

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 | 81 |
| 4.2.3 | Matrice Ω inconnue | 84 |
| 4.3 | Exercices | 85 |
| 5 | Régression polynomiale et régression spline | 87 |
| 5.1 | Régression polynomiale | 87 |
| 5.2 | Régression spline | 91 |
| 5.2.1 | Introduction | 91 |
| 5.2.2 | Spline de régression | 92 |
| 5.3 | Spline de lissage | 96 |
| 5.4 | Exercices | 99 |
| II | Inférence | 101 |
| 6 | Inférence dans le modèle gaussien | 103 |
| 6.1 | Estimateurs du maximum de vraisemblance | 103 |

| | | |
|------------|---|------------|
| 6.2 | Nouvelles propriétés statistiques | 104 |
| 6.3 | Intervalles et régions de confiance | 106 |
| 6.4 | Prévision | 109 |
| 6.5 | Les tests d’hypothèses | 110 |
| 6.5.1 | Introduction | 110 |
| 6.5.2 | Test entre modèles emboîtés | 111 |
| 6.6 | Applications | 114 |
| 6.7 | Exercices | 118 |
| 6.8 | Notes | 121 |
| 6.8.1 | Intervalle de confiance : bootstrap | 121 |
| 6.8.2 | Test de Fisher pour une hypothèse linéaire quelconque | 123 |
| 6.8.3 | Propriétés asymptotiques | 125 |
| 7 | Variables qualitatives : ANCOVA et ANOVA | 129 |
| 7.1 | Introduction | 129 |
| 7.2 | Analyse de la covariance | 131 |
| 7.2.1 | Introduction : exemple des eucalyptus | 131 |
| 7.2.2 | Modélisation du problème | 133 |
| 7.2.3 | Hypothèse gaussienne | 135 |
| 7.2.4 | Exemple : la concentration en ozone | 136 |
| 7.2.5 | Exemple : la hauteur des eucalyptus | 140 |
| 7.3 | Analyse de la variance à 1 facteur | 142 |
| 7.3.1 | Introduction | 142 |
| 7.3.2 | Modélisation du problème | 143 |
| 7.3.3 | Interprétation des contraintes | 145 |
| 7.3.4 | Estimation des paramètres | 145 |
| 7.3.5 | Hypothèse gaussienne et test d’influence du facteur | 147 |
| 7.3.6 | Exemple : la concentration en ozone | 148 |
| 7.3.7 | Une décomposition directe de la variance | 152 |
| 7.4 | Analyse de la variance à 2 facteurs | 153 |
| 7.4.1 | Introduction | 153 |
| 7.4.2 | Modélisation du problème | 154 |
| 7.4.3 | Estimation des paramètres | 156 |
| 7.4.4 | Analyse graphique de l’interaction | 157 |
| 7.4.5 | Hypothèse gaussienne et test de l’interaction | 158 |
| 7.4.6 | Exemple : la concentration en ozone | 161 |
| 7.5 | Exercices | 162 |
| 7.6 | Note : identifiabilité et contrastes | 165 |
| III | Réduction de dimension | 167 |
| 8 | Choix de variables | 169 |
| 8.1 | Introduction | 169 |
| 8.2 | Notations | 171 |

