PratiqueR

Régression avec Python

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

edp sciences

Régression avec Python

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

Performant, évolutif, libre, gratuit, le monde du logiciel libre s'est imposé depuis une vingtaine d'années comme la base des outils de calculs et d'intelligence artificielle déportés dans le cloud ou en local. Parmi les langages de cet écosystème gratuit et multiplateformes R, python ou julia sont devenus des outils incontournables en machine learning, IA, optimisation ou statistiques tant dans les milieux académiques qu'industriels.

La collection « PratiqueR » répond à cette évolution récente et propose d'intégrer pleinement l'utilisation d'un langage 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 ces langages pour leur fiabilité, leur confort d'utilisation et leur extensibilité via des modules ou des packages.

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

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

Cette nouvelle édition se décompose en 4 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, elastic-net ou ridge) et sur composantes (PLS ou PCR) sont notamment proposées. Un dernier chapitre propose des algorithmes, basés sur des méthodes de ré-échantillonnage comme l'apprentissage/validation ou la validation croisée, qui permettent d'établir une comparaison entre toutes ces méthodes.

Enfin la quatrième et dernière partie se concentre sur les modèles linéaires généralisés et plus particulièrement sur les régressions logistiques et de Poisson avec ou sans technique de régularisation. Une section particulière est consacrée aux comparaisons de méthodes en classification supervisée. Elle introduit notamment des critères de performance pour « scorer » des individus comme les courbes ROC et lift et propose des stratégies de choix seuil (Youden, F1...) pour les classer. Ces notions sont ensuite mises en œuvre sur des données réelles afin de sélectionner une méthode de prévision parmi plusieurs algorithmes basés sur des modèles logistiques (régularisés ou non). Une dernière section aborde le problème des données déséquilibrées qui est souvent rencontré en régression binaire.

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 avec python figurent dans le corps du texte. Enfin chaque chapitre est complété par une suite d'exercices. Sur le site du livre, https://regression-avec-python.github.io/, se trouvent les codes, les données et les corrections des exercices.

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 824328
978-2-7598-2432-8
www.edpsciences.org
edp sciences

Régression avec Python

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

Régression avec Python

edp sciences

ISBN (papier) : 978-2-7598-2432-8 — ISBN (ebook) : 978-2-7598-3779-3

© 2025, EDP Sciences, 17, avenue du Hoggar, BP 112, Parc d'activités de Courtaboeuf, 91944 Les Ulis Cedex A

Imprimé en France

Tous droits de traduction, d'adaptation et de reproduction par tous procédés réservés pour tous pays. Toute reproduction ou représentation intégrale ou partielle, par quelque procédé que ce soit, des pages publiées dans le présent ouvrage, faite sans l'autorisation de l'éditeur est illicite et constitue une contrefaçon. Seules sont autorisées, d'une part, les reproductions strictement réservées à l'usage privé du copiste et non destinées à une utilisation collective, et d'autre part, les courtes citations justifiées par le caractère scientifique ou d'information de l'oeuvre dans laquelle elles sont incorporées (art. L. 122-4, L. 122-5 et L. 335-2 du Code de la propriété intellectuelle). Des photocopies payantes peuvent être réalisées avec l'accord de l'éditeur. S'adresser au : Centre français d'exploitation du droit de copie, 3, rue Hautefeuille, 75006 Paris. Tél. : 01 43 26 95 35.

Collection PratiqueR

dirigée par Pierre-André Cornillon et Eric Matzner-Løber

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

Comité éditorial

Eva Cantoni
Institut de recherche en statistique
& Département d'économétrie
Université de Genève, Suisse

Marie Chavent
Institut de Mathématique de Bordeaux,
Centre Inria de l'université de Bordeaux
Talence, France

Rémy Drouilhet
Laboratoire Jean Kuntzmann
Université Pierre Mendes France
Grenoble, France

Ana Karina Fermin Rodriguez
Laboratoire Modal'X
Université Paris Ouest
France

François Husson
Unité Pédagogique de Mathématiques
Appliquées, Institut Agro Rennes-Angers
France

Pierre Lafaye de Micheaux
School of Mathematics and Statistics
UNSW Sydney
Australie

Déjà paru dans la même collection :

Régression avec R, 3e éd. – Pierre-André Cornillon, Nicolas Hengartner, Eric Matzner-Løber, Laurent Rouvière, 2023
ISBN : 978-2-7598-3145-6 – EDP Sciences

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 l'édition *Régression avec R* dans le langage python.
Cette 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 quinze 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. Nous souhaitons remercier tout particulièrement tous nos étudiants qui ont suivi nos cours et qui nous ont aidé à clarifier nos propos. Nous souhaitons également remercier quelques collègues et amis pour leurs commentaires minutieux et avisés : C. Abraham, N. Chèze, M.-L. Grisoni, P. Lafaye de Micheaux, V. Lefieux, E. Le Pennec. Ils 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 surtout leur patience.

AVANT-PROPOS

Cette première édition de régression avec python est une adaptation de la version régression avec R. Un site web dédié au livre est proposé à l'url https://regression-avec-python.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 python. 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... 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 quatre parties, chacune constituée de deux à cinq 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. Les quatrième et cinquième chapitre abordent des extensions du modèle linéaire comme la corrélation des erreurs et les régressions polynomiales et splines.

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, 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.

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 . . . 209
9.3.3 Choix de λ par validation croisée . . . 211
9.4 Régularisation avec le module **scikitlearn** . . . 214
9.4.1 Estimation des paramètres . . . 215
9.4.2 Chemin de régularisation . . . 216
9.4.3 Choix du paramètre de régularisation α . . . 217
9.4.4 Mise en pratique . . . 219
9.5 Intégration de variables qualitatives . . . 219
9.6 Exercices . . . 222
9.7 Note : lars et lasso . . . 226

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 Modification de 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 *accuracy* . . . 336
15.2.2 Sensibilité (*recall*) 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 Modification de 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**

Première partie

Introduction au modèle linéaire

Chapitre 1

La régression linéaire simple

1.1 Introduction

L'origine du mot régression vient de Sir Francis Galton. En 1885, travaillant sur l'hérédité, il chercha à expliquer la taille des fils en fonction de celle des pères. Il constata que lorsque le père était plus grand que la moyenne, *taller than mediocrity*, son fils avait tendance à être plus petit que lui et, *a contrario*, que lorsque le père était plus petit que la moyenne, *shorter than mediocrity*, son fils avait tendance à être plus grand que lui. Ces résultats l'ont conduit à considérer sa théorie de *regression toward mediocrity*. Cependant, l'analyse de causalité entre plusieurs variables est plus ancienne et remonte au milieu du XVIII^{e} siècle. En 1757, R. Boscovich, né à Ragusa, l'actuelle Dubrovnik, proposa une méthode minimisant la somme des valeurs absolues entre un modèle de causalité et les observations. Ensuite Legendre, dans son célèbre article de 1805, « Nouvelles méthodes pour la détermination des orbites des comètes », introduisit la méthode d'estimation par moindres carrés des coefficients d'un modèle de causalité et donna le nom à la méthode. Parallèlement, Gauss publia en 1809 un travail sur le mouvement des corps célestes qui contenait un développement de la méthode des moindres carrés, qu'il affirmait utiliser depuis 1795 (Birkes & Dodge, 1993).
Dans ce chapitre, nous allons analyser la régression linéaire simple : nous pouvons la voir comme une technique statistique permettant de modéliser la relation linéaire entre une variable explicative (notée X) et une variable à expliquer (notée Y). Cette présentation dans un cas simple va nous permettre de bien comprendre les enjeux de cette méthode, les problèmes posés et les réponses apportées.

