

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 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	28
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	46
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	71
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
5	Régression polynomiale et régression spline	89
5.1	Régression polynomiale	89
5.2	Régression spline	93
5.2.1	Introduction	93
5.2.2	Spline de régression	94
5.3	Spline de lissage	98
5.4	Exercices	101
II	Inférence	103
6	Inférence dans le modèle gaussien	105
6.1	Estimateurs du maximum de vraisemblance	105

6.2	Nouvelles propriétés statistiques	106
6.3	Intervalles et régions de confiance	108
6.4	Prévision	111
6.5	Les tests d’hypothèses	112
6.5.1	Introduction	112
6.5.2	Test entre modèles emboîtés	113
6.6	Applications	116
6.7	Exercices	120
6.8	Notes	123
6.8.1	Intervalle de confiance : bootstrap	123
6.8.2	Test de Fisher pour une hypothèse linéaire quelconque	125
6.8.3	Propriétés asymptotiques	127
7	Variables qualitatives : ANCOVA et ANOVA	131
7.1	Introduction	131
7.2	Analyse de la covariance	133
7.2.1	Introduction : exemple des eucalyptus	133
7.2.2	Modélisation du problème	135
7.2.3	Hypothèse gaussienne	137
7.2.4	Exemple : la concentration en ozone	138
7.2.5	Exemple : la hauteur des eucalyptus	142
7.3	Analyse de la variance à 1 facteur	144
7.3.1	Introduction	144
7.3.2	Modélisation du problème	145
7.3.3	Interprétation des contraintes	147
7.3.4	Estimation des paramètres	147
7.3.5	Hypothèse gaussienne et test d’influence du facteur	148
7.3.6	Exemple : la concentration en ozone	150
7.3.7	Une décomposition directe de la variance	154
7.4	Analyse de la variance à 2 facteurs	155
7.4.1	Introduction	155
7.4.2	Modélisation du problème	156
7.4.3	Estimation des paramètres	158
7.4.4	Analyse graphique de l’interaction	159
7.4.5	Hypothèse gaussienne et test de l’interaction	160
7.4.6	Exemple : la concentration en ozone	163
7.5	Exercices	164
7.6	Note : identifiabilité et contrastes	167
III	Réduction de dimension	171
8	Choix de variables	173
8.1	Introduction	173
8.2	Notations	175

8.3	Choix incorrect de variables : conséquences	176
8.3.1	Biais des estimateurs	176
8.3.2	Variance des estimateurs	178
8.3.3	Erreur quadratique moyenne	179
8.3.4	Erreur quadratique moyenne de prévision	181
8.4	Critères classiques de choix de modèles	183
8.4.1	Tests entre modèles emboîtés	184
8.4.2	Le R^2	185
8.4.3	Le R^2 ajusté	186
8.4.4	Le C_p de Mallows	187
8.4.5	Vraisemblance et pénalisation	189
8.4.6	Liens entre les critères	191
8.5	Procédure de sélection	193
8.5.1	Recherche exhaustive	193
8.5.2	Recherche pas à pas	193
8.6	Exemple : la concentration en ozone	195
8.6.1	Variables explicatives quantitatives	195
8.6.2	Intégration de variables qualitatives	196
8.7	Exercices	197
8.8	Note : C_p et biais de sélection	199
9	Régularisation des moindres carrés : Ridge, Lasso et elastic-net	203
9.1	Introduction	203
9.2	Problème du centrage-réduction des variables	206
9.3	Ridge, lasso et elastic net	207
9.4	Propriétés des régressions Ridge et lasso	207
9.4.1	Interprétation géométrique	211
9.4.2	Simplification quand les X sont orthogonaux	212
9.4.3	Choix de λ par validation croisée	215
9.5	Régressions avec le package scikitlearn	218
9.5.1	Estimation des paramètres	218
9.5.2	Chemin de régularisation	219
9.5.3	Choix du paramètre de régularisation α	220
9.5.4	Mise en pratique	222
9.6	Intégration de variables qualitatives	223
9.7	Exercices	224
9.8	Note : lars et lasso	228
10	Régression sur composantes : PCR et PLS	231
10.1	Régression sur composantes principales (PCR)	232
10.1.1	Changement de base	232
10.1.2	Estimateurs des MCO	233
10.1.3	Choix de composantes/variables	234
10.1.4	Retour aux données d’origine	236
10.1.5	La régression sur composantes en pratique	237

