Modélisation statistique

Table des matières

Remarques

Bien que les diapositives illustrent l'implémentation des techniques statistiques et des modèles à l'aide de **SAS**, ces notes présentent le pendant **R** : visitez le site web du projet **R** pour télécharger le logiciel. L'interface graphique la plus populaire (et celle que je vous recommande) est RStudio Desktop.

Ce cours traite de modélisation des données et une citation célèbre de George Box dit que « tous les modèles sont faux, mais certains sont utiles ». Ce point de vue est réducteur ; Peter McCullagh et John Nelder (traduction libre) expliquent dans le préambule de leur livre

> La modélisation en science demeure, du moins partiellement, un art. Certains principes existent, en revanche, pour guider le modélisateur. Le premier est que tous les modèles sont faux; mais que **certains sont meilleurs** et **le modélisateur doit chercher le meilleur à sa portée**. En même temps, il est sage de reconnaître que la quête perpétuelle de la vérité n'est pas envisageable.

Et David R. Cox (traduction libre), de rajouter

> . . . il n'est pas utile de simplement énoncer que tout modèle est faux. L'idée même de modèle sous-tend une notion de simplification et d'idéalisation. L'idée qu'un système physique, biologique ou sociologique complexe puisse être décrit de manière exacte par quelques formules est franchement absurde. La construction de **représentations idéalisées qui capturent les aspects stables les plus importants du système** est néanmoins une partie essentielle de toute analyse scientifique et les modèles statistiques ne diffèrent pas en cela d'autres types de modèles.

1 Introduction à l'inférence statistique

Ce chapitre porte sur deux concepts fondamentaux pour la modélisation, à savoir les principes sous-jacents aux tests d'hypothèses et l'analyse exploratoire des données.

L'inférence statistique a pour but de tirer des conclusions formelles à partir de données. Dans le cadre de la recherche scientifique, le chercheur formule une hypothèse, collecte des données et conclut quant à la plausibilité de son hypothèse.

On distingue deux types de jeux de données : les données **expérimentales** sont typiquement collectées en milieu contrôlé suivant un protocole d'enquête et un plan d'expérience : elles servent à répondre à une question prédéterminée. L'approche expérimentale est désirable pour éviter le «jardin des embranchements» (une allégorie signifiant qu'un chercheur peut raffiner son hypothèse à la lumière des données, sans ajustement pour des variables confondantes), mais elle n'est pas toujours réalisable : par exemple, un économiste ne peut pas modifier les taux d'intérêts pour observer les impacts sur le taux d'épargne des consommateurs. Lorsque les données ont été collectées préalablement à d'autres fins, on parle de données **observationnelles**.

Par modèle, on entendra la spécification d'une loi aléatoire pour les données et une équation reliant les paramètres ou l'espérance conditionnelle d'une variable réponse Y à un ensemble de variables explicatives $\mathbf{X}$. Ce modèle peut servir à des fins de prédiction (modèle prédictif) ou pour tester des hypothèses de recherche concernant les effets de ces variables (modèle explicatif). Ces deux objectifs ne sont pas mutuellement exclusifs même si on fait parfois une distinction entre inférence et prédiction.

Un modèle prédictif permet d'obtenir des prédictions de la valeur de Y pour d'autres combinaisons de variables explicatives ou des données futures. Par exemple, on peut chercher à prédire la consommation énergétique d'une maison en fonction de la météo, du nombre d'habitants de la maison et de sa taille. La plupart des boîtes noires utilisées en apprentissage automatique tombent dans la catégorie des modèles prédictifs : ces modèles ne sont pas interprétables et ignorent parfois la structure inhérente aux données.

Par contraste, les modèles explicatifs sont souvent simples et interprétables ; les modèles de régressions sont fréquemment utilisés pour l'inférence. On se concentrera dans ce cours sur les modèles explicatifs. Par exemple, on peut chercher à déterminer

— Est-ce que les consommateurs sont prêts à dépenser davantage lorsqu'ils paient par crédit qu'en argent comptant ?
— Est-ce qu'il y a de la discrimination salariale envers les femmes professeurs d'un collège américain ?
— Études supérieures : est-ce que le prix en vaut la chandelle?.
— Quels sont les critères médicaux qui impactent les primes d'assurance maladies ?
— Qu'est-ce qui explique que les prix de l'essence soient plus élevés en Gaspésie qu'ailleurs au Québec ? Un rapport de surveillance des prix de l'essence en Gaspésie par la Régie de l'énergie se penche sur la question.
— Est-ce que les examens pratiques de conduite en Grande-Bretagne sont plus faciles dans les régions à faible densité ? Une analyse du journal britannique *The Guardian* laisse penser que c'est le cas.

— Est-ce le risque de transmission de la Covid augmente en fonction de la distanciation? Une (mauvaise) méta-analyse dit que oui (ou l'art de tirer des conclusions erronées à partir d'une étude bancale).

1.1 Tests d'hypothèse

Un test d'hypothèse statistique est une façon d'évaluer la preuve statistique provenant d'un échantillon afin de faire une décision quant à la population sous-jacente. Les étapes principales sont :

— définir les paramètres du modèle,
— formuler les hypothèses alternative et nulle,
— choisir et calculer la statistique de test,
— déterminer son comportement sous $\mathcal{H}_0$ (loi nulle),
— calculer la valeur-p,
— conclure dans le contexte du problème (rejeter ou ne pas rejeter $\mathcal{H}_0$).

Mon approche privilégiée pour présenter les tests d'hypothèse est de faire un parallèle avec un procès pour meurtre où vous êtes nommé juré.

— Le juge vous demande de choisir entre deux hypothèses mutuellement exclusives, coupable ou non-coupable, sur la base des preuves présentées.
— Votre postulat de départ repose sur la présomption d'innocence : vous condamnerez uniquement le suspect si la preuve est accablante. Cela permet d'éviter les erreurs judiciaires. L'hypothèse nulle $\mathcal{H}_0$ est donc *non-coupable*, et l'hypothèse alternative $\mathcal{H}_a$ est coupable. En cas de doute raisonnable, vous émettrez un verdict de non-culpabilité.
— La choix de la statistique de test représente la preuve. Plus la preuve est accablante, plus grande est la chance d'un verdict de culpabilité — le procureur a donc tout intérêt à bien choisir les faits présentés en cour. Le choix de la statistique devrait donc idéalement maximiser la preuve pour appuyer le postulat de culpabilité le mieux possible (ce choix reflète la **puissance** du test).
— En qualité de juré, vous analysez la preuve à partir de la jurisprudence et de l'avis d'expert pour vous assurer que les faits ne relèvent pas du hasard. Pour le test d'hypothèse, ce rôle est tenu par la loi sous $\mathcal{H}_0$: si la personne était innocente, est-ce que les preuves présentées tiendraient la route? des traces d'ADN auront davantage de poids que des ouï-dire (la pièce de théâtre *Douze hommes en colère* de Reginald Rose présente un bel exemple de procès où un des juré émet un doute raisonnable et convainc un à un les autres membres du jury de prononcer un verdict de non-culpabilité).
— Vous émettez un verdict, à savoir une décision binaire, où l'accusé est déclaré soit non-coupable, soit coupable. Si vous avez une valeur-p, disons P, pour votre statistique de test et que vous effectuez ce dernier à niveau α, la règle de décision revient à rejeter $\mathcal{H}_0$ si $P < \alpha$.

On s'attarde davantage sur ces définitions heuristiques et le vocabulaire employé pour parler de tests d'hypothèse. Le matériel de la section suivante a été préparé par Juliana Schulz.

1.1.1 Hypothèse

Dans les test statistique il y a toujours deux hypothèse : l'hypothèse nulle ($\mathcal{H}_0$) et l'hypothèse alternative ($\mathcal{H}_a$). Habituellement, l'hypothèse nulle est le « statu quo » et l'alternative est l'hypothèse que l'on cherche à démontrer. Un test d'hypothèse statistique nous permet de décider si nos données nous fournissent assez de preuves pour rejeter $\mathcal{H}_0$ en faveur de $\mathcal{H}_a$, selon un risque d'erreur spécifié. Généralement, les tests d'hypothèses sont exprimés en fonction de paramètres (de valeurs inconnues) du modèle sous-jacent, par ex. θ. Un test d'hypothèse bilatéral concernant un paramètre unidimensionnel θ s'exprimerait la forme suivante :

$$\mathcal{H}_0 : \theta = \theta_0 \qquad \text{versus} \qquad \mathcal{H}_a : \theta \neq \theta_0.$$

Ces hypothèses permettent de tester si θ est égal à une valeur numérique précise θ_0.

Par exemple, pour un test bilatéral concernant le paramètre d'un modèle de régression β_j associé à une variable explicative d'intérêt X_j, les hypothèses sont

$$\mathcal{H}_0 : \beta_j = \beta_j^0 \qquad \text{versus} \qquad \mathcal{H}_a : \beta_j \neq \beta_j^0,$$

où β_j^0 est une valeur précise qui est reliée à la question de recherche. Par exemple, si $\beta_j^0 = 0$ la question de recherche sous-jacente est : est-ce que la covariable X_j impacte la variable réponse d'intérêt Y une fois l'effet des autres variables pris en compte?

Remarque : il est possible d'imposer une direction dans les tests en considérant une hypothèse alternative de la forme $\mathcal{H}_a : \theta > \theta_0$ ou $\mathcal{H}_a : \theta < \theta_0$.

1.1.2 Statistique de test

Une statistique de test T est un fonctionnel des données qui résume l'information contenue dans les données pour θ. La forme de la statistique de test est choisie de façon à ce que son comportement sous $\mathcal{H}_0$, c'est-à-dire l'ensemble des valeurs que prend T si $\mathcal{H}_0$ est vraie et leur probabilité relative, soit connu. En effet, T est une variable aléatoire et sa valeur va changer selon l'échantillon. La **loi nulle** de la statistique de test nous permet de déterminer quelles valeurs de T sont plausibles si $\mathcal{H}_0$ est vraie. Plusieurs statistiques que l'on couvrira dans ce cours sont des **statistiques de Wald**, de la forme

$$T = \frac{\widehat{\theta} - \theta_0}{\mathrm{se}(\widehat{\theta})}$$

où $\widehat{\theta}$ est l'estimateur du paramètre θ, θ_0 la valeur numérique postulée (par ex., zéro) et $\mathrm{se}(\widehat{\theta})$ est l'estimateur de l'écart-type de $\widehat{\theta}$.

Par exemple, pour une hypothèse sur la moyenne d'une population de la forme

$$\mathcal{H}_0 : \mu = 0, \qquad \mathcal{H}_a : \mu \neq 0,$$

la statistique de test de Wald est

$$T = \frac{\overline{X} - 0}{S_n/\sqrt{n}}$$

où $\overline{X}$ est la moyenne de l'échantillon $X_1, \ldots, X_n$,

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{X_1 + \cdots + X_n}{n}$$

et l'erreur-type de la moyenne $\overline{X}$ est $S_n/\sqrt{n}$; l'écart-type S_n est un estimateur de σ, où

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Il convient de faire la différence entre procédures/formules et valeurs numériques. Un **estimateur** est une règle ou une formule utilisée pour calculer l'estimation d'un paramètre ou quantité d'intérêt selon des données observées. Par exemple, la moyenne d'échantillon $\overline{X}$ est un estimateur de la moyenne dans la population μ. Une fois qu'on a des données observées, on peut calculer un estimé de la moyenne empirique $\overline{x}$, c'est-à-dire, on obtient une valeur numérique. Autrement dit,

— un estimateur est une fonction de variables aléatoires et donc c'est aussi une variable aléatoire car sa valeur fluctue d'un échantillon à l'autre.
— l'estimé est la valeur numérique calculée sur un échantillon donné.

1.1.3 Loi nulle et valeur-*p*

La **valeur-*p*** nous permet de déterminer si la valeur observée de la statistique de test T est plausible sous $\mathcal{H}_0$. Plus précisément, la valeur-p est la probabilité, si $\mathcal{H}_0$ est vraie, que la statistique de test soit égale or plus extrême à ce qu'on observe. Supposons qu'on a un échantillon $X_1, \ldots, X_n$ et qu'on observe une valeur de la statistique de test de $T = t$. Pour un test d'hypothèse bilatéral $\mathcal{H}_0 : \theta = \theta_0$ vs. $\mathcal{H}_a : \theta \neq \theta_0$, la valeur-$p$ est $\mathsf{Pr}_0(|T| \geq |t|)$. Si la distribution de T est symétrique autour de zéro, la valeur-p vaut

$$p = 2 \times \mathsf{Pr}_0(T \geq |t|).$$

Prenons l'exemple d'un test d'hypothèse bilatéral pour la moyenne au population $\mathcal{H}_0 : \mu = 0$ contre $\mathcal{H}_a : \mu \neq 0$. Si l'échantillon provient d'une (population de) loi normale $\mathsf{No}(\mu, \sigma^2)$, on peut démontrer que, si $\mathcal{H}_0$ est vraie et donc $\mu = 0$), la statistique de test

$$T = \frac{\overline{X}}{S/\sqrt{n}}$$

suit une loi de Student-t avec $n-1$ degrés de liberté, dénotée St_{n-1}. À partir de cette loi nulle, on peut calculer la valeur-p (ou bien à partir d'une table ou d'un logiciel statistique). Puisque la distribution Student-t est symétrique autour de 0, on peut calculer la valeur-p comme $P = 2 \times \mathsf{Pr}(T > |t|)$, où $T \sim \mathsf{St}_{n-1}$.

1.1.4 Conclusion

La valeur-p nous permet de faire une décision quant aux hypothèses du test. Si $\mathcal{H}_0$ est vraie, la valeur-p suit une loi uniforme. Si la valeur-p est petite, ça veut dire que le fait d'observer une statistique de test égal ou encore plus extrême que $T = t$ est peu probable, et donc nous aurons tendance de croire que $\mathcal{H}_0$ n'est pas vraie. Il y a pourtant toujours un risque sous-jacent de commettre un erreur quand on prend une décision. En statistique, il y a deux types d'erreurs :

— erreur de type I : on rejette $\mathcal{H}_0$ alors que $\mathcal{H}_0$ est vraie
— erreur de type II : on ne rejette pas $\mathcal{H}_0$ alors que $\mathcal{H}_0$ est fausse

Ces deux erreurs ne sont pas égales : on cherche souvent à contrôler l'erreur de type I (une erreur judiciaire, condamner un innocent). Pour se prémunir face à ce risque, on fixe préalablement un niveau de tolérance. Plus notre seuil de tolérance α est grand, plus on rejette souvent l'hypothèse nulle même si cette dernière est vraie. La valeur de $\alpha \in (0,1)$ est la probabilité qu'on rejette $\mathcal{H}_0$ quand $\mathcal{H}_0$ est en fait vraie.

$$\alpha = \mathsf{Pr}_0\big(\text{ rejeter } \mathcal{H}_0\big).$$

Comme chercheur, on choisit ce niveau α ; habituellement 1%, 5% ou 10%. La probabilité de commettre une erreur de type I est α seulement si le modèle nul postulé pour $\mathcal{H}_0$ est correctement spécifié (sic) et correspond au modèle générateur des données.

