Pratique R

Régression avec R

2e édition

Pierre-André Cornillon
Nicolas Hengartner
Eric Matzner-Løber
Laurent Rouvière

edp sciences

Régression avec R – 2e édition

Pierre-André Cornillon – Nicolas Hengartner
Eric Matzner-Løber – Laurent Rouvière

Performant, évolutif, libre, gratuit et multiplateformes, le logiciel R s'est imposé depuis une dizaine d'années comme un outil de calcul statistique incontournable, tant dans les milieux académiques qu'industriels.

La collection « Pratique R » répond à cette évolution récente et propose d'intégrer pleinement l'utilisation de R dans des ouvrages couvrant les aspects théoriques et pratiques de diverses méthodes statistiques appliquées à des domaines aussi variés que l'analyse des données, la gestion des risques, les sciences médicales, l'économie, etc.

Elle s'adresse aux étudiants, enseignants, ingénieurs, praticiens et chercheurs de ces différents domaines qui utilisent quotidiennement des données dans leur travail et qui apprécient le logiciel R pour sa fiabilité et son confort d'utilisation.

La collection **Pratique R** est dirigée par Pierre-André Cornillon et Eric Matzner-Løber

Cet ouvrage expose, de manière détaillée avec exemples à l'appui, différentes façons de répondre à un des problèmes statistiques les plus courants : la régression.

Cette nouvelle édition se décompose en cinq parties. La première donne les grands principes des régressions simple et multiple par moindres carrés. Les fondamentaux de la méthode, tant au niveau des choix opérés que des hypothèses et leur utilité, sont expliqués. La deuxième partie est consacrée à l'inférence et présente les outils permettant de vérifier les hypothèses mises en œuvre. Les techniques d'analyse de la variance et de la covariance sont également présentées dans cette partie. Le cas de la grande dimension est ensuite abordé dans la troisième partie. Différentes méthodes de réduction de la dimension telles que la sélection de variables, les régressions sous contraintes (lasso, elasticnet ou ridge) et sur composantes (PLS ou PCR) sont notamment proposées. Un dernier chapitre propose des algorithmes (basé sur l'apprentissage/validation ou la validation croisée) qui permettent de comparer toutes ces méthodes. La quatrième partie se concentre sur les modèles linéaires généralisés et plus particulièrement sur les régressions logistique et de Poisson avec ou sans technique de régularisation. Une section particulière est consacrée au scoring en régression logistique. Enfin, la dernière partie présente l'approche non paramétrique à travers les splines, les estimateurs à noyau et des plus proches voisins.

La présentation témoigne d'un réel souci pédagogique des auteurs qui bénéficient d'une expérience d'enseignement auprès de publics très variés. Les résultats exposés sont replacés dans la perspective de leur utilité pratique grâce à l'analyse d'exemples concrets. Les commandes permettant le traitement des exemples sous **R** figurent dans le corps du texte. Enfin, chaque chapitre est complété par une suite d'exercices corrigés. Les codes, les données et les corrections des exercices se trouvent sur le site https://regression-avec-r.github.io/

Cet ouvrage s'adresse principalement à des étudiants de Master et d'écoles d'ingénieurs ainsi qu'aux chercheurs travaillant dans les divers domaines des sciences appliquées.

9 782759 820764

978-2-7598-2076-4
www.edpsciences.org

edp sciences

Régression avec R

2e édition

Pierre-André Cornillon, Nicolas Hengartner,
Eric Matzner-Løber et Laurent Rouvière

Régression avec R

2e édition

edp sciences

ISBN (papier) : 978-2-7598-2076-4 — ISBN (ebook) : 978-2-7598-2183-9

Imprimé en France

Collection Pratique R

dirigée par Pierre-André Cornillon et Eric Matzner-Løber

Université Rennes-2
et ENSAE formation continue Le Cepe, France

Comité éditorial

Déjà paru dans la même collection :

Calcul parallèle avec R
Vincent Miele, Violaine Louvet, 2016
ISBN : 978-2-7598-2060-3 – EDP Sciences

Séries temporelles avec R
Yves Aragon, 2016
ISBN : 978-2-7598-1779-5 – EDP Sciences

Psychologie statistique avec R
Yvonnick Noël, 2015
ISBN : 978-2-7598-1736-8 – EDP Sciences

Réseaux bayésiens avec R
Jean-Baptiste Denis, Marco Scutati, 2014
ISBN : 978-2-7598-1198-4 – EDP Sciences