1.1.1 Un exemple : la pollution de l'air

La pollution de l'air constitue actuellement une des préoccupations majeures de santé publique. De nombreuses études épidémiologiques ont permis de mettre en évidence l'influence sur la santé de certains composés chimiques comme le dioxyde de souffre (SO_2), le dioxyde d'azote (NO_2), l'ozone (O_3) ou des particules sous

forme de poussières contenues dans l'air. L'influence de cette pollution est notable sur les personnes sensibles (nouveau-nés, asthmatiques, personnes âgées). La prévision des pics de concentration de ces composés est donc importante. Nous nous intéressons plus particulièrement à la concentration en ozone. Nous possédons quelques connaissances *a priori* sur la manière dont se forme l'ozone, grâce aux lois régissant les équilibres chimiques. Un des catalyseurs de l'ozone est le rayonnement, cette variable est plus difficile à prévoir que la température et nous allons étudier la concentration de l'ozone, qui est fonction de la température ; plus la température est élevée, plus la concentration en ozone est importante. Cette relation très vague doit être améliorée afin de pouvoir prédire les pics d'ozone.
Afin de mieux comprendre ce phénomène, l'association Air Breizh (surveillance de la qualité de l'air en Bretagne) mesure depuis 1994 la concentration en O_3 (en μg/ml) toutes les 10 minutes. Le maximum journalier de la concentration en O_3 sera noté `O3`. Air Breizh collecte également à certaines heures de la journée des données météorologiques comme la température, la nébulosité, le vent... Les données sont disponibles en ligne (voir Avant-propos). Le tableau suivant donne les 5 premières mesures effectuées.

Individu	`O3`	`T12`
1	63.6	13.4
2	89.6	15
3	79	7.9
4	81.2	13.1
5	88	14.1

Tableau 1.1 – 5 données de température à 12 h et teneur maximale en ozone.

Nous allons donc chercher à expliquer le maximum de `O3` de la journée par la température à 12 h. Le but de cette régression est double :
— ajuster un modèle pour expliquer la concentration en `O3` en fonction de `T12` ;
— prédire les valeurs de concentration en `O3` pour de nouvelles valeurs de `T12`.
Avant toute analyse, il est intéressant de représenter les données.

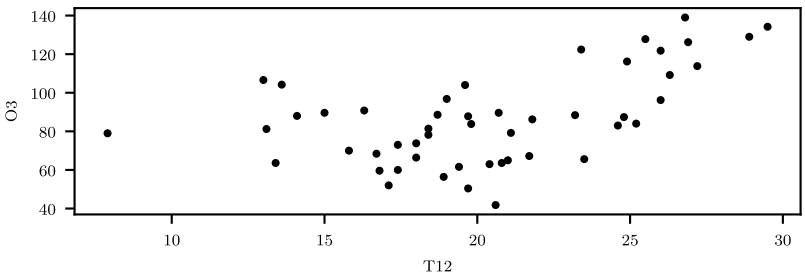


Fig. 1.1 – 50 données journalières de température et `O3`.

Chaque point du graphique (fig.1.1) représente, pour un jour donné, une mesure de la température à 12 h et le pic d'ozone de la journée.
Pour analyser la relation entre les x_i (température) et les y_i (ozone), nous allons chercher une fonction f telle que

$$y_i \approx f(x_i).$$

Pour définir $\approx$, il faut donner un critère quantifiant la qualité de l'ajustement de la fonction f aux données et une classe de fonctions $\mathcal{G}$ dans laquelle est supposée se trouver la vraie fonction inconnue. Le problème mathématique peut s'écrire de la façon suivante :

$$\underset{f \in \mathcal{G}}{\operatorname{argmin}} \sum_{i=1}^{n} l(y_i - f(x_i)), \tag{1.1}$$

où n représente le nombre de données à analyser et $l(.)$ est appelée *fonction de coût* ou encore *fonction de perte*.

1.1.2 Un second exemple : la hauteur des arbres

Cet exemple utilise des données fournies par l'UR2PI et le CIRAD forêt (voir Remerciements). Lorsque le forestier évalue la vigueur d'une forêt, il considère souvent la hauteur des arbres qui la composent. Plus les arbres sont hauts, plus la forêt ou la plantation produit. Afin de calculer le volume de l'arbre, il est nécessaire d'avoir sa hauteur et d'utiliser ensuite une formule du type « tronc de cône ». Cependant, mesurer la hauteur d'un arbre d'une vingtaine de mètres n'est pas aisé et demande un dendromètre. Ce type d'appareil mesure un angle entre le sol et le sommet de l'arbre. Il nécessite donc une vision claire de la cime de l'arbre et un recul assez grand afin d'avoir une mesure précise de l'angle et donc de la hauteur.
Dans certains cas, il est impossible de mesurer la hauteur, car ces deux conditions ne sont pas réunies, ou la mesure demande quelquefois trop de temps ou encore le forestier n'a pas de dendromètre. Il est alors nécessaire d'estimer la hauteur grâce à une mesure simple, la mesure de la circonférence à 1 mètre 30 du sol.
Nous possédons des mesures sur des eucalyptus dans une parcelle plantée et nous souhaitons à partir de ces mesures élaborer un modèle de prévision de la hauteur. Les eucalyptus étant plantés pour servir de matière première dans la pâte à papier, ils sont vendus au volume de bois. Il est donc important de connaître le volume et la hauteur, afin d'évaluer la réserve en matière première dans la plantation (ou volume sur pied total). Les surfaces plantées sont énormes, il n'est pas question de prendre trop de temps pour la mesure et prévoir la hauteur par la circonférence est une méthode permettant la prévision du volume sur pied. La parcelle d'intérêt est constituée d'eucalyptus de 6 ans, âge de « maturité » des eucalyptus, c'est-à-dire l'âge en fin de rotation avant la coupe. Dans cette parcelle, nous avons alors mesuré $n = 1429$ couples circonférence-hauteur. Le tableau suivant donne les 5 premières mesures effectuées.

Individu	ht	circ
1	18.25	36
2	19.75	42
3	16.50	33
4	18.25	39
5	19.50	43

Tableau 1.2 – Hauteur et circonférence (`ht` et `circ`) des 5 premiers eucalyptus.

Nous souhaitons donc expliquer la hauteur par la circonférence. Avant toute modélisation, nous représentons les données. Chaque point du graphique 1.2 représente une mesure du couple circonférence/hauteur sur un eucalyptus.

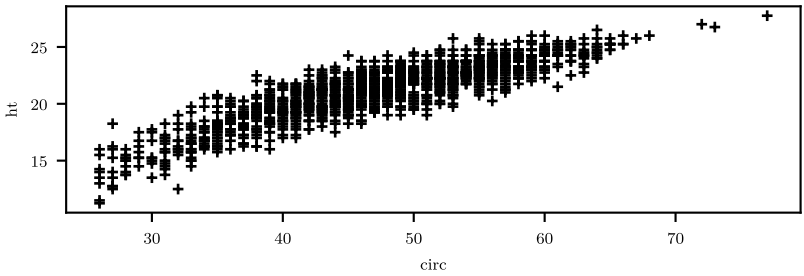


Fig. 1.2 – Représentation des mesures pour les $n = 1429$ eucalyptus mesurés.

Pour prévoir la hauteur en fonction de la circonférence, nous allons donc chercher une fonction f telle que

$$y_i \approx f(x_i)$$

pour chaque mesure $i \in \{1, \ldots, 1429\}$.
A nouveau, afin de quantifier le symbole $\approx$, nous allons choisir une classe de fonctions $\mathcal{G}$. Cette classe représente toutes les fonctions d'ajustement possible pour modéliser la hauteur en fonction de la circonférence. Puis nous cherchons la fonction de $\mathcal{G}$ qui soit la plus proche possible des données selon une fonction de coût. Cela s'écrit

$$\underset{f \in \mathcal{G}}{\operatorname{argmin}} \sum_{i=1}^{n} l(y_i - f(x_i)),$$

où n représente le nombre de données à analyser et $l(.)$ la *fonction de coût*.

Remarque
Le calcul du volume proposé ici est donc fait en deux étapes : dans la première on estime la hauteur et dans la seconde on utilise une formule de type « tronc de cône » pour calculer le volume avec la hauteur estimée et la circonférence. Une autre méthode de calcul de volume consiste à ne pas utiliser de formule incluant

la hauteur et prévoir directement le volume en une seule étape. Pour cela il faut calibrer le volume en fonction de la circonférence et il faut donc la mesure de nombreux volumes en fonction de circonférences, ce qui est très coûteux et difficile à réactualiser.

1.2 Modélisation mathématique

Nous venons de voir que le problème mathématique peut s'écrire de la façon suivante (voir équation 1.1) :

$$\underset{f\in\mathcal{G}}{\operatorname{argmin}} \sum_{i=1}^{n} l(y_i - f(x_i)),$$

où $l(.)$ est appelée *fonction de coût* et $\mathcal{G}$ un ensemble de fonctions données. Dans la suite de cette section, nous allons discuter du choix de la fonction de coût et de l'ensemble $\mathcal{G}$. Nous présenterons des graphiques illustratifs bâtis à partir de 10 données fictives de température et de concentration en ozone.