Le choix du statu quo (typiquement $\mathcal{H}_0$) s'explique plus facilement avec un exemple médical. Si vous voulez prouver qu'un nouveau traitement est meilleur que l'actuel (ou l'absence de traitement), vous devez démontrer hors de tout doute raisonnable que ce dernier ne cause pas de torts aux patients et offre une nette amélioration (pensez à Didier Raoult et ses allégations non-étayées voulant que l'hydrochloroquine, un antipaludique, soit efficace face au virus de la Covid19).

Décision \ vrai modèle	$\mathcal{H}_0$	$\mathcal{H}_a$
ne pas rejeter $\mathcal{H}_0$	✓	erreur de type II
rejeter $\mathcal{H}_0$	erreur de type I	✓

Pour prendre une décision, on doit comparer la valeur-p P avec le niveau du test α :

— si $P < \alpha$ on rejette $\mathcal{H}_0$,
— si $P \geq \alpha$ on ne rejette pas $\mathcal{H}_0$.

Attention à ne pas confondre niveau du test (probabilité fixée au préalable par l'expérimentateur) et la valeur-p (qui dépend de l'échantillon). Si vous faites un test à un niveau 5% la probabilité de faire une erreur de type I est de 5% par définition, quelque soit la valeur de la valeur-p. La valeur-p s'interprète comme la probabilité d'obtenir une valeur de la statistique de test égale ou même plus grande que celle qu'on a observée dans l'échantillon, si $\mathcal{H}_0$ est vraie.

1.1.5 Puissance statistique

Le but du test d'hypothèse est de prouver (hors de tout doute raisonnable) qu'une différence ou un effet est significatif : : par exemple, si une nouvelle configuration d'un site web (hypothèse alternative) permet d'augmenter les ventes par rapport au statu quo. Notre capacité à détecter cette amélioration dépend de la puissance du test : plus cette dernière est élevée, plus grande est notre capacité à rejeter $\mathcal{H}_0$ quand ce dernier est faux. Quand on ne rejette pas $\mathcal{H}_0$ et que $\mathcal{H}_a$ est en fait vraie, on commet une erreur de type II : cette dernière survient avec probabilité $1-\gamma$. La **puissance statistique** d'un test est la probabilité que le test rejette $\mathcal{H}_0$ alors que $\mathcal{H}_0$ est fausse, soit

$$\gamma = \mathsf{Pr}_a(\text{rejeter } \mathcal{H}_0)$$

Selon le choix de l'alternative, il est plus ou moins facile de rejeter l'hypothèse nulle en faveur de l'alternative.

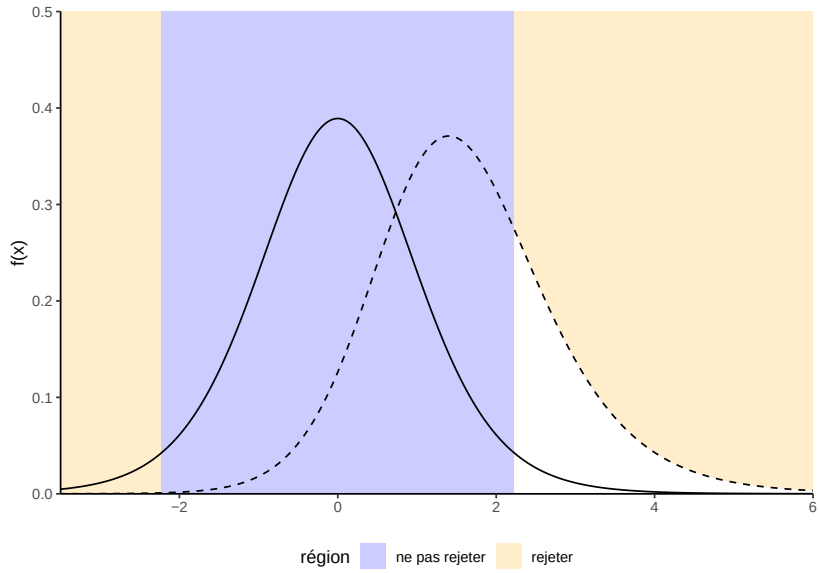

FIGURE 1 – Comparaison de la loi nulle (ligne pleine) et d'une alternative spécifique pour un test-t (ligne traitillée). La puissance correspond à l'aire sous la courbe de la densité de la loi alternative qui est dans la zone de rejet du test (en blanc).

On veut qu'un test ait une puissance élevée, c'est-à-dire, on veut que γ soit le plus près de 1 possible. Minimalement, la puissance du test devrait être α si on rejette l'hypothèse nulle une fraction α du temps quand cette dernière est vraie. La puissance dépend de plusieurs critères, à savoir :

— la taille de l'effet : plus la différence est grande entre la valeur du paramètre postulé θ_0 sous $\mathcal{H}_0$ et le comportement observé, plus il est facile de le détecter (voir Figure 3) ;

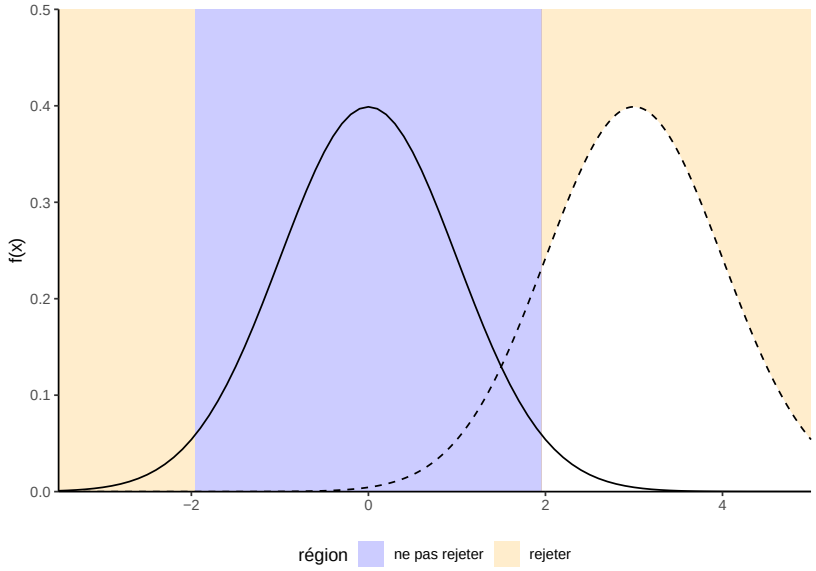

FIGURE 2 – Augmentation de la puissance suite à une augmentation de la différence de moyenne sous l'hypothèse alternative. La puissance est l'aire sous la courbe (blanc) de la loi alternative (ligne traitillée) ; cette dernière est plus décalée vers la droite par rapport à la loi nulle postulée (ligne pleine).

— la variabilité : moins les observations sont variables, plus il est facile de déterminer que la différence observée est significative (les grandes différences sont alors moins plausibles, comme l'illustre la Figure 2) ;
— la taille de l'échantillon : plus on a d'observations, plus notre capacité à détecter une différence significative augmente parce que l'erreur-type décroît avec la taille de l'échantillon à un rythme (ordinairement) de $n^{-1/2}$. La loi nulle devient aussi plus concentrée quand la taille de l'échantillon augmente.
— le choix de la statistique de test : par exemple, les statistiques basées sur les rangs n'utilisent pas les valeurs numériques qu'à travers le rang relatif. Ces tests sont donc moins puissants parce qu'ils n'utilisent pas toute l'information dans l'échantillon ; en contrepartie, ils sont souvent plus robustes en présence de valeurs aberrantes et si le modèle est mal spécifié. Les statistiques de test que nous choisirons sont souvent standards et parmi les plus puissantes qui soient, aussi on ne traitera pas de ce point davantage dans le cadre du cours.

Pour calculer la puissance d'un test, il faut choisir une alternative spécifique. Pour des exemples simples de statistiques, on peut obtenir une formule pour la puissance : par exemple, si on utilise un test-t pour un échantillon, la statistique $T = \sqrt{n}(\overline{X} - \mu_0)/S_n \sim \mathcal{T}_{n-1}$ et, si la vraie moyenne est $\Delta + \mu_0$, alors la loi alternative est Student-t, mais non-centrée avec paramètre de décalage Δ. Cette dérivation est l'exception plutôt que la règle et on détermine d'ordinaire la puissance à l'aide de

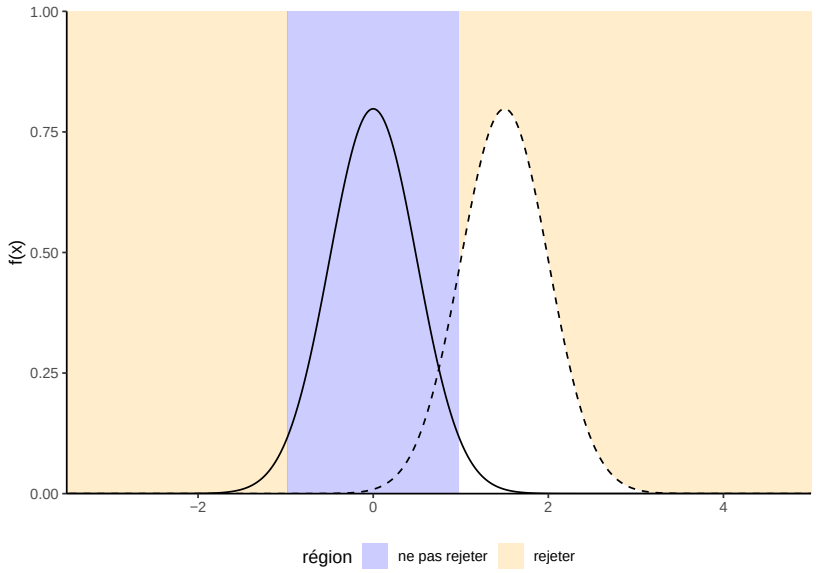

FIGURE 3 – Augmentation de la puissance suite à une augmentation de la taille de l'échantillon ou une diminution de l'écart-type de la population : la loi nulle (ligne pleine) est plus concentrée et la taille de la région de rejet diminue. La puissance est l'aire sous la courbe (blanc) de la loi alternative (ligne traitillée). Règle générale, la loi nulle change selon la taille de l'échantillon.

méthodes de Monte Carlo en simulant des observations d'une alternative donnée, en calculant la statistique de test sur le nouvel échantillon simulé et en calculant la valeur-p associée à notre hypothèse nulle de façon répétée. On calcule par la suite la proportion de tests qui mènent au rejet de l'hypothèse nulle à niveau α, ce qui correspond au pourcentage de valeurs-p inférieures à α.

1.1.6 Intervalle de confiance

Un **intervalle de confiance** est une manière alternative de rapporter les conclusions d'un test, en ce sens qu'on fournit une estimation ponctuelle de $\hat{\theta}$ avec une marge d'erreur. L'intervalle de confiance donne donc une indication de la variabilité de la procédure d'estimation. Un intervalle de confiance de Wald à $(1-\alpha)$ pour un paramètre θ est de la forme

$$\widehat{\theta} \pm \mathfrak{q}_{\alpha/2}\, \mathrm{se}(\widehat{\theta})$$

où $\mathfrak{q}_{\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi nulle de la statistique de Wald,

$$T = \frac{\widehat{\theta} - \theta}{\mathrm{se}(\widehat{\theta})},$$

et où θ représente la valeur du paramètre θ (supposé fixe, mais inconnu) de la population. Les bornes de l'intervalle de confiance sont aléatoires puisque $\widehat{\theta}$ et se$(\widehat{\theta})$ sont des variable aléatoires : leurs valeurs observées changent d'un échantillon à un autre.

Par exemple, pour un échantillon aléatoire $X_1, \ldots, X_n$ provenant d'une loi normale $\mathrm{No}(\mu, \sigma)$, l'intervalle de confiance à $(1-\alpha)$ pour la moyenne (dans la population) μ est

$$\overline{X} \pm t_{n-1,\alpha/2} \frac{S}{\sqrt{n}}$$

où $t_{n-1,\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi Student-t avec $n-1$ degrés de libertés.

Avant qu'on calcule l'intervalle de confiance, il y a une probabilité de $1-\alpha$ que θ soit contenu dans l'intervalle **aléatoire** symmétrique $(\widehat{\theta} - \mathrm{q}_{\alpha/2}\,\mathrm{se}(\widehat{\theta}), \widehat{\theta} + \mathrm{q}_{\alpha/2}\,\mathrm{se}(\widehat{\theta}))$, où $\widehat{\theta}$ dénote l'estimateur de θ. Une fois qu'on obtient un échantillon et qu'on calcule les bornes de l'intervalle de confiance, il n'y a plus de notion de probabilité : la vraie valeur du paramètre θ (inconnue) est soit contenue dans l'intervalle de confiance, soit pas. La seule interprétation de l'intervalle de confiance qui soit valable alors est la suivante : si on répète l'expérience plusieurs fois et qu'à chaque fois on calcule un intervalle de confiance à $1-\alpha$, alors une proportion de $(1-\alpha)$ de ces intervalles devraient contenir la vraie valeur de θ (de la même manière, si vous lancez une pièce de monnaie équilibrée, vous devriez obtenir grosso modo une fréquence de 50% de pile et 50% de face, mais chaque lancer donnera un ou l'autre de ces choix). Notre « confiance » est dans la procédure et non pas dans les valeurs numériques obtenues pour un échantillon donné.

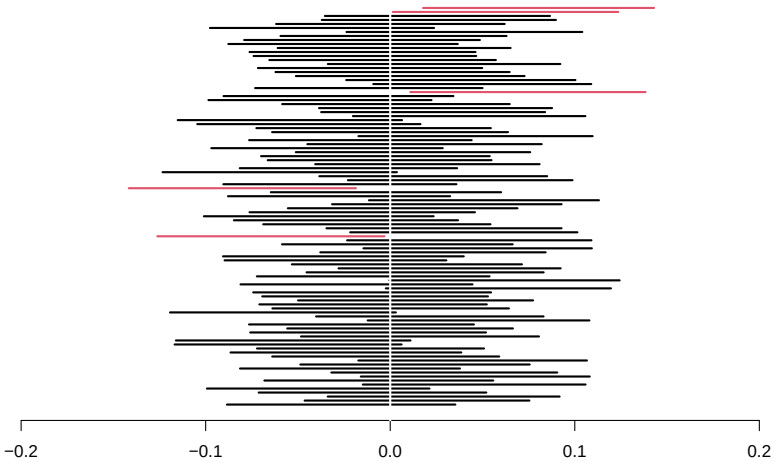

FIGURE 4 – Intervalles de confiance à 95% pour la moyenne d'une population normale $\mathrm{No}(0,1)$ pour 100 échantillons aléatoires. En moyenne, 5% de ces intervalles (en rouge) n'incluent pas la vraie valeur de la moyenne de zéro.