Analyse factorielle multiple avec R
Jérôme Pagès, 2013
ISBN : 978-2-7598-0963-9 – EDP Sciences

Méthodes de Monte-Carlo avec R
Christian P. Robert, George Casella, 2011
ISBN : 978-2-8178-0181-0 – Springer

REMERCIEMENTS

Cet ouvrage est l'évolution naturelle de la première édition de *Régression avec R*, elle-même issue de *Régression : Théorie et applications*.
Cette nouvelle édition s'appuie toujours sur des exemples concrets et elle n'existerait pas sans ceux-ci. Il est souvent difficile d'obtenir des données réelles pour tester ou présenter des méthodes. Et il est encore plus difficile d'obtenir l'autorisation de les publier. Or nous avons eu la chance d'avoir cette autorisation et des cohortes d'étudiants ont donc analysé des données de pollution et des données d'eucalyptus ! Nous souhaitons profiter de cette nouvelle édition pour renouveler nos sincères remerciement à M. Coron (Association Air Breizh), B. Mallet (CIRAD forêt) et J.-N. Marien (UR2PI) qui nous ont autorisé à utiliser et diffuser leurs données. Nous souhaitons bien sûr associer tous les membres de l'unité de recherche pour la productivité des plantations industrielles (UR2PI), passés ou présents. Les membres de cet organisme de recherche congolais gèrent de nombreux essais, tant génétiques que sylvicoles, et nous renvoyons toutes les personnes intéressées auprès de cet organisme ou auprès du CIRAD, département forêt (wwww.cirad.fr), qui est un des membres fondateurs et un participant actif au sein de l'UR2PI.
Plus de dix ans se sont écoulés depuis les premières versions de cet ouvrage et nous avons eu le plaisir de recevoir de nombreux retours pertinents sur les premières éditions. Les remaniements et l'ajout de nouveaux chapitres comme ceux consacrés au modèle linéaire généralisé, aux méthodes régularisées et à la régression non paramétrique nous ont incités à faire relire ces passages et à en rediscuter d'autres. Les commentaires minutieux et avisés de C. Abraham, N. Chèze, M.-L. Grisoni, P. Lafaye de Micheaux, V. Lefieux, E. Le Pennec nous ont ainsi permis d'améliorer les différents chapitres afin (nous l'espérons) de produire une nouvelle édition plus aboutie. Nous leurs adressons de chaleureux et sincères remerciements.
Nos remerciements vont également à N. Huilleret et C. Ruelle qui nous ont permis de mener à bien le projet de livre et d'édition. Enfin sans la reprise de la collection *Pratique R* par EDP Sciences, ce travail n'existerait pas. Merci donc à F. Citrini et S. Hosotte, pour leur temps, encouragements et patience. Nous remercions également EDP Sciences pour les relectures pertinentes et minutieuses de cet ouvrage.

AVANT-PROPOS

Cette seconde édition est une évolution de la version initiale publiée en 2009. Nous rappelons que cette première version s'inscrivait dans la continuation du livre *Régression : théorie et applications* paru chez Springer-Verlag (Paris). Cette nouvelle édition est plus qu'une mise à jour de la version initiale, la structure a été complètement repensée et de nouvelles parties sont apparues. Par ailleurs, un site web dédié au livre est proposé à l'url `https://regression-avec-r.github.io/`. On pourra notamment y trouver tous les jeux de données et les lignes de code utilisés dans chaque chapitre ainsi que les corrections des exercices.

L'objectif de cet ouvrage est de rendre accessible au plus grand nombre les différentes façons d'aborder un des problèmes auquel le statisticien est très souvent confronté : la *régression*. Les aspects théoriques et pratiques sont simultanément présentés. En effet, comme pour toute méthode statistique, il est nécessaire de comprendre précisément le modèle utilisé pour proposer des résultats pertinents sur des problèmes concrets. Si ces deux objectifs sont atteints, il sera alors aisé de transposer ces acquis à d'autres méthodes, moyennant un investissement modéré. Les grandes étapes – modélisation, estimation, choix de variables, examen de la validité du modèle choisi – restent les mêmes d'une méthode à l'autre. C'est dans cet esprit que cette nouvelle édition a été écrite.