1.2.1 Choix du critère de qualité et distance à la droite

De nombreuses fonctions de coût $l(.)$ existent, mais les deux principales utilisées sont les suivantes :

— $l(u) = u^2$ coût quadratique ;
— $l(u) = |u|$ coût absolu.

Ces deux fonctions sont représentées sur le graphique 1.3 :

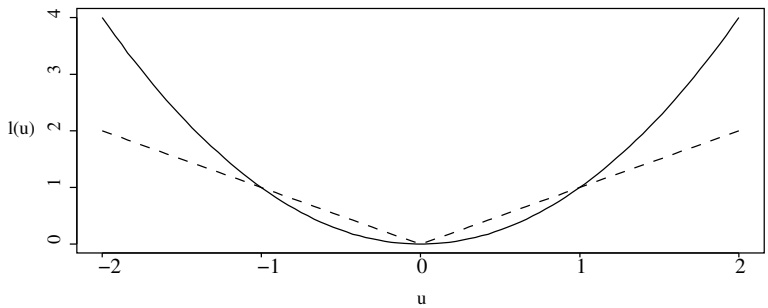


Fig. 1.3 – Coût absolu (pointillés) et coût quadratique (trait plein).

Ces fonctions sont positives, symétriques, elles donnent donc la même valeur lorsque l'erreur est positive ou négative et s'annulent lorsque u vaut zéro. De manière plus générale, l n'est pas obligatoirement symétrique mais est une fonction convexe positive et peut être vue comme la distance entre une observation (x_i, y_i) et son point correspondant sur la droite $(x_i, f(x_i))$ (voir fig. 1.4).

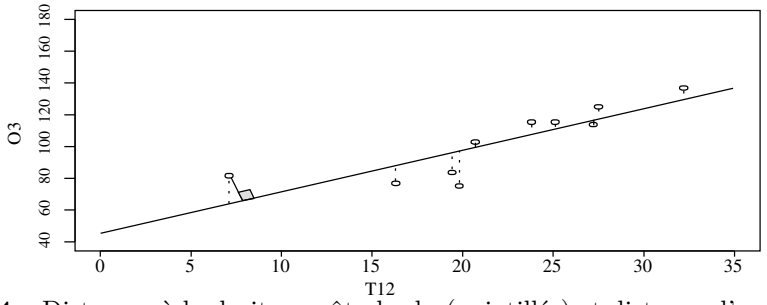


Fig. 1.4 – Distances à la droite : coût absolu (pointillés) et distance d'un point à une droite.

Par point correspondant, nous entendons « évalué » à la même valeur x_i. Nous aurions pu prendre comme critère à minimiser la somme des distances des points (x_i, y_i) à la droite [1] (voir fig. 1.4), mais ce type de distance n'entre pas dans le cadre des fonctions de coût puisqu'au point (x_i, y_i) correspond sur la droite un point $(x_i', f(x_i'))$ d'abscisse et d'ordonnée différentes.
Il est évident que, par rapport au coût absolu, le coût quadratique accorde une importance plus grande aux points qui restent éloignés de la droite ajustée, la distance étant élevée au carré (voir fig. 1.3). Sur l'exemple fictif, dans la classe $\mathcal{G}$ des fonctions linéaires, nous allons minimiser $\sum_{i=1}^{n}(y_i - f(x_i))^2$ (coût quadratique) et $\sum_{i=1}^{n}|y_i - f(x_i)|$ (coût absolu). Les droites ajustées sont représentées sur le graphique ci-dessous :

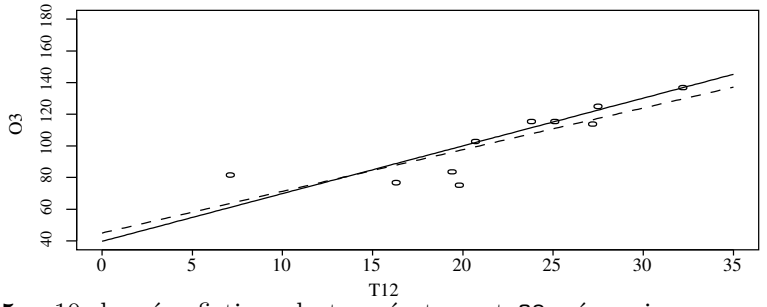


Fig. 1.5 – 10 données fictives de température et O3, régressions avec un coût absolu (trait plein) et quadratique (pointillés).

La droite ajustée avec un coût quadratique propose un compromis où aucun point n'est très éloigné de la droite : le coût quadratique est sensible aux points aberrants qui sont éloignés de la droite. Ainsi (fig. 1.5) le premier point d'abscisse approximative 7°C est assez éloigné des autres. La droite ajustée avec un coût quadratique lui accorde une plus grosse importance que l'autre droite et passe relativement donc plus près de lui. En enlevant ce point (de manière imaginaire),

1. La distance d'un point à une droite est la longueur de la perpendiculaire à cette droite passant par ce point.

la droite ajustée risque d'être très différente : le point est dit influent et le coût quadratique peu robuste. Le coût absolu est plus robuste et la modification d'une observation modifie moins la droite ajustée. Les notions de points influents, points aberrants, seront approfondies au chapitre 3.
Malgré cette non-robustesse, le coût quadratique est le coût le plus souvent utilisé, pour plusieurs raisons : historique, calculabilité, propriétés mathématiques. En 1800, il n'existait pas d'ordinateur et l'utilisation du coût quadratique permettait de calculer explicitement les estimateurs à partir des données. A propos de l'utilisation d'autres fonctions de coût, voici ce que disait Gauss (1809) : « Mais de tous ces principes, celui des moindres carrés est le plus simple : avec les autres, nous serions conduits aux calculs les plus complexes ». En conclusion, *seul le coût quadratique sera automatiquement utilisé dans la suite de ce livre, sauf mention contraire*. Les lecteurs intéressés par le coût absolu peuvent consulter le livre de Dodge & Rousson (2004).

1.2.2 Choix des fonctions à utiliser

Si la classe $\mathcal{G}$ est trop large, par exemple la classe des fonctions continues ($\mathcal{C}_0$), un grand nombre de fonctions de cette classe minimisent le critère (1.1). Ainsi toutes les fonctions de la classe qui passent par tous les points (interpolation), quand c'est possible, annulent la quantité $\sum_{i=1}^{n} l(y_i - f(x_i))$.

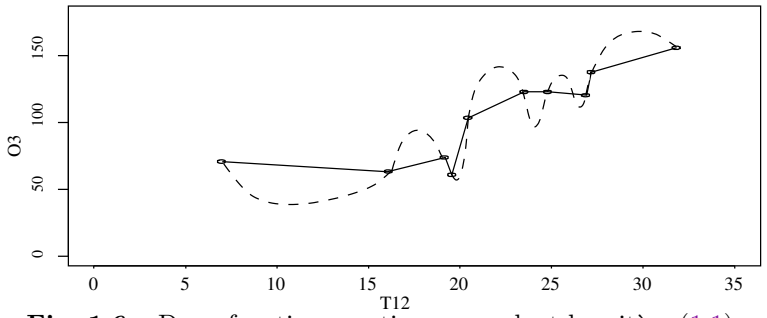


Fig. 1.6 – Deux fonctions continues annulant le critère (1.1).

La fonction continue tracée en pointillés sur la figure (fig. 1.6) semble inappropriée bien qu'elle annule le critère (1.1). La fonction continue tracée en traits pleins annule aussi le critère (1.1). D'autres fonctions continues annulent ce critère, donc la classe des fonctions continues est trop vaste. Ces fonctions passent par tous les points et c'est là leur principal défaut. Nous souhaitons plutôt une courbe, ne passant pas par tous les points, mais possédant un trajet harmonieux, sans trop de détours. Bien sûr le trajet sans aucun détour est la ligne droite et la classe $\mathcal{G}$ la plus simple sera l'ensemble des fonctions affines. Par abus de langage, on emploie le terme de fonctions linéaires. D'autres classes de fonctions peuvent être choisies et ce choix est en général dicté par une connaissance *a priori* du phénomène et (ou) par l'observation des données.