10.2	Régression aux moindres carrés partiels (PLS)	240
10.2.1	Algorithmes PLS	242
10.2.2	Choix de composantes/variables	242
10.2.3	Retour aux données d’origine	243
10.2.4	La régression PLS en pratique	245
10.3	Exercices	246
10.4	Notes	248
10.4.1	ACP et changement de base	248
10.4.2	Colinéarité parfaite : $ X'X = 0$	249
11	Comparaison des différentes méthodes, étude de cas réels	253
11.1	Erreur de prévision et validation croisée	253
11.2	Analyse de l’ozone	257
11.2.1	Préliminaires	257
11.2.2	Méthodes et comparaison	258
11.3	Transformation des variables : feature engineering	261
11.3.1	Modèle de prévision avec interactions	261
11.3.2	Modèle de prévision avec des polynômes	262
11.3.3	Modèle de prévision avec des splines	262
11.3.4	Modèle de prévision avec des splines et de l’interaction	263
11.3.5	Conclusion	264
IV	Le modèle linéaire généralisé	265
12	Régression logistique	267
12.1	Présentation du modèle	267
12.1.1	Exemple introductif	267
12.1.2	Modélisation statistique	268
12.1.3	Variables explicatives qualitatives, interactions	271
12.2	Estimation	273
12.2.1	La vraisemblance	274
12.2.2	Calcul des estimateurs : l’algorithme IRLS	275
12.2.3	Propriétés asymptotiques de l’EMV	277
12.3	Intervalles de confiance et tests	278
12.3.1	IC et tests sur les paramètres du modèle	278
12.3.2	Test sur un sous-ensemble de paramètres	280
12.3.3	Prévision	282
12.4	Adéquation du modèle	284
12.4.1	Le modèle saturé	285
12.4.2	Tests d’adéquation de la déviance et de Pearson	287
12.4.3	Analyse des résidus	289
12.5	Choix de variables	293
12.5.1	Tests entre modèles emboîtés	294
12.5.2	Procédures automatiques	294

12.6 Exercices	296
13 Régression de Poisson	303
13.1 Le modèle linéaire généralisé (GLM)	303
13.2 Exemple : modélisation du nombre de visites	306
13.3 Régression Log-linéaire	309
13.3.1 Le modèle	309
13.3.2 Estimation	310
13.3.3 Tests et intervalles de confiance	311
13.3.4 Choix de variables	315
13.4 Exercices	316
14 Régularisation de la vraisemblance	321
14.1 Régressions ridge, lasso et elastic-net	321
14.2 Choix du paramètre de régularisation λ	326
14.3 Group-lasso	329
14.4 Exercices	331
15 Comparaison en classification supervisée	333
15.1 Prévision en classification supervisée	333
15.2 Performance d’une règle	335
15.2.1 Erreur de classification et <i>accuracy</i>	338
15.2.2 Sensibilité (<i>recall</i>) et taux de faux négatifs	339
15.2.3 Spécificité et taux de faux positifs	339
15.2.4 Mesure sur les tables de contingence	340
15.3 Performance d’un score	341
15.3.1 Courbe ROC	341
15.3.2 Courbe lift	343
15.4 Choix du seuil	344
15.4.1 Respect des proportions initiales	344
15.4.2 Maximisation d’indices ad hoc	344
15.4.3 Maximisation d’un coût moyen	345
15.5 Analyse des données chd	346
15.5.1 Les données	346
15.5.2 Méthodes et comparaison	346
15.6 Transformation des variables : feature engineering	354
15.6.1 Modèle de prévision avec interactions	354
15.6.2 Modèle de prévision avec des polynômes	354
15.7 Exercices	356
16 Données déséquilibrées	359
16.1 Données déséquilibrées et modèle logistique	359
16.1.1 Un exemple	359
16.1.2 Rééquilibrage pour le modèle logistique	361
16.1.3 Exemples de schéma de rééquilibrage	362

16.2 Stratégies pour données déséquilibrées	366
16.2.1 Quelques méthodes de rééquilibrage	367
16.2.2 Critères pour données déséquilibrées	372
16.3 Choisir un algorithme de rééquilibrage	375
16.3.1 Rééquilibrage et validation croisée	376
16.3.2 Application aux données d’images publicitaires	377
16.4 Exercices	381
A Rappels	383
A.1 Rappels d’algèbre	383
A.2 Rappels de probabilités	386
A.3 Rappels de Python	387
Bibliographie	393
Index	393
Notations	397
Fonctions et packages python	397
Fonctions et packages R	397