Si on s'intéresse seulement à la décision rejeter/ne pas rejeter $\mathcal{H}_0$, l'intervalle de confiance est équivalent à la valeur-p en ce sens qu'il mène à la même décision. L'intervalle de confiance donne en revanche l'ensemble des valeurs pour lesquelles la statistique de test ne fournit pas assez de preuves pour rejeter $\mathcal{H}_0$: pour un test à niveau α, on ne rejetterait aucune des valeurs contenues dans l'intervalle de confiance de niveau $1-\alpha$. Si la valeur-p est inférieure à α, la valeur postulée pour θ est donc hors de l'intervalle de confiance calculé. À l'inverse, la valeur-p ne donne la probabilité d'obtenir un résultat aussi extrême sous l'hypothèse nulle que pour une seule valeur numérique, mais permet de quantifier précisément à quel point le résultat est extrême.

Exemple 1.1 (Achat en ligne de milléniaux)**.** Supposons qu'une chercheuse veut faire une étude sur l'évolution des ventes en ligne au Canada. Elle postule que les membres de la génération Y fait plus d'achats en ligne que ceux des générations antérieures. Pour répondre à cette question, un sondage est envoyé à un échantillon aléatoire de $n = 500$ individus représentatif de la population avec 160 membres de la génération Y et 340 personnes plus âgées. La variable réponse est le montant d'achat effectués en ligne dans le mois dernier (en dollars).

Dans cet exemple, on s'intéresse à la différence entre le montant moyen des Y et celui des générations antérieures : la différence de moyenne observée dans l'échantillon est de 16.49 dollars et donc les milléniaux ont dépensé davantage. En revanche, notre échantillon est aléatoire et le montant d'achat en ligne varie d'un individu à l'autre (et d'un mois à l'autre) : ce n'est donc pas suffisant pour dire que la différence est significative.

La première étape de notre analyse consiste à définir les quantités d'intérêt et à formuler nos hypothèse en fonction de paramètres du modèle ; il convient également de définir ces derniers en fonction des variables en présence dans l'exemple. Ici, on considère un test pour la différence de moyenne dans les populations postulées μ_1 (pour la génération Y) et μ_2 (pour les générations antérieures) d'écart-type respectif σ_1 et σ_2. Comment déterminer quelle hypothèse on considère ? Comme statisticien, on se fait l'avocat du Diable : l'hypothèse d'intérêt du chercheur est l'hypothèse alternative et ici, $\mathcal{H}_a : \mu_1 > \mu_2$, où μ_1 représente la moyenne des achats mensuels des milléniaux. L'hypothèse nulle comprend toutes les autres valeurs pour la différence de moyenne, soit $\mathcal{H}_0 : \mu_1 \leq \mu_2$. Il suffit néanmoins de considérer le cas $\mu_1 = \mu_2$ (pourquoi ?)

La deuxième étape consiste à choisir une statistique de test. S'il n'y a aucune différence de moyenne entre les groupes, alors $\overline{X}_1 - \overline{X}_2$ a moyenne zéro et la différence de moyenne a une variance de $\sigma_1^2/n_1 + \sigma_2^2/n_2$. Ici, on considère la statistique de **?** pour une différence de moyenne entre deux échantillons :

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^{1/2}},$$

où $\overline{X}_i$ est la moyenne empirique dans l'échantillon i ($i = 1, 2$) et S_i^2 est la variance empirique et n_i la taille de l'échantillon du groupe i. La statistique est utilisée pour calculer la différence de

TABLE 2 – Aperçu des données `renfe`.

prix	type	classe	tarif	dest	duree	jour
143.4	AVE	Preferente	Promo	Barcelone-Madrid	190	6
181.5	AVE	Preferente	Flexible	Barcelone-Madrid	190	2
86.8	AVE	Preferente	Promo	Barcelone-Madrid	165	7
86.8	AVE	Preferente	Promo	Barcelone-Madrid	190	7
69.0	AVE-TGV	Preferente	Promo	Barcelone-Madrid	175	4

moyennes de deux échantillons de variance potentiellement différente. La valeur de la statistique dans l'échantillon est $T = 2.76$, mais on obtiendrait une valeur différente avec un autre échantillon. Il convient donc de déterminer si cette valeur est compatible avec notre hypothèse nulle en la comparant à la loi nulle sous $\mathcal{H}_0$ de T. On effectuera le test à niveau $\alpha = 0.05$.

La troisième étape est l'obtention d'un étalon de mesure pour déterminer si notre résultat est extrême ou inattendu. Vous remarquerez que la statistique de Welch a moyenne zéro et variance un sous l'hypothèse nulle que $\mu_1 = \mu_2$: standardiser une statistique permet d'obtenir un objet dont on connaît le comportement pour de grands échantillons et obtenir une quantité sans unité de mesure. La dérivation de la loi nulle est hors objectifs du cours, aussi cette dernière vous sera donnée dans tous les cas qu'on considère. Asymptotiquement, T suit une loi normale $\mathsf{No}(0,1)$, mais il existe une meilleure approximation pour n petit; on compare le comportement de T à l'aide d'une loi de Student (à l'aide de l'approximation de **?**).

La dernière étape consiste à obtenir une valeur-p, soit la probabilité d'observer un résultat aussi extrême sous $\mathcal{H}_0$: l'avantage de la valeur-p est que cette valeur est une probabilité (dans $[0,1]$) et qu'elle suit une loi uniforme sous $\mathcal{H}_0$. Puisque nous avons une hypothèse alternative unilatérale, on regarde la probabilité sous $\mathcal{H}_0$ que $\mathsf{Pr}(T > t)$. La valeur-p vaut 0.0031 et donc, à niveau 5%, on rejette l'hypothèse nulle pour conclure que la génération Y dépense davantage en ligne que les générations antérieures.

Exemple 1.2 (Prix de billets de trains à grande vitesse espagnols)**.** La compagnie nationale de chemin de fer Renfe gère les trains régionaux et les trains à haute vitesse dans toute l'Espagne. Les prix des billets vendus par Renfe sont aggrégés par une compagnie. On s'intéresse ici à une seule ligne, Madrid–Barcelone. Notre question scientifique est la suivante : est-ce que le prix des billets pour un aller (une direction) est plus chère pour un retour? Pour ce faire, on considère un échantillon de 10000 billets entre les deux plus grandes villes espagnoles. On s'intéresse au billets de TGV vendus (AVE) au tarif Promotionnel. Notre statistique de test sera simplement la différence de moyenne entre les deux échantillons : la différence entre le prix en euros d'un train Madrid–Barcelone (μ_1) et le prix d'un billet Barcelone–Madrid (μ_2) est $\mu_1 - \mu_2$ et notre hypothèse nulle est qu'il n'y a aucune différence de prix, soit $\mathcal{H}_0 : \mu_1 - \mu_2 = 0$.

On utilise de nouveau le test de Welch pour deux échantillons en filtrant les données pour ne conserver que les billets au tarif Promo : la moyenne des billets Barcelone-Madrid est 82.11 euros, ceux pour Madrid-Barcelone 82.56 et la valeur de la statistique de Welch est -1.33. Si on utilise l'approximation normale, on obtient une valeur-p de 0.18.

Plutôt que d'utiliser la loi asymptotique (qui est valide pour de grands échantillons à cause du théorème central limite), on peut considérer une approximation sous une hypothèse moins restrictive en supposant que les données sont échangeables. Sous l'hypothèse nulle, il n'y aucune différence entre les deux destinations et les étiquettes pour la destination (une variable catégorielle binaire) sont arbitraires. On pourrait considérer les mêmes données, mais avec une permutation des variables explicatives : c'est ce qu'on appelle un test de permutation. On va recréer deux groupes de taille identique à notre échantillon original, mais en changeant les observations. On recalcule la statistique de test sur ces nouvelle données (si on a une poignée d'observations, il est possible de lister toutes les permutations possibles ; typiquement, il suffit de considérer un grand nombre de telles permutations, disons 9999). Pour chaque nouveau jeu de données, on calculera la statistique de test et on calculera le rang de notre statistique par rapport à cette référence. Si la valeur de notre statistique observée sur l'échantillon original est extrême en comparaison, c'est autant de preuves contre l'hypothèse nulle.

La valeur-p du test de permutation, 0.186, est la proportion de statistiques plus extrêmes que celle observée. Cette valeur-p est quasi-identique à celle de l'approximation de Satterthwaite, à savoir 0.182 (la loi Student-t est numériquement équivalente à une loi standard normale avec autant de degrés de liberté), tel que représenté dans la Figure 5. Malgré que notre échantillon soit très grand, avec $n = 8059$ observations, la différence n'est pas jugée significative. Avec un échantillon de deux millions de billets, on pourrait estimer précisément la moyenne (au centime près) : la différence de prix entre les deux destinations et cette dernière deviendrait statistiquement significative. Elle n'est pas en revanche pertinente (une différence de 0.28 euros sur un prix moyen de 82.56 euros est quantité négligeable).

1.2 Analyse exploratoire de données

L'analyse exploratoire, comme son nom l'indique, est une étape préliminaire à la modélisation servant à l'acquisition d'une meilleure compréhension des données. Une connaissance rudimentaire des graphiques est nécessaire et on s'attardera aux rudiments de la visualisation graphique. Plusieurs ouvrages abordent ces notions (en anglais) .

— Chapitre 3 de ***R*** *for Data Science* par Garrett Grolemund et Hadley Wickham
— Section 1.6 du livre *Introductory Statistics with Randomization and Simulation* d'OpenIntro
— *Fundamentals of Data Visualization* par Claus O. Wilke
— Chapitre 1 de *Data Visualization: A practical introduction* par Kieran Healy

Si l'analyse exploratoire est souvent négligée dans les cours de statistique (parce qu'elle n'a pas de fondement mathématique), elle n'en est pas moins importante car elle nous sert à interpréter les

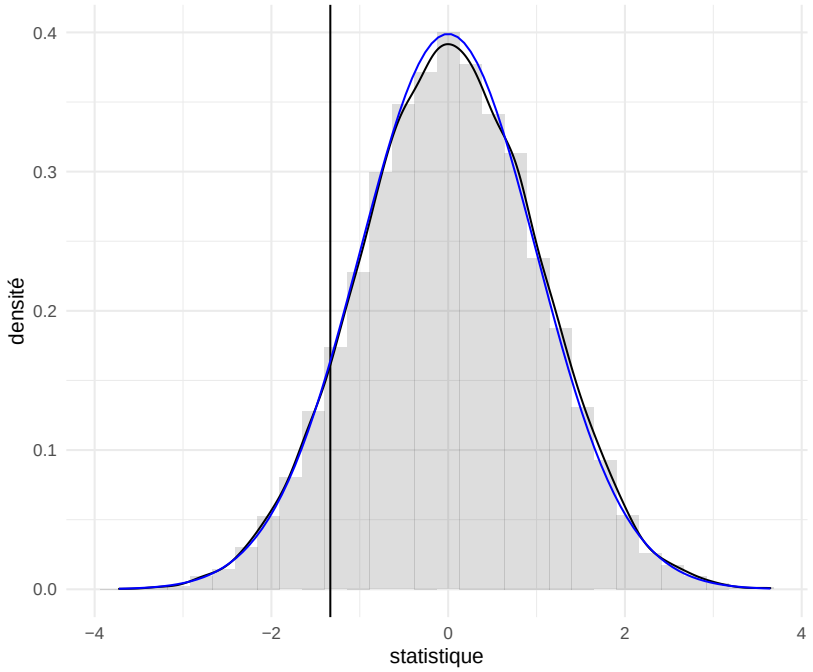

FIGURE 5 – Approximation par permutation de la loi nulle de la statistique de test de Welch (histogramme et trait noir) et loi asymptotique normale standard (trait bleu) pour le prix de billets de trains AVE au tarif promotionnel entre Madrid et Barcelone. La valeur de la statistique de test de l'échantillon original est représentée par un trait vertical.

données dans le contexte du problème et à nous assurer que notre analyse ou notre traitement de ces dernières est cohérent. Le sujet est difficile à cerner, puisque c'est davantage un art qu'une approche rigoureuse ; Grolemund et Wickham parlent même « d'état d'esprit ». Le but de l'analyse exploratoire graphique est d'extraire des informations utiles, le plus souvent par le biais d'une série de questions qui sont raffinées au fur et à mesure que progresse l'analyse. On s'intéresse particulièrement aux relations et interactions entre différentes variables et la distribution empirique de chaque variable. Les étapes majeures sont :

1. Formuler des questions sur les données
2. Chercher des réponses à ces questions à l'aide de statistiques descriptives, de tableaux de fréquence ou de contingence et de graphiques.
3. Raffiner nos questions, et utiliser les trouvailles pour peaufiner notre analyse

Dans un rapport, un résumé des caractéristiques les plus importantes devrait être inclut pour que le lecteur ou la lectrice puisse valider son interprétation des données.

1.2.1 Soignez votre travail

Si vous incluez un graphique (ou un tableau), il est important d'ajouter une légende qui décrit le graphique et le résume, les noms de variables (avec les unités) sur les axes, mais aussi de soigner le rendu et le formatage pour obtenir un produit fini propre, lisible et cohérent : en particulier, votre description devrait coïncider avec le rendu. Votre graphique raconte une histoire, aussi prenez-soin que cette dernière soit nécessaire et attrayante.

1.2.2 Types de variables

— Une **variable** représente une caractéristique de la population d'intérêt, par exemple le sexe d'un individu, le prix d'un article, etc.
— une **observation**, parfois appelée donnée, est un ensemble de mesures collectées sous des conditions identiques, par exemple pour un individu ou à un instant donné.

Le choix de modèle statistique ou de test dépend souvent du type de variables collectées. Les variables peuvent être de plusieurs types : quantitatives (discrètes ou continues) si elles prennent des valeurs numériques, qualitatives (binaires, nominales ou ordinales) si elles sont décrites par un adjectif ; je préfère le terme catégorielle, plus évocateur.

Les modèles de régression servent à expliquer des variables quantitatives en fonction d'autres caractéristiques.

— une variable discrète prend un nombre dénombrable de valeurs ; ce sont souvent des variables de dénombrement ou des variables dichotomiques.
— une variable continue peut prendre (en théorie) une infinité de valeurs, même si les valeurs mesurées sont arrondies ou mesurées avec une précision limitée (temps, taille, masse, vitesse, salaire). Dans bien des cas, nous pouvons considérer comme continues des variables discrètes si elles prennent un assez grand nombre de valeurs.