Ainsi une étude de régression linéaire simple débute toujours par une représentation graphique des observations (x, y), appelée nuage de points. Cette première représentation permet de savoir si le modèle linéaire est pertinent. Le graphique (fig. 1.7) représente trois nuages de points différents.

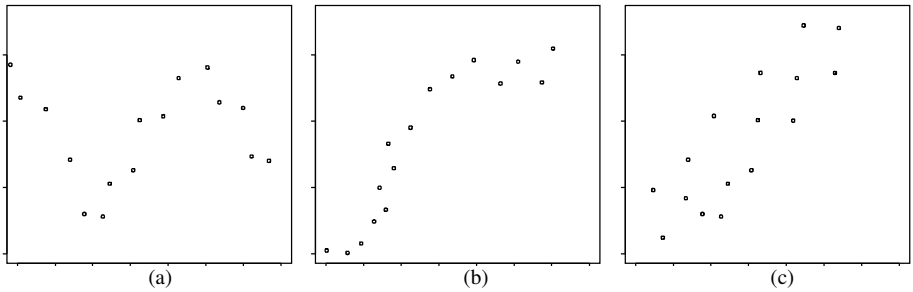


Fig. 1.7 – Exemples de nuages de points.

Au vu du graphique, il semble inadéquat de proposer une régression linéaire pour les graphiques (a) et (b), le tracé présentant une forme sinusoïdale ou sigmoïdale. Par contre, la modélisation par une droite de la relation entre X et Y pour le graphique (c) semble correspondre à la réalité de la liaison. Dans la suite de ce chapitre, nous prendrons $\mathcal{G} = \left\{ f : f(x) = ax + b, \quad (a, b) \in \mathbb{R}^2 \right\}$.

1.3 Modélisation statistique

Lorsque nous ajustons par une droite les données, nous supposons implicitement que leur relation était de la forme

$$Y = \beta_1 + \beta_2 X.$$

Dans l'exemple de l'ozone, nous supposons donc un modèle où la concentration d'ozone dépend linéairement de la température. Nous savons pertinemment que toutes les observations mesurées ne sont pas sur la droite. D'une part, il est irréaliste de croire que la concentration de l'ozone dépend linéairement de la température et de la température seulement. D'autre part, les mesures effectuées dépendent de la précision de l'appareil de mesure, de l'opérateur et il peut arriver que pour des valeurs identiques de la variable X, nous observions des valeurs différentes pour Y.
Nous supposons alors que la concentration d'ozone dépend linéairement de la température mais cette liaison est perturbée par un « bruit ». Nous supposons en fait que les données suivent le modèle suivant :

$$Y = \beta_1 + \beta_2 X + \varepsilon. \tag{1.2}$$

L'équation (1.2) est appelée **modèle de régression linéaire** et dans ce cas précis **modèle de régression linéaire simple**. Les β_j, appelés les paramètres du modèle (constante de régression et coefficient de régression), sont fixes mais inconnus,

et nous voulons les estimer. La quantité notée ε est appelée bruit, ou erreur, et est aléatoire et inconnue.
Afin d'estimer les paramètres inconnus du modèle, nous mesurons dans le cadre de la régression simple une seule variable explicative ou variable exogène X et une variable à expliquer ou variable endogène Y. La variable X est souvent considérée comme non aléatoire au contraire de Y. Nous mesurons alors n observations de la variable X, notées x_i, où i varie de 1 à n, et n valeurs de la variable à expliquer Y notées y_i.
Nous supposons que nous avons collecté n couples de données (x_i, y_i) où y_i est la réalisation de la variable aléatoire Y_i. Par abus de notation, nous confondrons la variable aléatoire Y_i et sa réalisation, l'observation y_i. Avec la notation ε_i, nous confondrons la variable aléatoire avec sa réalisation. Suivant le modèle (1.2), nous pouvons écrire

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i, \qquad i = 1, \cdots, n$$

où

- les x_i sont des observations connues non aléatoires et non toutes identiques ;
- les paramètres β_j, $j = 1, 2$ du modèle sont inconnus ;
- les ε_i sont les réalisations inconnues d'une variable aléatoire ;
- les y_i sont les observations d'une variable aléatoire.

1.4 Estimateurs des moindres carrés

Définition 1.1 (estimateurs des MC)
On appelle estimateurs des moindres carrés (MC) de β_1 et β_2, les estimateurs $\hat{\beta}_1$ et $\hat{\beta}_2$ obtenus par minimisation de la quantité

$$S(\beta_1, \beta_2) = \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i)^2 = \|Y - \beta_1 \mathbb{1} - \beta_2 X\|^2,$$

où $\mathbb{1}$ est le vecteur de $\mathbb{R}^n$ dont tous les coefficients valent 1. Les estimateurs peuvent également s'écrire sous la forme suivante :

$$(\hat{\beta}_1, \hat{\beta}_2) = \underset{(\beta_1, \beta_2) \in \mathbb{R} \times \mathbb{R}}{\operatorname{argmin}} S(\beta_1, \beta_2).$$

1.4.1 Calcul des estimateurs de β_j, quelques propriétés

La fonction $S(\beta_1, \beta_2)$ est strictement convexe. Si elle admet un point singulier, celui-ci correspond à l'unique minimum. Annulons les dérivées partielles, nous obtenons un système d'équations appelées équations normales :

$$\begin{cases} \dfrac{\partial S(\hat{\beta}_1, \hat{\beta}_2)}{\partial \beta_1} &= -2 \displaystyle\sum_{i=1}^{n} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0, \\ \dfrac{\partial S(\hat{\beta}_1, \hat{\beta}_2)}{\partial \beta_2} &= -2 \displaystyle\sum_{i=1}^{n} x_i (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0. \end{cases}$$

La première équation donne

$$\hat{\beta}_1 n + \hat{\beta}_2 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

et nous avons un estimateur de l'ordonnée à l'origine

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}, \tag{1.3}$$

où $\bar{x} = \sum_{i=1}^{n} x_i/n$. La seconde équation donne

$$\hat{\beta}_1 \sum_{i=1}^{n} x_i + \hat{\beta}_2 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i.$$

En remplaçant $\hat{\beta}_1$ par son expression (1.3), nous avons une première écriture de

$$\hat{\beta}_2 = \frac{\sum x_i y_i - \sum x_i \bar{y}}{\sum x_i^2 - \sum x_i \bar{x}},$$

et en utilisant astucieusement la nullité de la somme $\sum(x_i - \bar{x})$, nous avons d'autres écritures pour l'estimateur de la pente de la droite

$$\hat{\beta}_2 = \frac{\sum x_i(y_i - \bar{y})}{\sum x_i(x_i - \bar{x})} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})(x_i - \bar{x})} = \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2}. \tag{1.4}$$

Pour obtenir ce résultat, nous supposons qu'il existe au moins deux points d'abscisses différentes. Cette hypothèse notée $\mathcal{H}_1$ s'écrit $x_i \neq x_j$ pour au moins deux individus. Elle permet d'obtenir l'unicité des coefficients estimés $\hat{\beta}_1, \hat{\beta}_2$.

Une fois déterminés les estimateurs $\hat{\beta}_1$ et $\hat{\beta}_2$, nous pouvons estimer la droite de régression par la formule

$$\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X.$$

Si nous évaluons la droite aux points x_i ayant servi à estimer les paramètres, nous obtenons des $\hat{y}_i$ et ces valeurs sont appelées les valeurs ajustées. Si nous évaluons la droite en d'autres points, les valeurs obtenues seront appelées les valeurs prévues ou prévisions. Représentons les points initiaux et la droite de régression estimée. La droite de régression passe par le centre de gravité du nuage de points $(\bar{x}, \bar{y})$ (fig. 1.8) comme l'indique l 'équation (1.3).

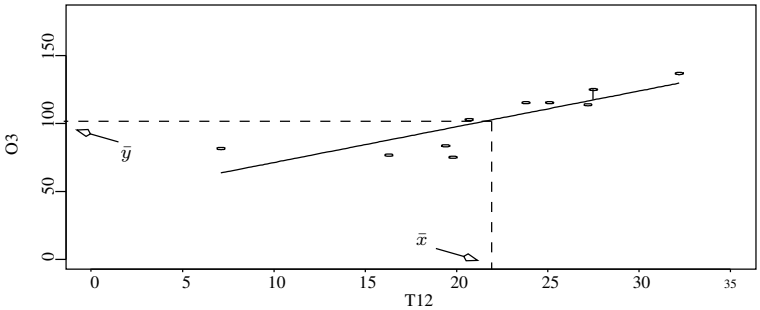


Fig. 1.8 – Nuage de points, droite de régression et centre de gravité.