Nous avons donc souhaité un livre avec toute la rigueur scientifique possible mais dont le contenu et les idées ne soient pas noyés dans les démonstrations et les lignes de calculs. Pour cela, seules quelques démonstrations, que nous pensons importantes, sont conservées dans le corps du texte. Les autres résultats sont démontrés à titre d'exercice. Des exercices, de difficultés variables, sont proposés en fin de chapitre. La présence de † indique des exercices plus difficiles. Des questions de cours sous la forme de QCM sont aussi proposées afin d'aider aux révisions du chapitre. Les corrections sont fournies sur le site du livre.

Afin que les connaissances acquises ne restent pas uniquement théoriques, nous avons intégré des exemples traités avec le logiciel libre R. Grâce aux commandes rapportées dans le livre, le lecteur pourra ainsi se familiariser avec le logiciel et retrouver les mêmes résultats que ceux donnés dans le livre. Nous encourageons donc les lecteurs à utiliser les données et les codes afin de s'approprier la théorie mais aussi la pratique.

Cet ouvrage s'adresse aux étudiants des filières scientifiques, élèves ingénieurs, chercheurs dans les domaines appliqués et plus généralement à toutes les personnes confrontés à un problème de régression. Il utilise notamment les notions de modèle, estimateur, biais-variance, intervalle de confiance, test... Pour les lecteurs peu à l'aise avec ces concepts, le livre de Lejeune (2004) pourra constituer une aide précieuse pour certains paragraphes. Cet ouvrage nécessite la connaissance des bases du calcul matriciel : définition d'une matrice, somme, produit, inverse, ainsi que valeurs propres et vecteurs propres. Des résultats classiques sont toutefois rappelés en annexes afin d'éviter de consulter trop souvent d'autres ouvrages.

Le livre se décompose en cinq parties, chacune constituée de deux à quatre chapitres. La première pose les fondamentaux du problème de régression et montre, à

travers quelques exemples, comment on peut l'aborder à l'aide d'un modèle linéaire simple d'abord, puis multiple. Les problèmes d'estimation ainsi que la géométrie associée à la méthode des moindres carrés sont proposés dans les deux premiers chapitres de cette partie. Le troisième chapitre propose les principaux diagnostics qui permettent de s'assurer de la validité du modèle tandis que le dernier présente quelques stratégies à envisager lorsque les hypothèses classiques du modèle linéaire ne sont pas vérifiées.
La seconde partie aborde la partie inférentielle. Il s'agit d'une des parties les plus techniques et calculatoires de l'ouvrage. Cette partie permet, entre autres, d'exposer précisément les procédures de tests et de construction d'intervalles de confiance dans le modèle linéaire. Elle décrit également les spécificités engendrées par l'utilisation de variables qualitatives dans ce modèle.
La troisième partie est consacrée à un problème désormais courant en régression : la réduction de la dimension. En effet, face à l'augmentation conséquente des données, nous sommes de plus en plus confrontés à des problèmes où le nombre de variables est (très) grand. Les techniques standards appliquées à ce type de données se révèlent souvent peu performantes et il est nécessaire de trouver des alternatives. Nous présentons tout d'abord les techniques classiques de choix de variables qui consistent à se donner un critère de performance et à rechercher à l'aide de procédures exhaustives ou pas à pas le sous-groupe de variables qui optimise le critère donné. Nous présentons ensuite les approches régularisées de type Ridge-Lasso qui consistent à trouver les estimateurs qui optimisent le critère des moindres carrés pénalisés par une fonction de la norme des paramètres. Le troisième chapitre propose de faire la régression non pas sur les variables initiales mais sur des combinaisons linéaires de celles-ci. Nous insistons sur la régression sur composantes principales (PCR) et la régression *Partial Least Square* (PLS). A ce stade, nous disposons de plusieurs algorithmes qui répondent à un même problème de régression. Il devient important de se donner une méthode qui permette d'en choisir un automatiquement (on ne laisse pas l'utilisateur décider, ce sont les données qui doivent choisir). Nous proposons un protocole basé sur la minimisation de risques empiriques calculés par des algorithmes de type validation croisée qui permet de choisir l'algorithme le plus approprié pour un problème donné.
Dans la quatrième partie, entièrement nouvelle, nous présentons le modèle linéaire généralisé. Cette partie généralise les modèles initiaux, qui permettaient de traiter uniquement le cas d'une variable à expliquer continue, à des variables à expliquer binaire (régression logistique) ou de comptage (régression de Poisson). Nous insistons uniquement sur les spécificités associées à ces types de variables, la plupart des concepts étudiés précédemment s'adaptent directement à ces cas nouveaux.
Enfin, la cinquième et dernière partie est dédiée à une introduction à l'estimation non paramétrique. Cette partie présente brièvement les estimateurs de type moyennes locales à travers les exemples des splines, estimateurs à noyau et des plus proches voisins. Elle inclut également une discussion sur les avantages et inconvénients d'une telle modélisation face aux modèles paramétriques étudiés précédemment.