Les variables catégorielles représentent un ensemble fini de possibilités. On les regroupe en deux types, pour lesquels on ne fera pas de distinction : nominales s'il n'y a pas d'ordre entre les modalités (sexe, couleur, pays d'origine) ou ordinale (échelle de Likert, tranche salariale). La codification des modalités des variables catégorielle est arbitraire ; en revanche, on préservera l'ordre lorsqu'on représentera graphiquement les variables ordinales. Lors de l'estimation, chaque variable catégorielle doit est transformée en un ensemble d'indicateurs binaires : il est donc essentiel de déclarer ces dernières dans votre logiciel statistique, surtout si elles sont parfois encodées dans la base de données à l'aide de valeurs entières.

1.2.3 Graphiques

Le principal type de graphique pour représenter la distribution d'une variable catégorielle est le diagramme en bâtons, dans lequel la fréquence de chaque catégorie est présentée sur l'axe des ordonnées (y) en fonction de la modalité, sur l'axe des abscisses (x), et ordonnées pour des

variables ordinales. Cette représentation est en tout point supérieur au diagramme en camembert, une engeance répandu qui devrait être honnie (notamment parce que l'humain juge mal les différences d'aires, qu'une simple rotation change la perception du graphique et qu'il est difficile de mesurer les proportions) — ce n'est pas de la tarte!

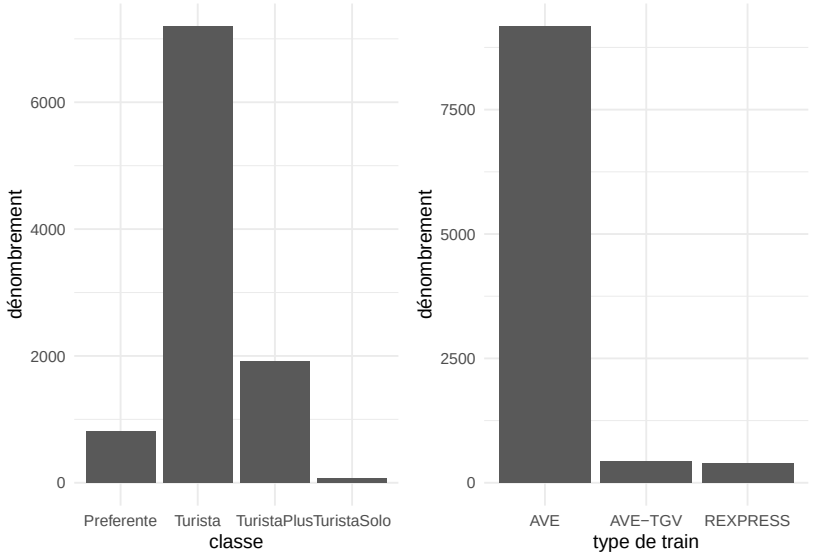

FIGURE 6 – Diagramme en bâtons pour la classe des billets de trains du jeu de données Renfe.

Puisque les variables continues peuvent prendre autant de valeurs distinctes qu'il y a d'observations, on ne peut simplement compter le nombre d'occurrence par valeur unique. On regroupera plutôt dans un certain nombre d'intervalle, en discrétisant l'ensemble des valeurs en classes pour obtenir un histogramme. Le nombre de classes dépendra du nombre d'observations si on veut que l'estimation ne soit pas impactée par le faible nombre d'observations par classe : règle générale, le nombre de classes ne devrait pas dépasser $\sqrt{n}$, où n est le nombre d'observations de l'échantillon. On obtiendra la fréquence de chaque classe, mais si on normalise l'histogramme (de façon à ce que l'aire sous les bandes verticales égale un), on obtient une approximation discrète de la fonction de densité. Faire varier le nombre de classes permet parfois de faire apparaître des caractéristiques de la variable (notamment la multimodalité, l'asymmétrie et les arrondis).

Puisque qu'on groupe les observations en classe pour tracer l'histogramme, il est difficile de voir l'étendue des valeurs que prenne la variable : on peut rajouter des traits sous l'histogramme pour représenter les valeurs uniques prises par la variable, tandis que la hauteur de l'histogramme nous renseigne sur leur fréquence relative.

Une boîte à moustaches(*boxplot*) représente graphiquement cinq statistiques descriptives.

— La boîte donne les 1e, 2e et 3e quartiles q_1, q_2, q_3. Il y a donc 50% des observations sont

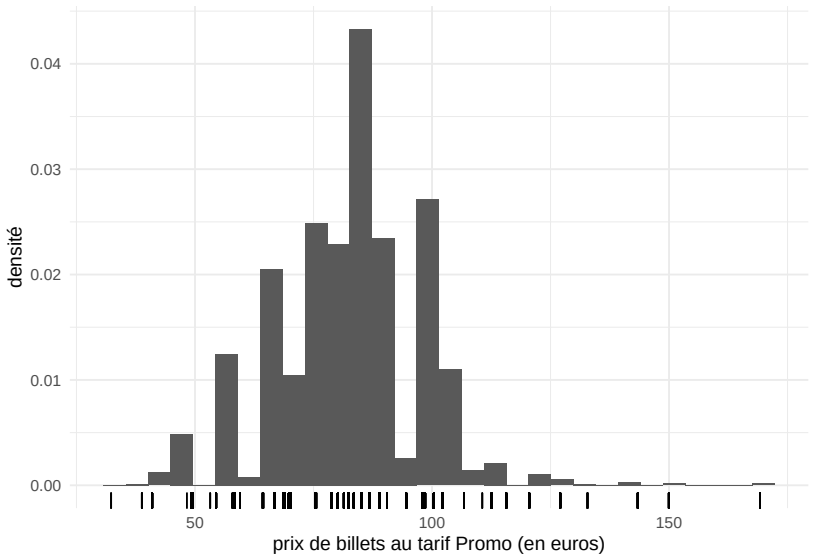

FIGURE 7 – Histogramme du prix des billets au tarif Promo de trains du jeu de données Renfe

au-dessus/en-dessous de la médiane q_2 qui sépare en deux la boîte.

— La longueur des moustaches est moins de 1.5 fois l'écart interquartile $q_3 - q_1$ (tracée entre 3e quartile et le dernier point plus petit que $q_3 + 1.5(q_3 - q_1)$, etc.)

— Les observations au-delà des moustaches sont encerclées. Notez que plus le nombre d'observations est élevé, plus le nombres de valeurs aberrantes augmente. C'est un défaut de la boîte à moustache, qui a été conçue pour des jeux de données qui passeraient pour petits selon les standards actuels.

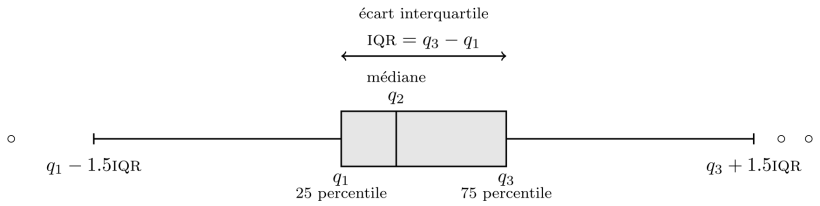

FIGURE 8 – Boîte à moustache.

On peut représenter la distribution d'une variable réponse continue en fonction d'une variable catégorielle en traçant une boîte à moustaches pour chaque catégorie et en les disposant côte-à-côte. Une troisième variable catégorielle peut être ajoutée par le biais de couleurs, comme dans la Figure 9.

Si on veut représenter la covariabilité de deux variables continues, on utilise un nuage de points où

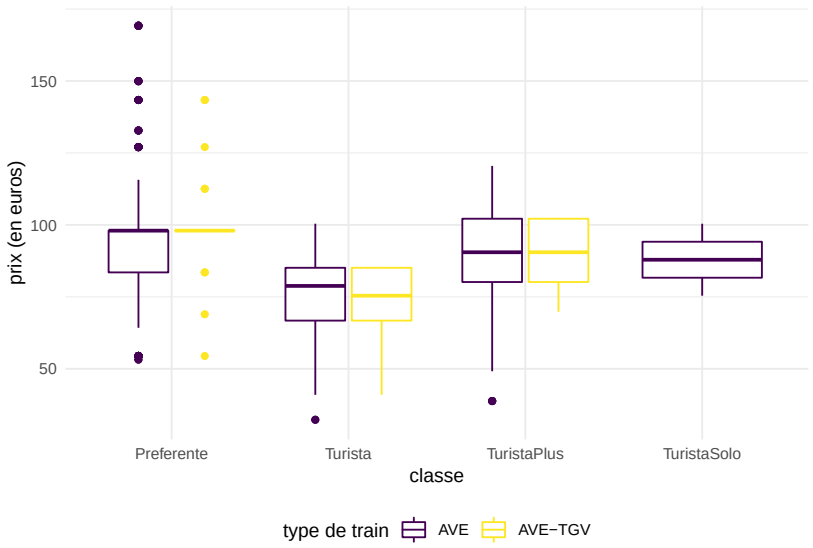

FIGURE 9 – Boîte à moustaches du prix des billets au tarif Promo en fonction de la classe pour le jeu de données Renfe.

chaque variable est représentée sur un axe et chaque observation donne la coordonnée des points. Si la représentation graphique est dominée par quelques valeurs très grandes, une transformation des données peut être utile : vous verrez souvent des données positives à l'échelle logarithmique.

Plutôt que de décrire plus en détail le processus de l'analyse exploratoire, on présente un exemple qui illustre le cheminement habitue sur les données de trains de la Renfe introduites précédemment.

Exemple 1.3 (Analyse exploratoire des trains Renfe)**.** La première étape consisterait à lire la description de la base de données. Le jeu de données
$texttttrenfe$ contient les variables suivantes

- `prix` : prix du billet (en euros) ;
- `dest` : indicateur binaire du trajet, soit de Barcelone vers Madrid (`0`) ou de Madrid vers Barcelone (`1`) ;
- `tarif` : variable catégorielle indiquant le tarif du billet, un parmi `AdultoIda`, `Promo` et `Flexible` ;
- `classe` : classe du billet, soit `Preferente`, `Turista`, `TuristaPlus` ou `TuristaSolo` ;
- `type` : variable catégorielle indiquant le type de train, soit Alta Velocidad Española (`AVE`), soit Alta Velocidad Española conjointement avec TGV (un partenariat entre la SNCF et Renfe pour les trains à destination ou en provenance de Toulouse) `AVE-TGV`, soit les trains régionaux `REXPRESS` ; seuls les trains étiquetés `AVE` ou `AVE-TGV` sont des trains à grande vitesse.
- `duree` : longueur annoncée du trajet (en minutes) ;

TABLE 3 – Nombre d'observations du jeu de données `renfe` par classe.

classe	n
Preferente	809
Turista	7197
TuristaPlus	1916
TuristaSolo	78

TABLE 4 – Nombre d'observations du jeu de données `renfe` par type de billets.

type	n
AVE	9174
AVE-TGV	429
REXPRESS	397

— `jour` entier indiquant le jour de la semaine du départ allant de dimanche (`1`) à samedi (`7`).

Il n'y a pas de valeurs manquantes et un aperçu des données (`head(renfe)`) montre qu'elles sont en format long, ce qui veut dire que chaque ligne contient une seule valeur pour la variable réponse, ici le prix d'un billet de train. On entame l'analyse exploratoire avec des questions plutôt vagues, par exemple

1. Quels sont les facteurs déterminant le prix et le temps de parcours ?
2. Est-ce que le temps de parcours est le même peut importe le type de train ?
3. Quelles sont les caractéristiques distinctives des types de train ?
4. Quelles sont les principales différences entre les tarifs ?

À l'exception de `prix` et de `duree`, toutes les variables explicatives sont catégorielles. On porte une attention particulière à `jour` pour éviter les mauvaises surprises ultérieures.

En analysant le nombre de trains dans les catégories, on remarque qu'il y a autant de billets de type `REXPRESS` que le nombre de billets au tarif `AdultoIda`. On peut faire le décompte par catégorie avec un tableau de contingence, qui compte le nombre respectif dans chaque sous-catégorie. Dans

TABLE 5 – Nombre d'observations du jeu de données `renfe` par tarif.

tarif	n
AdultoIda	397
Flexible	1544
Promo	8059

TABLE 6 – Tableau croisé du décompte par tarif et type de billets du jeu de données `renfe`.

tarif	type	n
AdultoIda	REXPRESS	397
Flexible	AVE	1446
Flexible	AVE-TGV	98
Promo	AVE	7728
Promo	AVE-TGV	331

TABLE 7 – Statistiques descriptives de la durée de trajet et du prix du billet par type de billet et destination pour le jeu de données `renfe`.

type	dest	durée moyenne	écart-type	prix moyen
REXPRESS	Barcelone-Madrid	544	0	43.2
REXPRESS	Madrid-Barcelone	562	0	43.2

la base de données Renfe, tous les billets pour les RegioExpress sont vendus au tarif `AdultoIda` en classe `Turista`. Le nombre de billets est minime, à peine 397 sur 10000. Cela suggère une nouvelle question : pourquoi ces trains sont-ils si peu populaires ?

On remarque également que seulement 17 temps de parcours sont affichés sur les billets (`renfe %>% distinct(duree)` ou `unique(renfe$duree)`). On peut donc penser que la durée affichée sur le billet (en minutes) est le temps de trajet annoncé. La majeure partie (15 sur 17) des temps de parcours sont sous la barre des 3h15, hormis deux qui dépassent les 9h ! Selon Google Maps, les deux villes sont distantes de 615km par la route, 500km à vol d'oiseau. Cela implique que, vraisemblablement, certains trains dépassent les 200km/h, tandis que d'autres vont plutôt à 70km/h. Quels sont ces trains plus lent ? La variable `type` codifie probablement ce fait, et permet de voir que ce sont les trains RegioExpress qui sont dans cette catégorie.

Aller de Madrid à Barcelone à l'aide d'un train régulier prend 18 minutes de plus. Avec plus de 9h de trajet, pas étonnant donc que ces billets soient peu courus. Encore plus frappant, on note que le prix des billets est fixe : 43.25 euros peu importe que le trajet soit aller ou retour. C'est probablement la trouvaille la plus importante jusqu'à maintenant, car les billets de train de type RegioExpress ne forment pas un échantillon : il n'y a aucune variabilité ! On aurait pu également découvrir cette anomalie en traçant une boîte à moustaches du prix en fonction du type de train.

On pourrait soupçonner que les trains étiquetés `AVE` soient plus rapides, sachant que c'est l'acronyme de *Alta Velocidad Española*, littéralement haute vitesse espagnole. Qu'en est-il des distinctions entre les deux types de trains étiquetés AVE ? Selon le site de la SNCF, les trains AVE-TGV sont des partenariats entre la Renfe et la SNCF et effectuent des liaisons entre la France et l'Espagne.