Nous avons réalisé une expérience et avons mesuré n valeurs (x_i, y_i). A partir de ces n valeurs, nous avons obtenu un estimateur de β_1 et de β_2. Si nous refaisions une expérience, nous mesurerions n nouveaux couples de données (x_i, y_i). A partir de ces données, nous aurions un nouvel estimateur de β_1 et de β_2. Les estimateurs sont fonction des données mesurées et changent donc avec les observations collectées (fig. 1.9). Les vraies valeurs de β_1 et β_2 sont inconnues et ne changent pas.

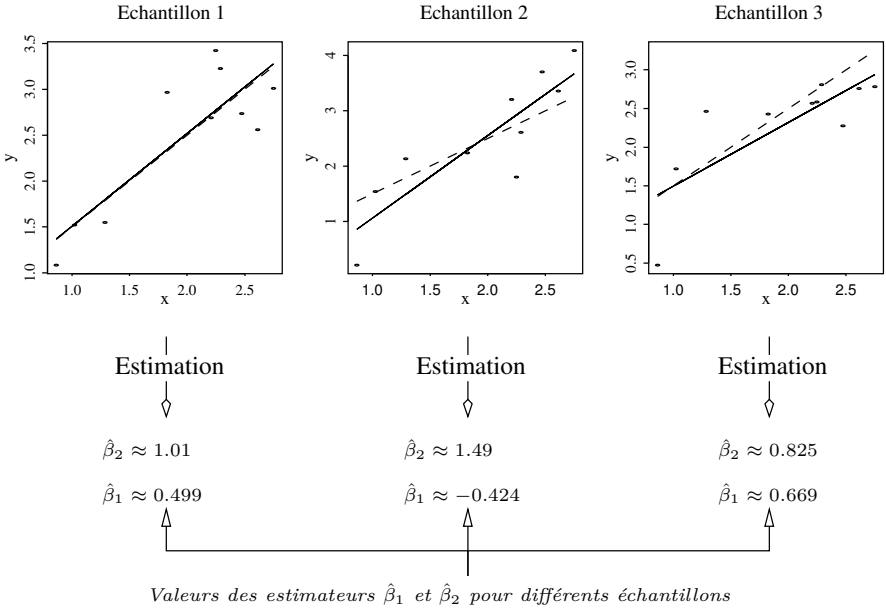


Fig. 1.9 – Exemple de la variabilité des estimations. Le vrai modèle est $Y = X + 0.5 + \varepsilon$, où ε est choisi comme suivant une loi $\mathcal{N}(0, 0.25)$. Nous avons ici 3 répétitions de la mesure de 10 points (x_i, y_i), ou 3 échantillons de taille 10. Le trait en pointillé représente la vraie droite de régression et le trait plein son estimation.

Le statisticien cherche en général à vérifier que les estimateurs utilisés admettent certaines propriétés comme :

— un estimateur $\hat{\beta}$ est-il sans biais ? Par définition $\hat{\beta}$ est sans biais si $\mathbb{E}(\hat{\beta}) = \beta$. En moyenne sur toutes les expériences possibles de taille n, l'estimateur $\hat{\beta}$ moyen sera égal à la valeur inconnue du paramètre. En français, cela signifie qu'en moyenne $\hat{\beta}$ « tombe » sur β ;
— un estimateur $\hat{\beta}$ est-il de variance minimale parmi les estimateurs d'une classe définie ? En d'autres termes, parmi tous les estimateurs de la classe, l'estimateur utilisé admet-il parmi toutes les expériences la plus petite variabilité ?

Pour cela, nous supposons une seconde hypothèse notée $\mathcal{H}_2$: les erreurs sont centrées, de même variance (homoscédasticité) et non corrélées entre elles. Elle permet de calculer les propriétés statistiques des estimateurs. $\mathcal{H}_2$: $\mathbb{E}(\varepsilon_i) = 0$, pour $i = 1, \cdots, n$ et $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \delta_{ij}\sigma^2$, où $\mathbb{E}(\varepsilon)$ est l'espérance de ε, $\mathrm{Cov}(\varepsilon_i, \varepsilon_j)$ est la covariance entre ε_i et ε_j et $\delta_{ij} = 1$ lorsque $i = j$ et $\delta_{ij} = 0$ lorsque $i \neq j$. Nous avons la première propriété de ces estimateurs (voir exercice 1.2).

Proposition 1.1 (Biais des estimateurs)
$\hat{\beta}_1$ et $\hat{\beta}_2$ estiment sans biais β_1 et β_2, c'est-à-dire que $\mathbb{E}(\hat{\beta}_1) = \beta_1$ et $\mathbb{E}(\hat{\beta}_2) = \beta_2$.

Les estimateurs $\hat{\beta}_1$ et $\hat{\beta}_2$ sont sans biais, nous allons nous intéresser à leur variance. Afin de montrer que ces estimateurs sont de variances minimales dans leur classe, nous allons d'abord calculer leur variance (voir exercices 1.3, 1.4). C'est l'objet de la prochaine proposition.

Proposition 1.2 (Variances de $\hat{\beta}_1$ et $\hat{\beta}_2$)
Les variances et covariance des estimateurs des paramètres valent :

$$\begin{aligned} \mathrm{V}(\hat{\beta}_2) &= \frac{\sigma^2}{\sum(x_i - \bar{x})^2} \\ \mathrm{V}(\hat{\beta}_1) &= \frac{\sigma^2 \sum x_i^2}{n\sum(x_i - \bar{x})^2} \\ \mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2) &= -\frac{\sigma^2 \bar{x}}{\sum(x_i - \bar{x})^2}. \end{aligned}$$

Cette proposition nous permet d'envisager la précision de l'estimation en utilisant la variance. Plus la variance est faible, plus l'estimateur sera précis. Pour avoir des variances petites, il faut avoir un numérateur petit et (ou) un dénominateur grand. Les estimateurs seront donc de faibles variances lorsque :
— la variance σ^2 est faible. Cela signifie que la variance de Y est faible et donc les mesures sont proches de la droite à estimer ;
— la quantité $\sum(x_i - \bar{x})^2$ est grande, les mesures x_i doivent être dispersées autour de leur moyenne ;
— la quantité $\sum x_i^2$ ne doit pas être trop grande, les points doivent avoir une faible moyenne en valeur absolue. En effet, nous avons

$$\frac{\sum x_i^2}{\sum(x_i - \bar{x})^2} = \frac{\sum x_i^2 - n\bar{x}^2 + n\bar{x}^2}{\sum(x_i - \bar{x})^2} = 1 + \frac{n\bar{x}^2}{\sum(x_i - \bar{x})^2}.$$

L'équation (1.3) indique que la droite des MC passe par le centre de gravité du nuage $(\bar{x}, \bar{y})$. Supposons $\bar{x}$ positif, alors si nous augmentons la pente, l'ordonnée à l'origine va diminuer et vice versa. Nous retrouvons donc le signe négatif pour la covariance entre $\hat{\beta}_1$ et $\hat{\beta}_2$.
Nous terminons cette partie concernant les propriétés par le théorème de Gauss-Markov qui indique que, parmi tous les estimateurs linéaires sans biais, les estimateurs des MC possèdent la plus petite variance (voir exercice 1.5).

Théorème 1.1 (Gauss-Markov)
Parmi les estimateurs sans biais linéaires en Y, les estimateurs $\hat{\beta}_j$ sont de variance minimale.

1.4.2 Résidus et variance résiduelle

Nous avons estimé β_1 et β_2. La variance σ^2 des ε_i est le dernier paramètre inconnu à estimer. Pour cela, nous allons utiliser les résidus : ce sont des estimateurs des erreurs inconnues ε_i.

Définition 1.2 (Résidus)
Les résidus sont définis par

$$\hat{\varepsilon}_i = y_i - \hat{y}_i$$

où $\hat{y}_i$ est la valeur ajustée de y_i par le modèle, c'est-à-dire $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$.