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 . . . 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 . . . 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

Première partie

Introduction au modèle linéaire

Notations

β Vecteur de $\mathbb{R}^p$ de coordonnées $(\beta_1, \ldots, \beta_p)$, page 32

$\hat{\beta}_{(i)}$ Estimateur de β dans le modèle linéaire privé de l'observation i, page 52

$\beta_{\bar{j}}$ Vecteur β privé de sa j^{e} coordonnée, page 64

$\text{Cov}(X, Y)$ Covariance entre X et Y, *i.e.* $\mathbb{E}\{(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))'\}$, page 14

$c_{n-p}(1-\alpha)$ Fractile de niveau $(1-\alpha)$ d'une loi de χ^2 à $(n-p)$ ddl, page 21

ddl Degré de liberté, page 20

$\mathbb{E}(X)$ Espérance de X, page 14

$\mathcal{F}_{p,n-p}$ Loi de Fisher à p ddl au numérateur et $(n-p)$ degrés de liberté au dénominateur, page 20

$f_{(p,n-p)}(1-\alpha)$ Fractile de niveau $(1-\alpha)$ d'une loi de Fisher à $(p, n-p)$ ddl, page 20

$\mathcal{H}_2$ $\mathbb{E}(\varepsilon_i) = 0$ pour $i = 1, \cdots, n$ et $\text{Cov}(\varepsilon_i, \varepsilon_j) = \delta_{ij}\sigma^2$, page 38

I_n ou I Matrice identité d'ordre n ou d'ordre dicté par le contexte, page 38

i.i.d. Indépendants et identiquement distribués, page 93

$\Im(X)$ Image de X (matrice $n \times p$) sous-espace de $\mathbb{R}^n$ engendré par les p colonnes de X : $\Im(X) = \{z \in \mathbb{R}^n : \exists \alpha \in \mathbb{R}^p, z = X\alpha\}$, page 35

$\mathcal{N}(0, \sigma^2)$ Loi normale d'espérance nulle et de variance σ^2, page 19

P_X Matrice de projection orthogonale sur $\Im(X)$, page 35

$\Pr(Y \leq y)$ Probabilité que Y soit inférieur ou égal à y, page 189

R^2 Coefficient de détermination, page 18

SCE Somme des carrés expliquée par le modèle, page 18

SCR Somme des carrés résiduelle, page 18

SCT Somme des carrés totale, page 18

$\hat{\sigma}_{(i)}$ Estimateur de σ dans le modèle linéaire privé de l'observation i, page 52

$\mathcal{T}_{n-p}$ Loi de Student à $(n-p)$ degrés de liberté, page 20

$t_{n-p}(1-\alpha/2)$ Fractile de niveau $(1-\alpha/2)$ d'une loi $\mathcal{T}_{n-p}$, page 20

VC Validation croisée, page 52

X $X = (X_1|X_2|\ldots|X_p)$ matrice du plan d'expérience, page 32

x_i' i^{e} ligne de X, page 32

$|\xi|$ Cardinal de ξ un sous-ensemble d'indice de $\{1, 2, \ldots, p\}$, page 162

X_j j^{e} colonne de X, page 32

$X_{\bar{j}}$ Matrice X privée de sa j^{e} colonne, page 64

$\hat{y}_i$ Ajustement de l'individu i, page 15

$\hat{y}_i^p$ Prévision de l'individu i, page 16

$\hat{y}_\xi^p$ Prévision de l'individu $x^\star$ dans le modèle ayant ξ variables explicatives, page 168

$\hat{Y}_\xi^p$ Prévision des $n^\star$ individus de la matrice $X^\star$ dans le modèle à ξ variables, page 168

$\hat{y}(x_\xi)$ Ajustement de l'individu i dans le modèle ayant ξ variables explicatives, page 166

$\hat{Y}(X_\xi)$ Ajustement des n individus de la matrice X dans le modèle à ξ variables, page 166