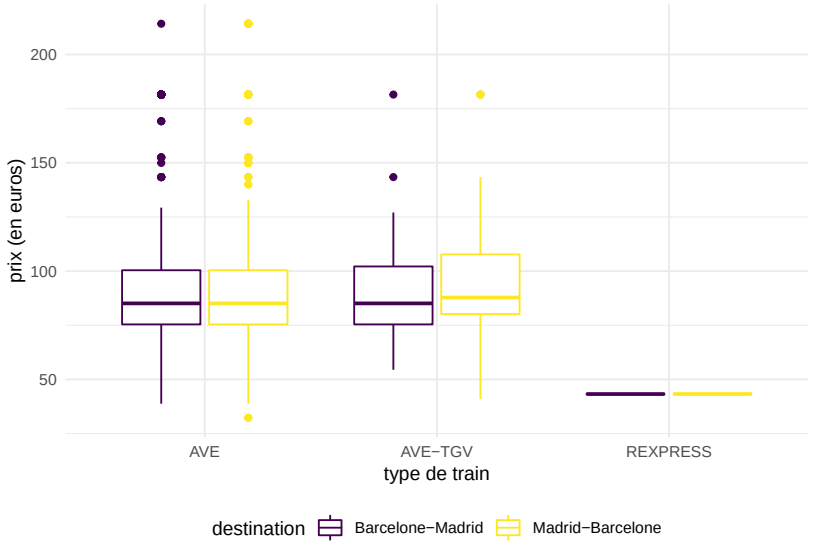

FIGURE 10 – Boîte à moustaches du prix de billets de train de Renfe en fonction de la destination et du type de train.

TABLE 8 – Statistiques descriptives de la durée de trajet et du prix du billet par type de billet et destination pour les trains grande vitesse du jeu de données `renfe`.

type	dest	durée moyenne	écart-type	prix moyen
AVE	Barcelone-Madrid	171	19.8	87.4
AVE	Madrid-Barcelone	170	20.8	88.2
AVE-TGV	Barcelone-Madrid	175	16.8	87.0
AVE-TGV	Madrid-Barcelone	179	20.2	90.6

Les prix sont beaucoup plus élevés, en moyenne plus de deux fois plus que les trains régionaux. Les écarts de prix importants (l'écart type est de 20 euros) indique qu'il y a peut-être d'autres sources d'hétérogénéité, mais on pourrait soupçonner que la Renfe pratique la tarification dynamique. Il y un seul temps de parcours prévu pour les trains AVE-TGV. On ne note pas de différence de prix notable selon la direction ou le type de train grande vitesse, mais peut-être que les tarifs ou la classe disponibles diffèrent selon que le train ou non est en partenariat avec la compagnie française.

On n'a pas encore considéré le tarif et la classe des billets, hormis pour les trains RegioExpress. On voit dans la Figure 12 une forte différente dans l'hétérogénéité des prix selon le tarif; le tarif Promo prend plusieurs valeurs distinctes, tandis que les tarifs AdultoIda et Flexible semblent ne prendre que quelques valeurs. La première classe (`Preferente`) est plus chère et il y a moins d'observations dans ce groupe. La classe Turista est la classe la moins dispendieuse et la plus

TABLE 9 – Décompte des billets au tarif Flexible par prix et classe pour le jeu de données `renfe`.

prix	classe	n
108	Turista	1050
108	TuristaSolo	67
127	Turista	285
127	TuristaSolo	9
129	TuristaPlus	31
140	Preferente	2
152	TuristaPlus	10
182	Preferente	78
214	Preferente	12

populaire. `TuristaPlus` offre plus de confort, tandis que `TuristaSolo` permet d'obtenir un siège individuel.

Côté tarif, Promo et PromoPlus permette d'obtenir des rabais pouvant aller jusqu'à respectivement 70% et 65%. Les annulations et changements ne sont pas possibles avec Promo, mais disponibles avec PromoPlus moyennant une pénalité équivalent à 30-20% du prix du billet. Le tarif Flexible est disponible au même prix que les billets réguliers, avec des bénéfices additionnels.

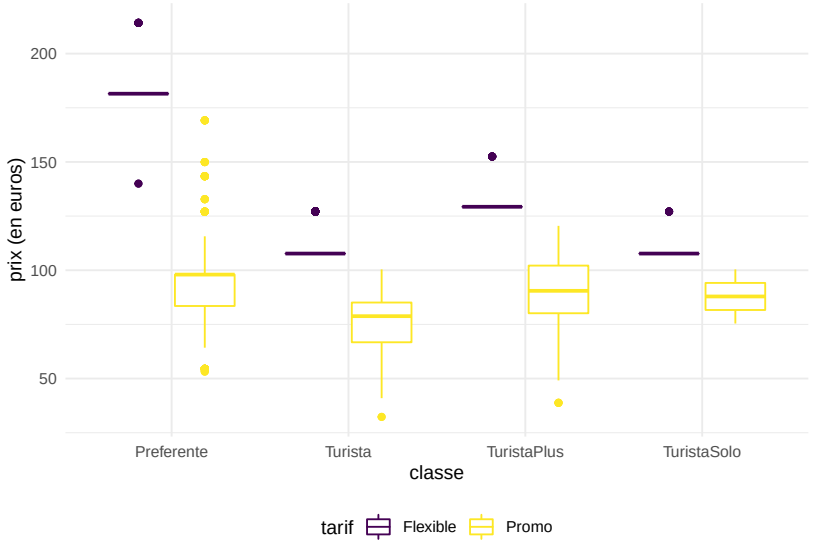

FIGURE 11 – Boîte à moustaches du prix en fonction du tarif et de la classe de billets de trains à haute vitesse de la Renfe.

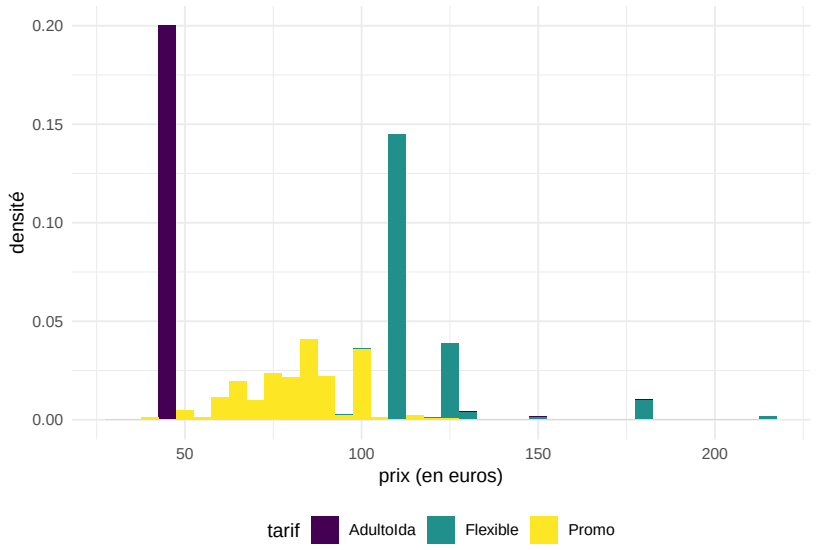

FIGURE 12 – Histogrammes du prix en fonction du tarif de billets de trains de la Renfe.

On note que la répartition des prix pour les billets de classe Flexible est inhabituelle : notre boîte à moustachesest écrasée et l'écart interquartile semble nul, même si quelques valeurs inexpliquées sont aussi présentes. L'écrasante majorité des billets Flexibles sont en classe Turista, donc ça pourrait être dû à un (trop) faible nombre de billets dans chaque catégorie. On peut rejeter cette hypothèse en calculant le nombre de trains au tarif Flexible pour les différents types de billets. Ni la durée, ni le type de train, ni la destination n'expliquent pas pourquoi le prix de certains billets Flexibles est plus faible ou élevés. Le prix des billets Promo est plus faible, et les billets au tarif Preferente (la première classe) sont plus élevés.

On peut résumer notre brève analyse exploratoire :

— plus de 91% des trains sont des trains à grande vitesse AVE.
— le temps de trajet dépend du type de train : les trains à grande vitesse mettent 3h20 au maximum pour relier Madrid et Barcelone.
— les temps de trajets sont ceux annoncés (variable discrète avec 17 valeurs uniques, dont 13 pour les trains AVE)
— le prix de trains RegioExpress est fixe (43.25€) ; tous ces billets sont dans la classe Turista et au tarif Adulto Ida. 57% de ces trains vont de Barcelone à Madrid. La durée du trajet pour les RegioExpress est de 9h22 de Barcelona à Madrid, 18 minutes de plus que dans l'autre direction.
— les billets en classe`Preferente` sont plus chers et moins fréquents. La classe `Turista` est la classe la moins dispendieuse et la plus populaire. `TuristaPlus` offre plus de confort, tandis que `TuristaSolo` permet d'obtenir un siège individuel.

— selon le site web de la Renfe, les billets au tarif `Flexible` « viennent avec des offres additionnelles qui permettent au passagers d'échanger leurs billets ou annuler s'ils manquent leurs trains. » ; en contrepartie, ces billets sont plus chers et leur tarif est fixe sauf une poignée de billets dont le prix reste inexpliqué.
— la distribution des prix des billets de TGV au tarif `Promo` est plus ou moins symmétrique, tandis que les billets au tarif `Flexible` apparaissent tronqués à gauche (le prix minimum pour ces billets est 107.7€ dans l'échantillon).
— la Renfe pratique la tarification dynamique pour les billets au tarif promotionnel `Promo` : ces derniers peuvent être jusqu'à 70% moins chers que les billets à prix régulier lorsqu'achetés via l'agence officielle ou le site de Renfe. Ces billets ne peuvent être ni remboursés, ni échangés.
— il n'y a pas d'indication à effet de quoi les prix varient selon la direction du trajet.

2 Régression linéaire

On entend par régression linéaire un modèle pour l'espérance conditionnelle d'une variable réponse Y (ou régressande) en fonction de p variables explicatives (appelées parfois régresseurs ou covariables) à l'aide d'une équation de la forme

$$\mathsf{E}(Y \mid \mathbf{X}) = \beta_0 + \beta_1 \mathrm{X}_1 + \cdots + \beta_p \mathrm{X}_p.$$

Le fait que la moyenne est conditionnelle aux valeurs de $\mathbf{X}$ implique simplement que l'on considère les régresseurs comme constant, ou connus à l'avance.

En pratique, tout modèle est une approximation de la réalité, aussi on ajoute un terme d'erreur qui sert à tenir compte du fait qu'aucune relation linéaire exacte ne lie $\mathbf{X}$ et Y, ou que les mesures de Y contiennent des erreurs. Ce terme d'erreur aléatoire ε servira de base à l'inférence car il permettra de quantifier l'adéquation entre notre modèle et les données.

On peut réécrire le modèle linéaire en terme de l'erreur pour un échantillon aléatoire de taille n : dénotons par Y_i la valeur de Y pour le sujet i, et X_{ij} la valeur de la je variable explicative du sujet i. Le modèle de régression linéaire est

$$Y_i = \beta_0 + \beta_1 \mathrm{X}_{i1} + \ldots + \beta_p \mathrm{X}_{ip} + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{1}$$

où ε_i est le terme d'erreur additive. Si aucune hypothèse sur la loi aléatoire de l'erreur n'est spécifiée, on fixe néanmoins l'espérance du terme d'erreur à zéro car on postule qu'il n'y a pas d'erreur systématique, c'est-à-dire que $\mathsf{E}(\varepsilon_i \mid \boldsymbol{X}_i) = 0$ ($i = 1, \ldots, n$).

La flexibilité du modèle linéaire vient de sa formulation : on spécifie l'espérance conditionnelle d'une variable continue comme **combinaison linéaire de variables explicatives**, dont le choix est arbitraire. Il est important de remarquer que ce modèle est linéaire dans les coefficients $\boldsymbol{\beta} \in \mathbb{R}_{p+1}$, pas dans les variables explicatives ! les covariables sont quelconques et peuvent être des fonctions (non)-linéaires d'autres variables explicatives, par exemple $\mathrm{X} = \log(\texttt{annees})$, $\mathrm{X} = \texttt{puissance}^2$ ou

$X = I_{\texttt{homme}} \cdot I_{\texttt{titulaire}}$. C'est ce qui fait la flexibilité du modèle linéaire : ce dernier est principalement employé aux fins suivantes :

1. Comprendre comment et dans quelle mesure les variables explicatives $\mathbf{X}$ influencent la moyenne de la réponse Y (description).
2. Quantifier l'influence des variables explicatives $\mathbf{X}$ sur la régressande Y et tester leur significativité.
3. Prédire les valeurs de Y pour de nouveaux ensembles de covariables $\mathbf{X}$.

2.1 Introduction

Le modèle linéaire est sans conteste le modèle statistique le plus couramment employé. Le terme « modèle linéaire » est trompeur : une grande panoplie de tests statistiques (tests-t, analyse de variance, test de Wilcoxon ou de Kruskal–Wallis) peut être calculée à l'aide d'un modèle linéaire, tandis que des modèles aussi divers que les arbres aléatoires, la régression en composantes principales et les réseaux de neurones multicouches ne sont en réalité que de bêtes modèles linéaires. Ce qui change d'un modèle à l'autre est simplement la méthode d'optimisation (moindres carrés ordinaires, optimisation sous contrainte ou par descente de gradient stochastique), de même que le choix des variables explicatives (bases de spline pour la régression nonparamétrique, variables indicatrices pour les arbres, fonctions d'activations pour les réseaux de neurones). Ce chapitre porte sur la formulation de modèles linéaires, l'interprétation des coefficients et les tests usuels reliés à ces modèles. Certains modèles bien connus, comme l'analyse de variance, seront présentés comme cas spéciaux du modèle de régression linéaire.

Afin de rendre plus tangible le concept et les notions qui touchent aux modèles linéaires, on présentera ces notions dans le cadre d'un exemple. On s'intéresse à la discrimination salariale dans un collège américain, au sein duquel une étude a été réalisée pour investiguer s'il existait des inégalités salariales entre hommes et femmes. Le jeu de données contient les variables suivantes

— `salaire` : salaire de professeurs pendant l'année académique 2008–2009 (en milliers de dollars USD).
— `echelon` : échelon académique, soit adjoint (`adjoint`), aggrégé (`aggrege`) ou titulaire (`titulaire`).
— `domaine` : variable catégorielle indiquant le champ d'expertise du professeur, soit appliqué (`applique`) ou théorique (`theorique`).
— `sexe` : indicateur binaire pour le sexe, `homme` ou `femme`.
— `service` : nombre d'années de service.
— `annees` : nombre d'années depuis l'obtention du doctorat.

Une analyse exploratoire des données est de mise avant d'ébaucher un modèle. Si le salaire augmente au fil des ans, on voit que l'hétérogénéité change en fonction de l'échelon et qu'il y a une relation claire entre ce dernier et le nombre d'années de service (les professeurs n'étant éligibles à des promotions qu'après un certain nombre d'années). Les professeurs adjoints qui ne sont pas

TABLE 10 – Tableau de contingence donnant le nombre de professeurs du collège par sexe et par échelon académique.

	adjoint	aggrege	titulaire
femme	11	10	18
homme	56	54	248

promus sont généralement mis à la porte, aussi il y a moins d'occasions pour que les salaires varient sur cette échelle.