Nous avons la propriété suivante (voir exercice 1.6).

Proposition 1.3
Dans un modèle de régression linéaire simple, la somme des résidus est nulle.

Intéressons-nous maintenant à l'estimation de σ^2 et construisons un estimateur sans biais $\hat{\sigma}^2$ (voir exercice 1.7) :

Proposition 1.4 (Estimateur de la variance du bruit)
La statistique $\hat{\sigma}^2 = \sum_{i=1}^{n} \hat{\varepsilon}_i^2/(n-2)$ est un estimateur sans biais de σ^2.

1.4.3 Prévision

Un des buts de la régression est de proposer des prévisions pour la variable à expliquer Y. Soit x_{n+1} une nouvelle valeur de la variable X, nous voulons prédire y_{n+1}. Le modèle indique que

$$y_{n+1} = \beta_1 + \beta_2 x_{n+1} + \varepsilon_{n+1}$$

avec $\mathbb{E}(\varepsilon_{n+1}) = 0$, $\mathrm{V}(\varepsilon_{n+1}) = \sigma^2$ et $\mathrm{Cov}(\varepsilon_{n+1}, \varepsilon_i) = 0$ pour $i = 1, \cdots, n$. Nous pouvons prédire la valeur correspondante grâce au modèle estimé

$$\hat{y}_{n+1}^p = \hat{\beta}_1 + \hat{\beta}_2 x_{n+1}.$$

En utilisant la notation $\hat{y}^p_{n+1}$ nous souhaitons insister sur la notion de prévision : la valeur pour laquelle nous effectuons la prévision, ici la $(n+1)^{\text{e}}$, n'a pas servi dans le calcul des estimateurs. Remarquons que cette quantité sera différente de la valeur ajustée, notée $\hat{y}_i$, qui elle fait intervenir la i^{e} observation.

Deux types d'erreurs vont entacher notre prévision, l'une due à la non-connaissance de ε_{n+1} et l'autre due à l'estimation des paramètres.

Proposition 1.5 (Variance de la prévision $\hat{y}^p_{n+1}$)
La variance de la valeur prévue de $\hat{y}^p_{n+1}$ vaut

$$\mathrm{V}\left(\hat{y}^p_{n+1}\right) = \sigma^2\left(\frac{1}{n} + \frac{(x_{n+1}-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right).$$

La variance de $\hat{y}^p_{n+1}$ (voir exercice 1.8) nous donne une idée de la stabilité de l'estimation. En prévision, on s'intéresse généralement à l'erreur que l'on commet entre la vraie valeur à prévoir y_{n+1} et celle que l'on prévoit $\hat{y}^p_{n+1}$. L'erreur peut être simplement résumée par la différence entre ces deux valeurs, c'est ce que nous appellerons l'erreur de prévision. Cette erreur de prévision permet de quantifier la capacité du modèle à prévoir. Nous avons sur ce thème la proposition suivante (voir exercice 1.8).

Proposition 1.6 (Erreur de prévision)
L'erreur de prévision, définie par $\hat{\varepsilon}^p_{n+1} = y_{n+1} - \hat{y}^p_{n+1}$ satisfait les propriétés suivantes :

$$\begin{aligned}\mathbb{E}(\hat{\varepsilon}^p_{n+1}) &= 0\\ \mathrm{V}(\hat{\varepsilon}^p_{n+1}) &= \sigma^2\left(1 + \frac{1}{n} + \frac{(x_{n+1}-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right).\end{aligned}$$

Remarque
La variance augmente lorsque x_{n+1} s'éloigne du centre de gravité du nuage. Effectuer une prévision lorsque x_{n+1} est « loin » de $\bar{x}$ est donc périlleux, la variance de l'erreur de prévision peut alors être très grande !

1.5 Interprétations géométriques

1.5.1 Représentation des individus

Pour chaque individu, ou observation, nous mesurons une valeur x_i et une valeur y_i. Une observation peut donc être représentée dans le plan, nous dirons alors que $\mathbb{R}^2$ est l'espace des observations. $\hat{\beta}_1$ correspond à l'ordonnée à l'origine alors que $\hat{\beta}_2$ représente la pente de la droite ajustée. Cette droite minimise la somme des carrés des distances verticales des points du nuage à la droite ajustée (fig. 1.10).

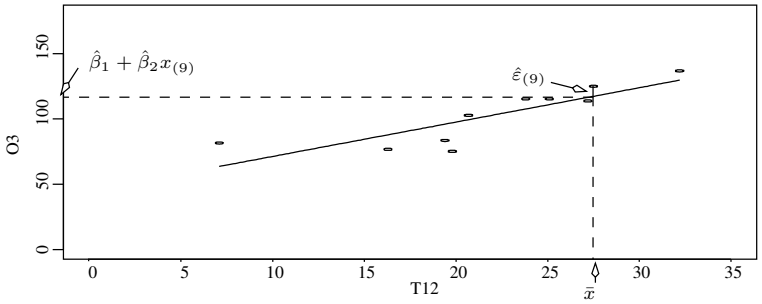


Fig. 1.10 – Représentation des individus.

Les couples d'observations (x_i, y_i) avec $i = 1, \ldots, n$ ordonnées suivant les valeurs croissantes de x sont notés $(x_{(i)}, y_{(i)})$. Nous avons représenté la neuvième valeur de x et sa valeur ajustée $\hat{y}_{(9)} = \hat{\beta}_1 + \hat{\beta}_2 x_{(9)}$ sur le graphique, ainsi que le résidu correspondant $\hat{\varepsilon}_{(9)}$.

1.5.2 Représentation des variables

Nous pouvons voir le problème d'une autre façon. Nous mesurons n couples de points (x_i, y_i). La variable X et la variable Y peuvent être considérées comme deux vecteurs possédant n coordonnées. Le vecteur X (respectivement Y) admet pour coordonnées les observations $x_1, x_2, \ldots, x_n$ (respectivement $y_1, y_2, \ldots, y_n$). Ces deux vecteurs d'observations appartiennent au même espace $\mathbb{R}^n$: l'espace des variables. Nous pouvons donc représenter les données dans l'espace des variables. Le vecteur $\mathbb{1}$ est également un vecteur de $\mathbb{R}^n$ dont toutes les composantes valent 1. Les 2 vecteurs $\mathbb{1}$ et X engendrent un sous-espace de $\mathbb{R}^n$ de dimension 2. Nous avons supposé que $\mathbb{1}$ et X ne sont pas colinéaires grâce à $\mathcal{H}_1$ mais ces vecteurs ne sont pas obligatoirement orthogonaux. Ces vecteurs sont orthogonaux lorsque $\bar{x}$, la moyenne des observations $x_1, x_2, \ldots, x_n$ vaut zéro.
La régression linéaire peut être vue comme la projection orthogonale du vecteur Y dans le sous-espace de $\mathbb{R}^n$ engendré par $\mathbb{1}$ et X, noté $\Im(X)$ (voir fig. 1.11).

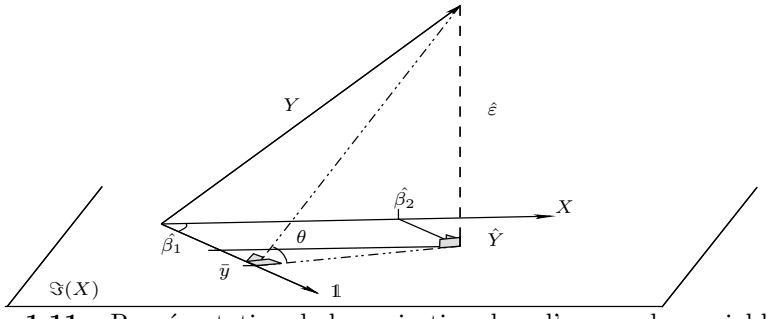


Fig. 1.11 – Représentation de la projection dans l'espace des variables.

Les coefficients $\hat{\beta}_1$ et $\hat{\beta}_2$ s'interprètent comme les composantes de la projection orthogonale notée $\hat{Y}$ de Y sur ce sous-espace.