| | | |
|-----------|---|------------|
| 8.3 | Choix incorrect de variables : conséquences | 172 |
| 8.3.1 | Biais des estimateurs | 172 |
| 8.3.2 | Variance des estimateurs | 174 |
| 8.3.3 | Erreur quadratique moyenne | 175 |
| 8.3.4 | Erreur quadratique moyenne de prévision | 177 |
| 8.4 | Critères classiques de choix de modèles | 179 |
| 8.4.1 | Tests entre modèles emboîtés | 180 |
| 8.4.2 | Le R^2 | 181 |
| 8.4.3 | Le R^2 ajusté | 182 |
| 8.4.4 | Le C_p de Mallows | 183 |
| 8.4.5 | Vraisemblance et pénalisation | 185 |
| 8.4.6 | Liens entre les critères | 187 |
| 8.5 | Procédure de sélection | 189 |
| 8.5.1 | Recherche exhaustive | 189 |
| 8.5.2 | Recherche pas à pas | 189 |
| 8.6 | Exemple : la concentration en ozone | 191 |
| 8.6.1 | Variables explicatives quantitatives | 191 |
| 8.6.2 | Intégration de variables qualitatives | 192 |
| 8.7 | Exercices | 193 |
| 8.8 | Note : C_p et biais de sélection | 195 |
| 9 | Régularisation des moindres carrés : Ridge, Lasso et elastic-net | 199 |
| 9.1 | Introduction | 199 |
| 9.2 | Problème du centrage-réduction des variables | 202 |
| 9.3 | Propriétés des régressions Ridge et lasso | 203 |
| 9.3.1 | Interprétation géométrique | 207 |
| 9.3.2 | Simplification quand les X sont orthogonaux | 208 |
| 9.3.3 | Choix de λ par validation croisée | 211 |
| 9.4 | Régularisation avec le module <code>scikitlearn</code> | 213 |
| 9.4.1 | Estimation des paramètres | 214 |
| 9.4.2 | Chemin de régularisation | 215 |
| 9.4.3 | Choix du paramètre de régularisation α | 216 |
| 9.4.4 | Mise en pratique | 218 |
| 9.5 | Intégration de variables qualitatives | 218 |
| 9.6 | Exercices | 221 |
| 9.7 | Note : lars et lasso | 225 |
| 10 | Régression sur composantes : PCR et PLS | 229 |
| 10.1 | Régression sur composantes principales (PCR) | 230 |
| 10.1.1 | Changement de base | 230 |
| 10.1.2 | Estimateurs des MCO | 231 |
| 10.1.3 | Choix de composantes/variables | 232 |
| 10.1.4 | Retour aux données d’origine | 234 |
| 10.1.5 | La régression sur composantes en pratique | 235 |
| 10.2 | Régression aux moindres carrés partiels (PLS) | 238 |

| | | |
|-----------|---|------------|
| 10.2.1 | Algorithmes PLS | 240 |
| 10.2.2 | Choix de composantes/variables | 240 |
| 10.2.3 | Retour aux données d’origine | 241 |
| 10.2.4 | La régression PLS en pratique | 243 |
| 10.3 | Exercices | 244 |
| 10.4 | Notes | 246 |
| 10.4.1 | ACP et changement de base | 246 |
| 10.4.2 | Colinéarité parfaite : $ X'X = 0$ | 247 |
| 11 | Comparaison des différentes méthodes, étude de cas réels | 251 |
| 11.1 | Erreur de prévision et validation croisée | 251 |
| 11.2 | Analyse de l’ozone | 255 |
| 11.2.1 | Préliminaires | 255 |
| 11.2.2 | Méthodes et comparaison | 256 |
| 11.3 | Transformation des variables : feature engineering | 259 |
| 11.3.1 | Modèle de prévision avec interactions | 259 |
| 11.3.2 | Modèle de prévision avec des polynômes | 260 |
| 11.3.3 | Modèle de prévision avec des splines | 260 |
| 11.3.4 | Modèle de prévision avec interactions et splines | 261 |
| 11.3.5 | Conclusion | 262 |
| IV | Le modèle linéaire généralisé | 263 |
| 12 | Régression logistique | 265 |
| 12.1 | Présentation du modèle | 265 |
| 12.1.1 | Exemple introductif | 265 |
| 12.1.2 | Modélisation statistique | 266 |
| 12.1.3 | Variables explicatives qualitatives, interactions | 269 |
| 12.2 | Estimation | 271 |
| 12.2.1 | La vraisemblance | 271 |
| 12.2.2 | Calcul des estimateurs : l’algorithme IRLS | 273 |
| 12.2.3 | Propriétés asymptotiques de l’EMV | 274 |
| 12.3 | Intervalles de confiance et tests | 275 |
| 12.3.1 | IC et tests sur les paramètres du modèle | 276 |
| 12.3.2 | Test sur un sous-ensemble de paramètres | 277 |
| 12.3.3 | Prévision | 280 |
| 12.4 | Adéquation du modèle | 282 |
| 12.4.1 | Le modèle saturé | 283 |
| 12.4.2 | Tests d’adéquation de la déviance et de Pearson | 285 |
| 12.4.3 | Analyse des résidus | 288 |
| 12.5 | Choix de variables | 292 |
| 12.5.1 | Tests entre modèles emboîtés | 292 |
| 12.5.2 | Procédures automatiques | 293 |
| 12.6 | Exercices | 295 |