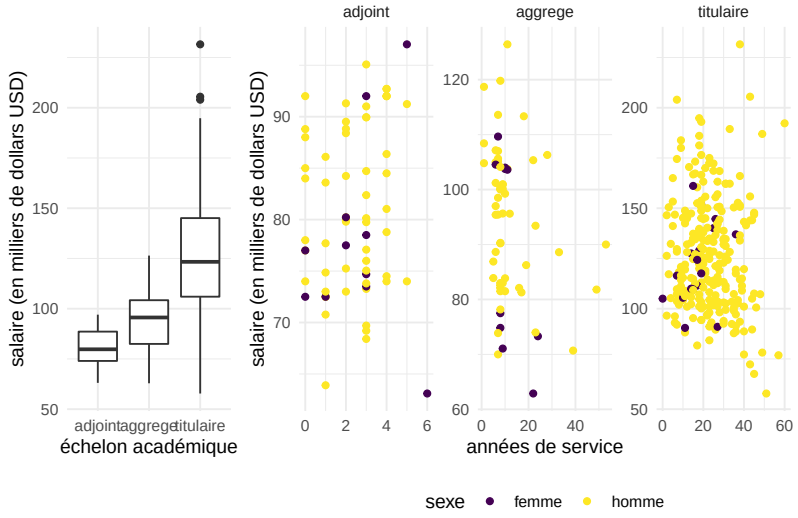

FIGURE 13 – Analyse exploratoire des données `college` : répartition des salaires en fonction de l'échelon et du nombre d'années de service

Ainsi, le salaire augmente avec les années, mais la variabilité croît également. Il y a peu de femmes dans l'échantillon : moins d'information signifie moins de puissance pour détecter de petites différences de salaire. Si on fait un tableau de contingence de l'échelon et du sexe, on peut calculer la proportion relative homme/femme dans chaque échelon : 16% des profs adjoints, 16% pour les aggrégés, mais seulement 7% des titulaires alors que ces derniers sont mieux payés en moyenne.

Le modèle linéaire simple n'inclut qu'une variable explicative et consiste en une droite d'équation $y = \beta_0 + \beta_1 \mathrm{X}$ qui passe à travers un nuage de points. La Figure 14 montre la droite de régression dans le nuage de points formé par les couples $\{\mathrm{X}_i, y_i\}$, où y_i est le `salaire` et X est `service`.

Une infinité de droites pourraient passer dans le nuage de points ; il faut donc choisir la meilleure

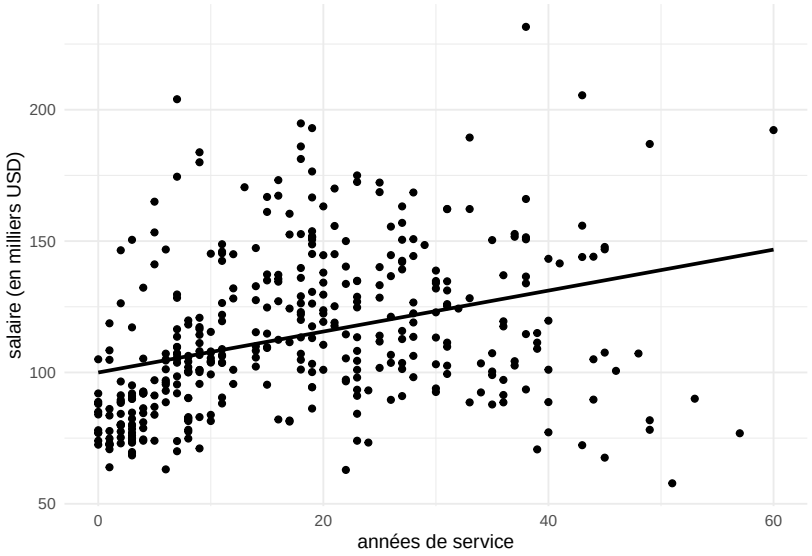

FIGURE 14 – Régression linéaire simple pour le salaire en fonction des années de service ; la droite satisfait le critère des moindres carrés.

droite (selon un critère donné). La section aborde le choix de ce critère et l'estimation des paramètres de l'équation de la droite.

2.2 Moindres carrés ordinaires

Les estimateurs des moindres carrés ordinaires $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \ldots, \widehat{\beta}_p)$ sont les paramètres qui minimisent simultanément la distance euclidienne entre les observations Y_i et les **valeurs ajustées**

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 \mathrm{X}_{i1} + \cdots + \widehat{\beta}_p \mathrm{X}_{ip}, \qquad i = 1, \ldots, n.$$

En d'autres mots, les estimateurs des moindres carrés sont la solution du problème d'optimization convexe

$$\widehat{\boldsymbol{\beta}} = \min_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2 = \min_{\boldsymbol{\beta}} \| \boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta} \|^2$$

Ce système d'équation a une solution explicite qui est plus facilement exprimée en notation matricielle. Soit les matrices et vecteurs

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} \mathrm{X}_{11} & \mathrm{X}_{12} & \cdots & \mathrm{X}_{1p} \\ \mathrm{X}_{21} & \mathrm{X}_{22} & \cdots & \mathrm{X}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{X}_{n1} & \mathrm{X}_{n2} & \cdots & \mathrm{X}_{np} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$$

Le modèle en notation matricielle s'écrit de manière compacte,

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon};$$

chaque ligne de la matrice correspond à l'équation (1) avec une observation par ligne. L'estimateur des moindres carrés ordinaires résoud le problème d'optimisation non-contraint

$$\widehat{\boldsymbol{\beta}} = \min_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}} (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}).$$

Une preuve est fournie dans l'Annexe. Si le rang de la matrice $\mathbf{X}$ est dimension $n \times (p+1)$ est de rang $p+1$, l'unique solution du problème d'optimisation est

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{Y}. \tag{2}$$

Que représente les moindres carrés en deux dimensions ? L'estimateur est celui qui minimise la somme du carré des résidus ordinaires. Le ie **résidu ordinaire** $e_i = y_i - \widehat{y}_i$ est la distance *verticale* entre un point y_i et la valeur ajustée $\widehat{y}_i$, soit les traits bleus de la Figure 15.

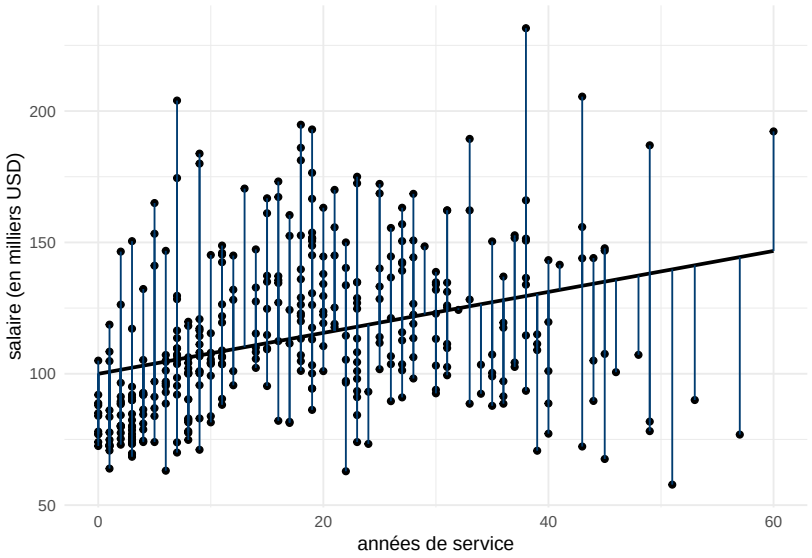

FIGURE 15 – Illustration des résidus ordinaires ajoutés à la droite de régression.

Remarque (Géométrie des moindres carrés). Si on considère les n observations comme un vecteur (colonne), le terme $\mathbf{X}\widehat{\boldsymbol{\beta}}$ correspond à la projection sur l'espace linéaire engendré par les colonnes de la matrice $\mathbf{X}$, $\mathcal{S}_{\mathbf{X}}$ du vecteur de réponse $\boldsymbol{y}$. Les résidus ordinaires sont donc orthogonaux à $\mathcal{S}_{\mathbf{X}}$ par construction et les résidus sont orthogonaux aux valeurs ajustées, $\boldsymbol{e}^\top \widehat{\boldsymbol{y}} = 0$. Une conséquence directe est que la corrélation linéaire entre $\boldsymbol{e}$ et $\widehat{\boldsymbol{y}}$ est zéro ; cette propriété nous servira dans les diagnostics graphiques.

Remarque (Complexité du calcul des moindres carrés ordinaires). Tangente : en apprentissage automatique, on utilise souvent un algorithme du gradient (stochastique) pour estimer les estimés des moindres carrés ordinaires. Or, à moins d'avoir des tailles d'échantillons n ou un nombre de covariables p subséquent (pensez échelle Google), une solution approximative ne devrait pas être préférée à la solution exacte! D'un point de vue numérique, l'opération la plus coûteuse est le calcul de l'inverse de la matrice $\mathbf{X}^\top\mathbf{X}$, qui de dimension $(p+1)\times(p+1)$. Règle générale, on n'inverse pas directement cette matrice car ce n'est pas la façon la plus numériquement stable d'obtenir la solution. **R** utilise la décomposition QR qui a une complexité de $\mathrm{O}(np^2)$ (l'ordre du nombre de flops ou d'opérations pour le calcul). Une alternative plus coûteuse, mais plus stable numériquement, est la décomposition en valeurs singulières (même ordre en terme de calculs).

Mais trève de disgression mathématique : tout bon logiciel calculera pour vous les estimés des moindres carrés. Retenez que l'on minimise une forme quadratique qui admet une solution explicite et unique pour autant que les colonnes de $\mathbf{X}$ ne soient pas colinéaires. Si vous avez plus d'une variable explicative, les valeurs ajustées seront situées sur un hyperplan (peu commode à représenter graphiquement). Maîtriser le langage associé à la régression (notamment les résidus ordinaires, les valeurs ajustées, etc.) est nécessaire pour la continuation.

2.3 Interprétation des paramètres du modèles

Que représentent les paramètres $\boldsymbol{\beta}$ du modèle linéaire? Dans le cas simple présenté dans la Figure 14 où l'équation de la droite est de la forme $\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 \mathrm{X}_1$, β_0 est l'ordonnée à l'origine (la valeur moyenne de Y quand $\mathrm{X}_1 = 0$) et β_1 est la pente, soit l'augmentation moyenne de Y quand X_1 augmente d'une unité.

Dans certains cas, l'interprétation de l'ordonnée à l'origine n'est pas valide car c'est un **non-sens** : la valeur $\mathrm{X}_1 = 0$ n'est pas plausible (par exemple, si X_1 est la taille d'un humain). De même, il peut arriver qu'il n'y ait pas d'observations dans le voisinage de $\mathrm{X}_1 = 0$, même si cette valeur est plausible ; on parle alors d'extrapolation.

Si les colonnes de $\mathbf{X}$ sont arbitraires, il est d'usage d'inclure une constante : cela revient à inclure $\mathbf{1}_n$ comme colonne de la matrice de plan d'expérience $\mathbf{X}$. Parce que les résidus sont orthogonaux aux colonnes de $\mathbf{X}$, leur moyenne est zéro, $n^{-1}\mathbf{1}_n^\top \boldsymbol{e} = \bar{\boldsymbol{e}} = 0$. En général, on peut obtenir des résidus centrés en incluant comme régresseurs dans la matrice $\mathbf{X}$ des vecteurs colonnes qui sont collinéaires avec $\mathbf{1}_n$.

Dans notre exemple, l'équation de la droite ajustée de la Figure 14 est

$$\widehat{\texttt{salaire}} = 99.975 + 0.78\texttt{service}.$$

Ainsi, le salaire moyen d'un nouveau professeur serait 99974.653 dollars, tandis que l'augmentation moyenne annuelle du salaire est 779.569 dollars.

Si la variable réponse Y doit être *continue*, il n'y a aucune restriction pour les variables explicatives. Le cas des variables explicatives binaires est illustratif : ces variables sont encodées numériquement

à l'aide de 0/1. Considérons par exemple le sexe des professeurs de l'étude, par exemple

$$\texttt{sexe} = \begin{cases} 0, & \text{pour les hommes,} \\ 1, & \text{pour les femmes.} \end{cases}$$

L'équation du modèle linéaire simple qui n'inclut que cette variable catégorielle à deux niveaux, `sexe`, s'écrit $\texttt{salaire} = \beta_0 + \beta_1 \texttt{sexe} + \varepsilon$. Posons μ_0 le salaire moyen des hommes et μ_1 celui des femmes. L'ordonnée à l'origine β_0 s'interprète comme d'ordinaire : c'est le salaire moyen quand $\texttt{sexe} = 0$, autrement dit $\beta_0 = \mu_0$. On peut écrire l'équation de l'espérance conditionnelle pour chacune des catégories,

$$\mathsf{E}(\texttt{salaire} \mid \texttt{sexe}) = \begin{cases} \beta_0, & \texttt{sexe} = 0 \text{ (homme)}, \\ \beta_0 + \beta_1 & \texttt{sexe} = 1 \text{ (femme)}. \end{cases}$$

Un modèle linéaire qui contient uniquement une variable binaire X comme régresseur équivaut à spécifier une moyenne différente pour deux groupes ; la moyenne des femmes est $\mathsf{E}(\texttt{salaire} \mid \texttt{sexe} = 1) = \beta_0 + \beta_1 = \mu_1$ et $\beta_1 = \mu_1 - \mu_0$ représente la différence entre la moyenne des hommes et celles des femmes. L'estimateur des moindres carrés $\widehat{\beta}_0$ est la moyenne du salaire des hommes de l'échantillon et $\widehat{\beta}_1$ est la différence des moyennes empiriques entre femmes et hommes. Cette paramétrisation en terme d'**effets différentiels** est particulièrement utile si on veut tester s'il y a une différence moyenne de salaire entre les deux sexe car cela revient à tester $\mathcal{H}_0 : \beta_1 = 0$. Si on voulait obtenir directement la moyenne, il faudrait remplacer la matrice de plan d'expérience $[\mathbf{1}_n, \texttt{sexe}]$ par $[\mathbf{1}_n - \texttt{sexe}, \texttt{sexe}]$ pour obtenir un modèle équivalent. Règle générale, il n'est pas recommandé de retirer l'ordonnée à l'origine même si l'espace linéaire engendré par les colonnes de $\mathbf{X}$ contient $\mathbf{1}_n$.