Remarque
Les vecteurs $\mathbb{1}$ et X de normes respectives $\sqrt{n}$ et $\sqrt{\sum_{i=1}^{n} x_i^2}$ ne forment pas une base orthogonale. Afin de savoir si ces vecteurs sont orthogonaux, calculons leur produit scalaire. Le produit scalaire est la somme du produit terme à terme des composantes des deux vecteurs et vaut ici $\sum_{i=1}^{n} x_i \times 1 = n\bar{x}$. Les vecteurs forment une base orthogonale lorsque la moyenne de X est nulle. En effet $\bar{x}$ vaut alors zéro et le produit scalaire est nul. Les vecteurs n'étant en général pas orthogonaux, cela veut dire que $\hat{\beta}_1\mathbb{1}$ n'est pas la projection de Y sur la droite engendrée par $\mathbb{1}$ et que $\hat{\beta}_2 X$ n'est pas la projection de Y sur la droite engendrée par X. Nous reviendrons sur cette différence au chapitre suivant.

Un modèle, que l'on qualifiera de bon, possédera des estimations $\hat{y}_i$ proches des vraies valeurs y_i. Sur la représentation dans l'espace des variables (fig. 1.11) la qualité peut être évaluée par l'angle θ. Cet angle est compris entre -90 degrés et 90 degrés. Un angle proche de -90 degrés ou de 90 degrés indique un modèle de mauvaise qualité. Le cosinus carré de θ est donc une mesure possible de la qualité du modèle et cette mesure varie entre 0 et 1.
Le théorème de Pythagore nous donne directement que

$$\begin{aligned} \|Y - \bar{y}\mathbb{1}\|^2 &= \|\hat{Y} - \bar{y}\mathbb{1}\|^2 + \|\hat{\varepsilon}\|^2 \\ \sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}\hat{\varepsilon}_i^2 \\ \mathrm{SCT} &= \mathrm{SCE} + \mathrm{SCR}, \end{aligned}$$

où SCT (respectivement SCE et SCR) représente la somme des carrés totale (respectivement expliquée par le modèle et résiduelle).

Le coefficient de détermination R^2 est défini par

$$\mathrm{R}^2 = \frac{\mathrm{SCE}}{\mathrm{SCT}} = \frac{\|\hat{Y} - \bar{y}\mathbb{1}\|^2}{\|Y - \bar{y}\mathbb{1}\|^2},$$

c'est-à-dire la part de la variabilité expliquée par le modèle sur la variabilité totale. De nombreux logiciels multiplient cette valeur par 100 afin de donner un pourcentage.

Remarques
Dans ce cas précis, R^2 est le carré du coefficient de corrélation empirique entre les x_i et les y_i et correspond au cosinus carré de l'angle θ :
— si $\mathrm{R}^2 = 1$, le modèle explique tout, l'angle θ vaut donc zéro, Y est dans $\Im(X)$ c'est-à-dire que $y_i = \beta_1 + \beta_2 x_i$;
— si $\mathrm{R}^2 = 0$, cela veut dire que $\sum(\hat{y}_i - \bar{y})^2 = 0$ et donc que $\hat{y}_i = \bar{y}$. Le modèle de régression linéaire est inadapté ;

— si R^2 est proche de zéro, cela veut dire que Y est quasiment dans l'orthogonal de $\Im(X)$, le modèle de régression linéaire est inadapté, la variable X utilisée n'explique pas la variable Y.

1.6 Inférence statistique

Jusqu'à présent, nous avons pu, en choisissant une fonction de coût quadratique, ajuster un modèle de régression, à savoir calculer $\widehat{\beta}_1$ et $\widehat{\beta}_2$. Grâce aux coefficients estimés, nous pouvons donc prédire, pour chaque nouvelle valeur x_{n+1} une valeur de la variable à expliquer $\widehat{y}^p_{n+1}$ qui est tout simplement le point sur la droite ajustée correspondant à l'abscisse x_{n+1}. En ajoutant l'hypothèse $\mathcal{H}_2$, nous avons pu calculer l'espérance et la variance des estimateurs. Ces propriétés permettent d'appréhender de manière grossière la qualité des estimateurs proposés. Le théorème de Gauss-Markov permet de juger de la qualité des estimateurs parmi une classe d'estimateurs : les estimateurs linéaires sans biais. Enfin ces deux hypothèses nous ont aussi permis de calculer l'espérance et la variance de la valeur prédite $\widehat{y}^p_{n+1}$.
Cependant, nous souhaitons en général connaître la loi des estimateurs afin de calculer des intervalles ou des régions de confiance ou effectuer des tests. Il faut donc introduire une hypothèse supplémentaire concernant la loi des ε_i. L'hypothèse $\mathcal{H}_2$ devient

$$\mathcal{H}_3 \left\{ \begin{array}{ll} \varepsilon_i & \sim \mathcal{N}(0, \sigma^2) \\ \varepsilon_i & \text{sont indépendants} \end{array} \right.$$

où $\mathcal{N}(0, \sigma^2)$ est une loi normale d'espérance nulle et de variance σ^2. Le modèle de régression devient le modèle paramétrique $\left\{\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n}, \mathcal{N}(\beta_1 + \beta_2 x, \sigma^2)\right\}$, où β_1, β_2, σ^2 sont à valeurs dans $\mathbb{R}$, $\mathbb{R}$ et $\mathbb{R}^+$ respectivement. La loi des ε_i étant connue, nous en déduisons la loi des y_i. Toutes les preuves de cette section seront détaillées au chapitre 6.
Nous allons envisager dans cette section les propriétés supplémentaires des estimateurs qui découlent de l'hypothèse $\mathcal{H}_3$ (normalité et indépendance des erreurs) :
— lois des estimateurs $\hat{\beta}_1$, $\hat{\beta}_2$ et $\hat{\sigma}^2$;
— intervalles de confiance univariés et bivariés ;
— loi des valeurs prévues $\hat{y}^p_{n+1}$ et intervalle de confiance.
Cette partie est plus technique que les parties précédentes. Afin de faciliter la lecture, considérons les notations suivantes :

$$\begin{array}{rclcrcl} \sigma^2_{\hat{\beta}_1} & = & \sigma^2 \left(\dfrac{\sum x_i^2}{n \sum (x_i - \bar{x})^2} \right), & & \hat{\sigma}^2_{\hat{\beta}_1} & = & \hat{\sigma}^2 \left(\dfrac{\sum x_i^2}{n \sum (x_i - \bar{x})^2} \right), \\ \sigma^2_{\hat{\beta}_2} & = & \dfrac{\sigma^2}{\sum (x_i - \bar{x})^2}, & & \hat{\sigma}^2_{\hat{\beta}_2} & = & \dfrac{\hat{\sigma}^2}{\sum (x_i - \bar{x})^2}, \end{array}$$

où $\hat{\sigma}^2 = \sum \hat{\varepsilon}_i^2/(n-2)$. Cet estimateur est donné au théorème 1.4. Notons que les estimateurs de la colonne de gauche ne sont pas réellement des estimateurs.

En effet puisque σ^2 est inconnu, ces estimateurs ne sont pas calculables avec les données. Cependant, ce sont eux qui interviennent dans les lois des estimateurs $\hat{\beta}_1$ et $\hat{\beta}_2$ (voir proposition 1.7). Les estimateurs donnés dans la colonne de droite sont ceux qui sont utilisés (et utilisables) et ils consistent simplement à remplacer σ^2 par $\hat{\sigma}^2$. Les lois des estimateurs sont données dans la proposition suivante.