| | |
|---|------------|
| 13 Régression de Poisson | 301 |
| 13.1 Le modèle linéaire généralisé (GLM) | 301 |
| 13.2 Exemple : modélisation du nombre de visites | 304 |
| 13.3 Régression Log-linéaire | 307 |
| 13.3.1 Le modèle | 307 |
| 13.3.2 Estimation | 308 |
| 13.3.3 Tests et intervalles de confiance | 309 |
| 13.3.4 Choix de variables | 313 |
| 13.4 Exercices | 314 |
| 14 Régularisation de la vraisemblance | 319 |
| 14.1 Régressions ridge, lasso et elastic-net | 319 |
| 14.2 Choix du paramètre de régularisation λ | 324 |
| 14.3 Group-lasso | 327 |
| 14.4 Exercices | 329 |
| 15 Comparaison en classification supervisée | 331 |
| 15.1 Prévision en classification supervisée | 331 |
| 15.2 Performance d’une règle | 333 |
| 15.2.1 Erreur de classification et <i>accuracy</i> | 336 |
| 15.2.2 Sensibilité (<i>recall</i>) et taux de faux négatifs | 337 |
| 15.2.3 Spécificité et taux de faux positifs | 337 |
| 15.2.4 Mesure sur les tables de contingence | 338 |
| 15.3 Performance d’un score | 339 |
| 15.3.1 Courbe ROC | 339 |
| 15.3.2 Courbe lift | 341 |
| 15.4 Choix du seuil | 342 |
| 15.4.1 Respect des proportions initiales | 342 |
| 15.4.2 Maximisation d’indices ad hoc | 342 |
| 15.4.3 Maximisation d’un coût moyen | 343 |
| 15.5 Analyse des données chd | 344 |
| 15.5.1 Les données | 344 |
| 15.5.2 Méthodes et comparaison | 344 |
| 15.6 Transformation des variables : feature engineering | 351 |
| 15.6.1 Modèle de prévision avec interactions | 352 |
| 15.6.2 Modèle de prévision avec des polynômes | 352 |
| 15.7 Exercices | 354 |
| 16 Données déséquilibrées | 357 |
| 16.1 Données déséquilibrées et modèle logistique | 357 |
| 16.1.1 Un exemple | 357 |
| 16.1.2 Rééquilibrage pour le modèle logistique | 359 |
| 16.1.3 Exemples de schéma de rééquilibrage | 360 |
| 16.2 Stratégies pour données déséquilibrées | 365 |
| 16.2.1 Quelques méthodes de rééquilibrage | 365 |

| | |
|---|------------|
| 16.2.2 Critères pour données déséquilibrées | 370 |
| 16.3 Choisir un algorithme de rééquilibrage | 373 |
| 16.3.1 Rééquilibrage et validation croisée | 374 |
| 16.3.2 Application aux données d’images publicitaires | 375 |
| 16.4 Exercices | 379 |
| A Rappels | 381 |
| A.1 Rappels d’algèbre | 381 |
| A.2 Rappels de probabilités | 384 |
| A.3 Modules Python | 385 |
| Bibliographie | 391 |
| Index | 395 |
| Notations | 403 |
| Fonctions et modules python | 405 |