Si on ajuste un modèle de régression linéaire pour les données `college`, on obtient un salaire moyen de $\widehat{\beta}_0 = 115.09$ milliers de dollars USD pour les hommes et une différence moyenne de salaire entre femmes et hommes de $\widehat{\beta}_1 = 14.088$ milliers de dollars. Puisque l'estimé est négatif, les femmes sont moins payés : ce modèle n'est en revanche pas suffisant pour déterminer s'il y a inéquité salariale : la Figure 14 montre que le nombre d'années de service et l'échelon académique impactent fortement le salaire, or il n'est pas dit que la répartition des sexes au sein des échelons est comparable (et ce n'est pas le cas).

Même si le modèle linéaire simple définit une droite, cette dernière n'a de sens qu'en 0 ou 1 ; la Figure 16 montre un estimé de la densité et la répartition des points (décalés) dans l'échantillon selon le sexe, avec la moyenne de chacun. On voit bien que la droite passe par la moyenne de chaque groupe.

Plus généralement, il est possible de considérer une variable catégorielle à k niveaux. Comme pour la variable binaire, on ajoute au modèle $k-1$ variables indicatrices en plus de l'ordonnée à l'origine : si on veut modéliser k moyennes, il est logique de n'inclure que k paramètres. On choisira comme dans l'exemple avec le sexe une **catégorie de référence** dont la moyenne sera encodée

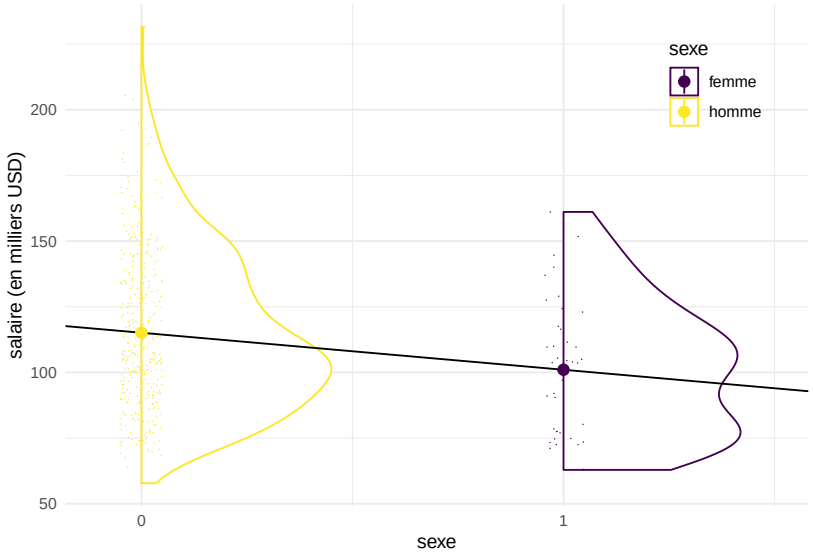

FIGURE 16 – Modèle linéaire simple pour les données `college` en fonction de la variable binaire sexe : bien que le modèle définisse une ligne, seule la valeur en 0/1 est réalisable.

par l'ordonnée à l'origine β_0. Les autres paramètres seront des effets différentiels relatifs à cette catégorie. Prenons pour exemple l'échelon académique, une variable catégorielle ordinale à trois niveaux (adjoint, aggrégé, titulaire). On ajoute deux variables binaires $\mathrm{X}_1 = \mathsf{I}(\texttt{echelon} = \texttt{aggrege})$ et $\mathrm{X}_2 = \mathsf{I}(\texttt{echelon} = \texttt{titulaire})$; l'élément i de la colonne X_1 vaut 1 si le professeur est aggrégé et zéro autrement. Le modèle linéaire

$$\texttt{salaire} \mid \texttt{echelon} = \beta_0 + \beta_1 \mathrm{X}_1 + \beta_2 \mathrm{X}_2 + \varepsilon,$$

et l'espérance conditionnelle du salaire s'écrit

$$\mathsf{E}(\texttt{salaire} \mid \texttt{echelon}) = \begin{cases} \beta_0, & \texttt{echelon} = \texttt{adjoint}, \\ \beta_0 + \beta_1 & \texttt{echelon} = \texttt{aggrege}, \\ \beta_0 + \beta_2 & \texttt{echelon} = \texttt{titulaire}, \end{cases}$$

Ainsi, β_1 (respectivement β_2) est la différence de salaire moyenne entre professeurs titulaires (respectivement aggrégés) et professeurs adjoints. Le choix de la catégorie de référence est arbitraire et le modèle ajusté est le même : seule l'interprétation des coefficients change. Pour une variable ordinale, il vaut mieux choisir la plus petite ou la plus grande des modalités pour faciliter les comparaisons.

Les modèles que nous avons ajusté jusqu'à maintenant ne sont pas adéquats parce qu'ils ignorent des variables qui sont importantes pour expliquer le modèle : la Figure 13 illustre en effet que

l'échelon est une composante essentielle pour expliquer les variations de salaire au sein du collège. On peut (et on doit) donc inclure plusieurs variables simultanément pour avoir un modèle adéquat. Avant de procéder, on considère l'interprétation des paramètres quand on utilise plus d'une variable explicative dans le modèle.

Soit le modèle $Y = \beta_0 + \beta_1 \mathrm{X}_1 + \cdots + \beta_p \mathrm{X}_p + \varepsilon$. L'ordonnée à l'origine β_0 représente la valeur moyenne de Y quand *toutes* les covariables du modèle sont égales à zéro,

$$\beta_0 = \mathsf{E}(Y \mid \mathrm{X}_1 = 0, \mathrm{X}_2 = 0, \ldots, \mathrm{X}_p = 0).$$

De nouveau, cette interprétation peut ne pas être sensée ou logique selon le contexte de l'étude. Le coefficient β_j ($j \geq 1$) peut quant à lui être interprété comme l'augmentation moyenne de l'espérance de la variable réponse Y quand X_j augmente d'une unité, toutes choses étant égales par ailleurs (*ceteris paribus*). Par exemple, l'interprétation de β_1 est

$$\begin{aligned}\beta_1 = \mathsf{E}(Y \mid \mathrm{X}_1 &= x_1 + 1, \mathrm{X}_2 = x_2, \ldots, \mathrm{X}_p = x_p) \\ &\quad - \mathsf{E}(Y \mid \mathrm{X}_1 = x_1, \mathrm{X}_2 = x_2, \ldots, \mathrm{X}_p = x_p) \\ &= \{\beta_0 + \beta_1(x_1 + 1) + \beta_2 x_2 + \cdots + \beta_p \mathrm{X}_p\} \\ &\quad - \{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p \mathrm{X}_p\}\end{aligned}$$

Il n'est pas toujours possible de fixer la valeur des autres colonnes de $\mathbf{X}$ si plusieurs colonnes contiennent des transformations ou des fonctions d'une même variable explicative. Par exemple, on pourrait par exemple considérer un polynôme d'ordre k (normalement, $k \leq 3$ en pratique),

$$Y = \beta_0 + \beta_1 \mathrm{X} + \beta_2 \mathrm{X}^2 + \ldots + \beta_k \mathrm{X}^k + \varepsilon.$$

Si l'on inclut un terme d'ordre k, X^k, il faut **toujours** inclure les termes d'ordre inférieur $1, \mathrm{X}, \ldots, \mathrm{X}^{k-1}$ pour l'interprétabilité du modèle résultant (autrement, cela revient à choisir un polynôme en imposant que certains coefficients soient zéros). L'interprétation des effets des covariables nonlinéaires (même polynomiaux) est complexe parce qu'on ne peut pas « fixer la valeur des autres variables » : l'effet d'une augmentation d'une unité de X *dépend de la valeur de cette dernière*.

Exemple 2.1 (Données automobile)**.** Considérons un modèle de régression linéaire pour l'autonomie d'essence en fonction de la puissance du moteur pour différentes voitures dont les caractéristiques sont données dans le jeu de données `automobiles`. Le modèle postulé incluant un terme quadratique est

$$\texttt{autonomie}_i = \beta_0 + \beta_1 \texttt{puissance}_i + \beta_2 \texttt{puissance}_i^2 + \varepsilon_i$$

Afin de comparer l'ajustement du modèle quadratique, on peut inclure également la droite ajustée du modèle de régression simple qui n'inclut que puissance.

À vue d'oeil, l'ajustement est meilleur pour le modèle quadratique : nous verrons plus tard à l'aide de test si cette observation est vérifiée statistiquement. On voit aussi dans la Figure 17 que l'autonomie

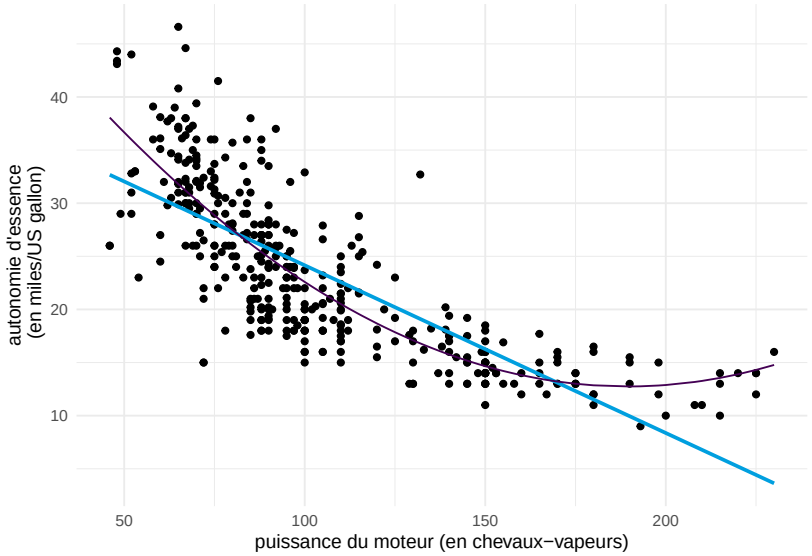

FIGURE 17 – Modèle de régression avec terme quadratique pour la puissance

d'essence décroît rapidement quand la puissance croît entre 0 et 189.35, mais semble remonter légèrement par la suite pour les voitures qui un moteur de plus de 200 chevaux-vapeurs, ce que le modèle quadratique capture. Prenez garde en revanche à l'extrapolation là où vous n'avez pas de données (comme l'illustre remarquablement bien le modèle cubique de Hassett pour le nombre de cas quotidiens de coronavirus).

La représentation graphique du modèle polynomial de degré 2 présenté dans la Figure 17 peut sembler contre-intuitive, mais c'est une projection en 2D d'un plan 3D de coordonnées $\beta_0 + \beta_1 x - y + \beta_2 z = 0$, où $x = \texttt{puissance}$, $z = \texttt{puissance}^2$ et $y = \texttt{autonomie}$. La physique et le bon-sens imposent la contrainte $z = x^2$, et donc les valeurs ajustées vivent sur une courbe dans un sous-espace du plan ajusté, représenté en gris dans la Figure 18.

FIGURE 18 – Représentation graphique 3D du modèle de régression linéaire pour les données `automobile`.

Remarque (Utilisation de bases polynomiales pour les effets nonlinéaires). Règle générale, on utilise des représentations flexibles (bases de splines) plutôt que des modèles polynomiaux pour le lissage si la relation entre une variable Y et une variable explicative X est nonlinéaire. Une compréhension de la physique du système à l'étude, ou bien un modèle théorique permet aussi de guider le choix des fonctions (non)linéaires à utiliser.

Le coefficient β_j est la contribution *marginale* de X_j quand les autres covariables sont incluses

TABLE 11 – Estimés des coefficients du modèle linéaire pour les données `college` (en dollars USD, arrondis à l'unité).

$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\beta}_4$	$\widehat{\beta}_5$
86596	-4771	-13473	14560	49160	-89

dans le modèle. On peut représenter graphiquement cet effet en projetant les vecteurs Y et X_j dans le complément orthogonal de $\mathbf{X}_{-j}$. Le diagramme de régression partielle est un diagnostic graphique qui illustre la valeur ajoutée de X_j : il montre en ordonnée (axe des y), les résidus du modèle de régression pour Y avec toutes les variables explicatives sauf X_j, et en abcisse (axe des x), les résidus de la régression de X_j sur les autres variables explicatives. La droite de régression qui satisfait le critère des moindre carrés pour ce nuage de points passe par $(0,0)$ et sa pente est $\hat{\beta}_j$. Ce diagnostic est particulièrement utile pour détecter l'impact de valeurs aberrantes ou la colinéarité.

Exemple 2.2 (Inéquité salariale dans un collège américain)**.** On considère les données `college` et un modèle de régression qui inclut le sexe, l'échelon académique, le nombre d'années de service et le domaine d'expertise (appliquée ou théorique).

Si on multiplie le salaire par mille, le modèle linéaire postulé s'écrit

$$\begin{aligned}\texttt{salaire} \times 1000 = \beta_0 &+ \beta_1 \texttt{sexe}_{\texttt{femme}} + \beta_2 \texttt{domaine}_{\texttt{theorique}} \\ &+ \beta_3 \texttt{echelon}_{\texttt{aggrege}} + \beta_4 \texttt{echelon}_{\texttt{titulaire}} + \beta_5 \texttt{service} + \varepsilon.\end{aligned}$$

L'interprétation des coefficients est la suivante :

- L'ordonnée à l'origine β_0 correspond au salaire moyen d'un professeur adjoint (un homme) qui vient de compléter ses études et qui travaille dans un domaine appliqué : on estime ce salaire à $\widehat{\beta}_0 = 86596$ dollars.
- toutes choses étant égales par ailleurs (même domaine, échelon et années depuis le dernier diplôme), l'écart de salaire entre un homme et un femme est estimé à $\widehat{\beta}_1 = -4771$ dollars.
- *ceteris paribus*, un(e) professeur(e) qui oeuvre dans un domaine théorique gagne β_2 dollars de plus qu'une personne du même sexe dans un domaine appliqué; on estime cette différence à -13473 dollars.
- *ceteris paribus*, la différence moyenne de salaire entre professeurs adjoints et aggrégés est estimée à $\widehat{\beta}_3 = 14560$ dollars.
- *ceteris paribus*, la différence moyenne de salaire entre professeurs adjoints et titulaires est de $\widehat{\beta}_4 = 49160$ dollars.
- au sein d'un même échelon, chaque année supplémentaire de service mène à une augmentation de salaire annuelle moyenne de $\widehat{\beta}_5 = -89$ dollars.

On voit que les femmes sont moins payées que les hommes : reste à savoir si cette différence est statistiquement significative. L'estimé de la surprime annuelle due à l'expérience est négative, un résultat contre-intuitif au vu de la Figure 14 qui montrait une augmentation notable du salaire avec les années. Cette représentation graphique est trompeuse : la Figure 13 montrait l'impact important de l'échelon académique. Une fois tous les autres facteurs pris en compte, le nombre d'années de service n'apporte que peu d'information au modèle et le diagramme de régression partielle de la Figure 19 illustre l'absence de corrélation entre salaire et la partie non expliquée par les autres covariables ; les gens avec un grand nombre d'années de service sont moins payés que certains de leurs collègues, ce qui explique la pente négative.