Proposition 1.7 (Lois des estimateurs : variance connue)
Les lois des estimateurs des MC sont :
(i) $\hat{\beta}_j \sim \mathcal{N}\left(\beta_j, \sigma^2_{\hat{\beta}_j}\right)$ *pour* $j = 1, 2$.
(ii) $\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} \sim \mathcal{N}\left(\beta, \sigma^2 V\right)$, $\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$ *et* $V = \dfrac{1}{\sum (x_i - \bar{x})^2} \begin{bmatrix} \sum x_i^2/n & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}$.
(iii) $\dfrac{(n-2)}{\sigma^2}\hat{\sigma}^2$ *suit une loi du* χ^2 *à* $(n-2)$ *degrés de liberté (ddl)* (χ^2_{n-2}).
(iv) $(\hat{\beta}_1, \hat{\beta}_2)$ *et* $\hat{\sigma}^2$ *sont indépendants.*

La variance σ^2 n'est pas connue en général, nous l'estimons par $\hat{\sigma}^2$. Les estimateurs des MC ont alors les propriétés suivantes :

Proposition 1.8 (Lois des estimateurs : variance estimée)
Lorsque σ^2 *est estimée par* $\hat{\sigma}^2$, *nous avons*
(i) pour $j = 1, 2$ $\dfrac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim \mathcal{T}_{n-2}$ *où* $\mathcal{T}_{n-2}$ *est une loi de Student à* $(n-2)$ *ddl.*
(ii) $\dfrac{1}{2\,\hat{\sigma}^2}(\hat{\beta} - \beta)' V^{-1} (\hat{\beta} - \beta) \sim \mathcal{F}_{2,n-2}$, *où* $\mathcal{F}_{2,n-2}$ *est une loi de Fisher à 2 ddl au numérateur et* $(n-2)$ *ddl au dénominateur.*

Ces dernières propriétés nous permettent de donner des intervalles de confiance (IC) ou des régions de confiance (RC) des paramètres inconnus. En effet, la valeur ponctuelle d'un estimateur est en général insuffisante et il est nécessaire de lui adjoindre un intervalle de confiance. Nous parlerons d'IC quand un paramètre est univarié et de RC quand le paramètre est multivarié.

Proposition 1.9 (IC et RC de niveau $1-\alpha$ pour les paramètres)
(i) Un IC *bilatéral de* β_j $(j \in \{1, 2\})$ *est donné par :*

$$\left[\hat{\beta}_j - t_{n-2}(1-\alpha/2)\hat{\sigma}_{\hat{\beta}_j}, \hat{\beta}_j + t_{n-2}(1-\alpha/2)\hat{\sigma}_{\hat{\beta}_j}\right] \tag{1.5}$$

où $t_{n-2}(1-\alpha/2)$ *est le fractile de niveau* $(1-\alpha/2)$ *d'une loi* $\mathcal{T}_{n-2}$.
(ii) Une RC des deux paramètres inconnus β *est donnée par l'équation suivante :*

$$\frac{1}{2\hat{\sigma}^2}\left[n(\hat{\beta}_1 - \beta_1)^2 + 2n\bar{x}(\hat{\beta}_1 - \beta_1)(\hat{\beta}_2 - \beta_2) + \sum x_i^2 (\hat{\beta}_2 - \beta_2)^2\right] \leq f_{(2,n-2)}(1-\alpha),$$

$f_{(2,n-2)}(1-\alpha)$ *est le fractile de niveau* $(1-\alpha)$ *d'une loi de Fisher à* $(2, n-2)$ *ddl.*

(iii) Un IC *de* σ^2 *est donné par :*

$$\left[\frac{(n-2)\hat{\sigma}^2}{c_{n-2}(1-\alpha/2)}, \frac{(n-2)\hat{\sigma}^2}{c_{n-2}(\alpha/2)}\right],$$

où $c_{n-2}(1-\alpha/2)$ représente le fractile de niveau $(1-\alpha/2)$ d'une loi du χ^2 à $(n-2)$ degrés de liberté.

Remarque
La propriété (ii) donne la RC simultanée des paramètres de la régression $\beta = (\beta_1, \beta_2)'$, appelée ellipse de confiance grâce à la loi du couple. Au contraire (i) donne l'IC d'un paramètre sans tenir compte de la corrélation entre $\hat{\beta}_1$ et $\hat{\beta}_2$. Il est donc délicat de donner une RC du vecteur (β_1, β_2) en juxtaposant les deux IC.

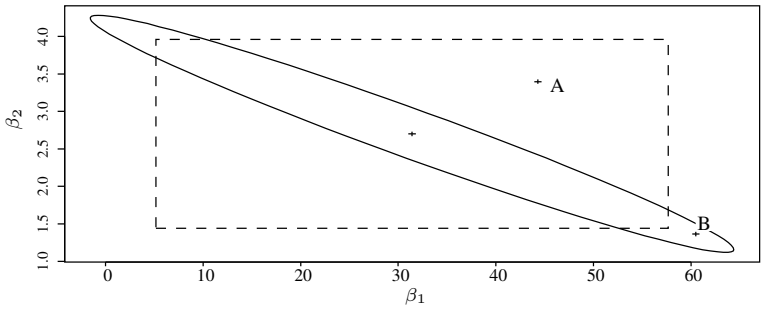


Fig. 1.12 – Comparaison entre ellipse et rectangle de confiance.

Un point peut avoir chaque coordonnée dans son IC respectif mais ne pas appartenir à l'ellipse de confiance. Le point A est un exemple de ce type de point. *A contrario*, un point peut appartenir à la RC sans qu'aucune de ses coordonnées n'appartienne à son IC respectif (le point B). L'ellipse de confiance n'est pas toujours calculée par les logiciels de statistique. Le rectangle de confiance obtenu en juxtaposant les deux intervalles de confiance peut être une bonne approximation de l'ellipse si la corrélation entre $\hat{\beta}_1$ et $\hat{\beta}_2$ est faible (fig. 1.12).

Nous pouvons donner un intervalle de confiance de la droite de régression.

Proposition 1.10 (IC pour $\mathbb{E}(y_i)$)
Un IC *de $\mathbb{E}(y_i) = \beta_1 + \beta_2 x_i$ est donné par :*

$$\left[\hat{y}_i \pm t_{n-2}(1-\alpha/2)\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_l - \bar{x})^2}}\right]. \tag{1.6}$$

En calculant les IC pour tous les points de la droite, nous obtenons une hyperbole de confiance. Lorsque x_i est proche de $\bar{x}$, le terme dominant de la variance est $1/n$, et dès que x_i s'éloigne de $\bar{x}$, le terme dominant est le terme au carré. Nous avons les mêmes résultats que ceux obtenus à la section (1.4.3). Enonçons le résultat permettant de calculer un intervalle de confiance pour une valeur prévue :

Proposition 1.11 (IC pour y_{n+1})
Un IC *de y_{n+1} est donné par :*

$$\left[\hat{y}^p_{n+1} \pm t_{n-2}(1-\alpha/2)\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum(x_i - \bar{x})^2}}\right]. \tag{1.7}$$

Plus le point à prévoir est éloigné de $\bar{x}$, plus la variance de la prévision et donc l'IC seront grands. Intuitivement, plus une observation est éloignée du centre de gravité, moins nous avons d'information sur elle. Lorsque x_{n+1} est à l'intérieur de l'étendue des x_i, le terme dominant de la variance est la valeur 1 et donc la variance est relativement constante. Lorsque x_{n+1} est en dehors de l'étendue des x_i, le terme dominant peut être le terme au carré, et la forme de l'intervalle sera à nouveau une hyperbole.

1.7 Exemples

La concentration en ozone

Nous allons traiter les 50 données journalières de concentration en ozone. La variable à expliquer est la concentration en ozone notée `O3` et la variable explicative est la température notée `T12`. Avant toute chose, nous devons importer les modules (voir A.3, p. 385) afin d'utiliser Python.

- Nous importons les données

```
ozone = pd.read_csv("ozone_simple.txt", header=0, sep=";")
```

- Nous représentons ensuite les données

```
plt.plot(ozone.T12, ozone.O3, "o")
plt.ylabel("O3")
plt.xlabel("T12")
```

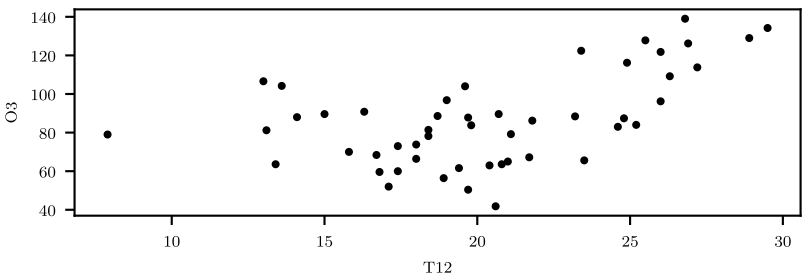


Fig. 1.13 – 50 données journalières de `T12` et `O3`.

Ce graphique permet de vérifier visuellement si une régression linéaire est pertinente. Autrement dit, il suffit de regarder si le nuage de points s'étire le long d'une droite. Bien qu'ici il semble que le nuage s'étire sur une première droite jusqu'à 22 ou 23 degrés C puis selon une autre droite pour les hautes valeurs de températures, nous pouvons tenter une régression linéaire simple.