poisson	295
12.1 Le modèle linéaire généralisé (GLM)	295
12.2 Exemple : modélisation du nombre de visites	298
12.3 Régression Log-linéaire	301
12.3.1 Le modèle	301
12.3.2 Estimation	302
12.3.3 Tests et intervalles de confiance	303
12.3.4 Choix de variables	308
12.4 Exercices	309
13 Régularisation de la vraisemblance	315
13.1 Régressions ridge et lasso	315
13.2 Choix du paramètre de régularisation λ	318
13.3 Group-lasso et elastic net	322
13.3.1 Group-lasso	322

13.3.2 Elastic net	324
13.4 Application : détection d'images publicitaires sur internet	325
13.4.1 Ajustement des modèles	325
13.4.2 Comparaison des modèles	327
13.5 Exercices	329
V Introduction à la régression non paramétrique	331
14 Introduction à la régression spline	333
14.1 Introduction	333
14.2 Régression spline	337
14.2.1 Introduction	337
14.2.2 Spline de régression	338
14.3 Spline de lissage	342
14.4 Exercices	345
15 Estimateurs à noyau et k plus proches voisins	347
15.1 Introduction	347
15.2 Estimateurs par moyennes locales	350
15.2.1 Estimateurs à noyau	350
15.2.2 Les k plus proches voisins	354
15.3 Choix des paramètres de lissage	355
15.4 Ecriture multivariée et fléau de la dimension	358
15.4.1 Ecriture multivariée	358
15.4.2 Biais et variance	359
15.4.3 Fléau de la dimension	361
15.5 Exercices	363
A Rappels	367
A.1 Rappels d'algèbre	367
A.2 Rappels de probabilités	370
Bibliographie	371
Index	375
Notations	383