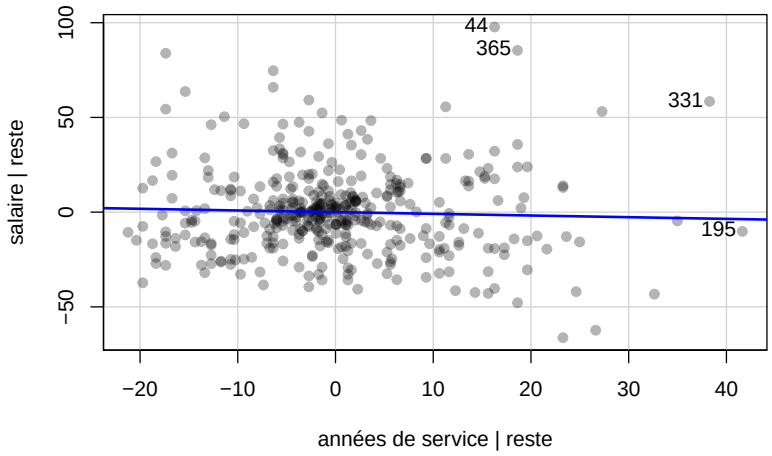

FIGURE 19 – Diagramme de régression partielle pour les années de service dans le modèle de régression linéaire pour les données `college`.

3 Inférence basée sur la vraisemblance

4 Modèles linéaires généralisés

Dans certains scénarios, on aimerait bâtir un modèle de régression avec une variable Y qui ne sera pas forcément continue : cela inclut des exemples de variable réponse binaire, entière ou non-négative. La régression linéaire et l'estimation par le biais des moindres carrés n'est dans cette optique pas idéale puisque notre droite ajusté ne respectera pas nécessairement ces contraintes.

Ce chapitre est une introduction aux modèles linéaire généralisés, une extension de la régression linéaire. On s'intéresse tout particulièrement à la modélisation de données binaires, de proportions

et de données de dénombrement. On considère la régression de Poisson et la régression logistique en se concentrant sur le cas où les observations sont indépendantes : les *modèles linéaires généralisés mixtes*, qui servent à l'analyse de données corrélées ou longitudinales, sont traités dans le cours MATH 80621.

4.1 Principes de base

Le point de départ est le même que pour la régression linéaire : on a un échantillon aléatoire simple d'observations postulées indépendantes, $(\boldsymbol{Y}, \mathbf{X})$, où Y est notre variable réponse et $\mathrm{X}_1, \ldots, \mathrm{X}_p$ sont les p variables explicatives qu'on suppose fixe (non-stochastiques). Une fois de plus, on cherche à construire un modèle pour la moyenne de la variable réponse à l'aide d'une combinaison linéaire des variables explicatives.

Soit $\mu_i = \mathsf{E}(Y_i \mid \mathbf{X}_i)$ l'espérance ou moyenne conditionnelle de Y_i sachant la valeur des variables explicatives. On dénote par η_i la combinaison linéaire de ces variables qui servira à modéliser la moyenne de la réponse,

$$\eta_i = \beta_0 + \beta_1 \mathrm{X}_{i1} + \cdots + \beta_p \mathrm{X}_{ip}.$$

Nos outils de base dans la constructions d'un modèle linéaire généralisé sont les suivants :

— Une loi de probabilité pour la variable aléatoire Y qui appartienne à la famille de lois exponentielles de dispersion (normale, binomiale, Poisson, gamma, …).
— Une composante déterministique, le **prédicteur linéaire** $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$, où $\mathbf{X}$ est une matrice $n \times (p+1)$ avec colonnes $\mathbf{1}_n, \mathbf{X}_1, \ldots, \mathbf{X}_p$ qui représentent l'ordonnée à l'origine et chacune des variables explicatives pour les n observations, de même que les coefficients $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$.
— Une fonction monotone g, appelée **fonction de liaison**, qui relie la moyenne de Y_i au prédicteur linéaire, $g(\mu_i) = \eta_i$.

5 Données corrélées et longitudinales

On couvrira dans ce chapitre une extension du modèle linéaire afin de relaxer le postulat d'indépendance entre observations. Ce dernier ne tient en effet pas la route si la base de données contient des mesures répétées sur une unité. On s'intéressera en particulier à l'analyse de données longitudinales, pour lesquelles on a à disposition de brèves séries chronologiques. La clé pour incorporer la dépendance entre mesures est la modélisation parsimonieuse de la structure de covariance, laquelle nous permettra également de traiter les données groupées et de traiter un cas particulier d'hétéroscédasticité pour ces mêmes données.

5.1 Données longitudinales

Dans les **données longitudinales**, aussi appelées données de panel ou mesures répétées, on a des mesures répétées sur la même variables réponse pour m unités ou individus, typiquement

à différents temps. Dans ce cadre, il est logique de supposer que les observations de différents individus sont indépendantes, mais pas les mesures pour un même individu. Voici quelques exemples de données longitudinales :

— Un sondage sur la satisfaction face au service à clientèle est envoyé aux clients d'une compagnie suite à un appel de leur part : la variable réponse est la moyenne d'échelles de Likert.
— On prend trois mesures de circonférence d'arbres à différentes hauteurs pour estimer le volume de bois d'une forêt.
— une étude suit une cohorte pour évaluer l'attitude de jeunes adolescents et leur attitudes envers la consommation de drogues et de stupéfiants, avec un suivi annuel.

Règle générale, on s'intéresse à la modélisation de la moyenne de la réponse en fonction d'autres variables explicatives. Si les mesures sont collectées dans le temps, on peut aussi représenter graphiquement l'évolution temporelle en identifiant la variable représentant le temps et en dessinant un diagramme à ligne brisé par individu : ce type de graphique est appelé **diagramme spaghetti**.

Dans ce qui suit, on s'intéresse plus en détail à deux exemples.

Exemple 5.1 (Étude longitudinale sur l'effet de la pollution atmosphérique sur la mortalité dans six villes américaines.)**.** L'article *An Association Between Air Pollution And Mortality In Six U.S. Cities* est le fruit du travail de chercheurs de l'Université de Harvard paru dans la revue *The New England Journal of Medicine* en 1993. L'étude se penchait sur le lien entre la pollution atmosphérique, en particulier l'effet des particules fines, et la mortalité en milieu urbain en suivant une cohorte de 8111 adultes de six villes américains pendant une période allant de 14 à 16 ans. Cette étude longitudinale avait démarrée en 1979. La base de données `fev1` contient 300 filles parmi les 13 379 enfants nés en 1967 ou ultérieurement dans six villes. La plupart des enfants ont été enrollés dans l'étude en première ou deuxième année primaire, soit vers l'âge de six ou sept ans. À chaque examen annuel, on a mesuré le volume d'air évacué durant la premier seconde d'une expiration, appelé volume expiratoire maximal ou VEM1.

Outre la variable réponse VEM1, on a également à disposition la taille, l'âge et l'identifiant de l'enfant. La Figure 20 montre les courbes de croissance du volume expiratoire maximal en fonction de l'âge de la personne. On peut remarquer que la croissance est presque linéaire, avec une diminution de la pente à partir de 14 ans. Si les profils semblent suivre la même tendance, on voit une forte fluctuation autour de la moyenne. Toutes les courbes ne démarrent ni ne finissent au même âge : une analyse exploratoire des identifiants des enfants nous indique qu'on a entre 1 et 12 mesures par enfant, pour un total de 1994 mesures. Les données sont stockées en format long : chaque ligne correspond à une mesure de VEM1, tandis qu'on recense plusieurs mesures par identifiant (`id`). L'âge de la fille (`age`) représente la variable temps, et sa taille (en mètres) évolue au fil des années. L'âge et la taille initiale lors de la première collecte de données sont fixes et identiques pour toutes les mesures d'une même personne.

Conceptuellement, on pourrait envisager que chaque fille a une courbe de croissance spécifique

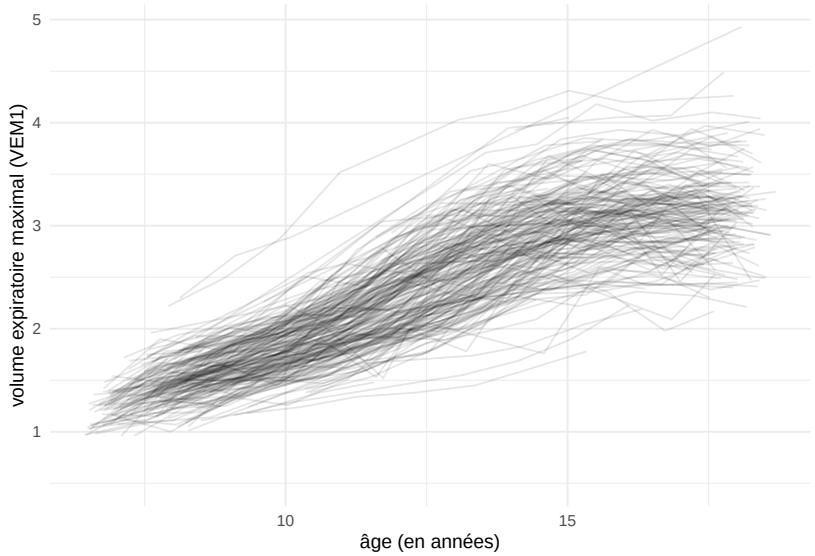

FIGURE 20 – Diagramme spaghetti des données de spirométrie de l'étude *Six Cities Study of Air Pollution and Health*.

de sa capacité respiratoire et qui est mesurée avec une erreur à cause de la variabilité de la prise de mesure. On peut également envisager une moyenne théorique due à des facteurs biologique, autour duquelle les courbes de croissance des différentes filles fluctuent.

Notre modèle statistique devra prendre en compte que les mesures sur un même sujet sont corrélées : par exemple, si une courbe de croissance est plus élevée que la moyenne globale, elle aura tendance à rester au dessus de cette moyenne peu importe l'âge de la personne. On pourrait également postuler que les fluctuations sont corrélées dans le temps : une fois la moyenne globale soustraite, les mesures d'une même personnes sont plus semblables quand elles sont rapprochées. En revanche, on peut raisonnablement supposer que les observations de filles différentes sont indépendantes. Ainsi, si on s'intéressait uniquement au comportement à un âge donné, on pourrait faire l'analyse d'une coupe transversale des données en calculant la moyenne empirique.

Exemple 5.2 (Programme thérapeutique *Beating the Blues*.)**.** *Beating the Blues* est un programme clinique qui vise à traiter les personnes atteintes d'anxiété et de dépression clinique légère ou modéré. Comparativement aux autres thérapies cognitivo-comportementales, les rencontres avec un(e) thérapeute sont remplacées par des capsules vidéos et une prise en charge en ligne pour amener les patients à prendre en charge leur guérison. Les personnes qui se sont joints à l'étude ont été assignés de manière aléatoire à deux traitements, soit *Beating the Blues* ou le traitement usuel. L'objectif dans ce exemple est de comparer l'efficacité relative des traitements, mais puisque les individus ne sont assignés qu'à l'un ou l'autre des traitements, on ne pourra que comparer les diffé-

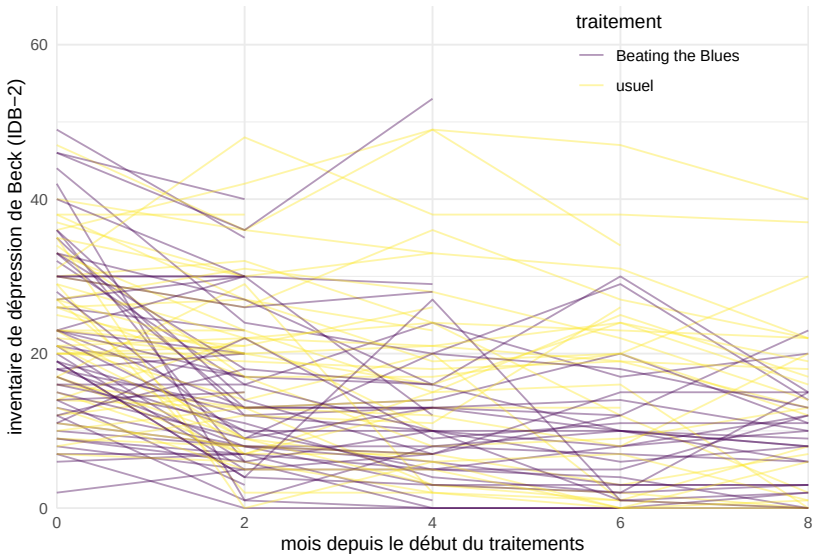

FIGURE 21 – Diagramme spaghetti pour les données de l'étude clinique sur la dépression et du programme comportemental *Beating the Blues*.

rences moyennes de traitement. Le diagramme spaghetti de la Figure 21 illustre le comportement erratique de l'évolution de la santé mentale des individus telle que mesurée à l'aide de l'inventaire de dépression de Beck, un questionnaire de 21 items. Si la décroissance semble en moyenne plus grande pour *Beating the Blues*, c'est loin d'être certain : il y a beaucoup de pertes de suivi et la variabilité inter-individus est énorme puisque le point de départ des individus est différent ; on pourrait penser à standardiser la courbe par la mesure initiale pour faciliter la comparaison. On remarque que tous les patients sont suivis à intervalle réguliers de deux mois et que les données sont strictement positives.

Les données longitudinales ne sont pas toujours stockées dans un format qui soit convenable pour l'analyse. En effet, il est commun d'enregistrer ces données en **format large**, auquel cas chaque ligne représente un individu différent et les colonnes contiennent à la fois des variables explicatives et les différentes répétitions de la variables réponse $y_1, \ldots, y_{n_i}$.

Une base de données en **format long** contient une seule valeur de la variable réponse par ligne, avec une colonne additionnelle indiquant l'identifiant (temporel) de l'observation. On peut passer de format large en format long en transformant les noms de colonnes en étiquettes de la variables d'identification. De même, on suppposera par la suite que les données sont ordonnées par individu, puis chronologiquement.

En format long, les variables explicatives qui sont fixes au sein d'une unité pour toutes les répétitions seront copiées sur chaque ligne. Il faut faire attention au calcul des statistiques descriptives : ces

dernières ne seront pas correctes sauf si on ne conserve qu'une copie pour chaque unité. En particulier, l'estimé de l'écart-type de la variable explicative sera sous-estimé parce qu'on gonfle artificiellement la taille de l'échantillon lors de la duplication.

Des données longitudinales ou mesures répétées comprennent plusieurs mesures pour une même unité, ce qui engendre de la corrélation entre les réponses d'une même unité.

Un diagramme spaghetti (ligne brisée en fonction du temps pour chaque unité) permet de visualiser l'évolution chronologique de la variable réponse.

Lorsqu'on analyse des données longitudinales, on doit identifier par le biais d'une analyse exploratoire plusieurs facteurs qui impacteront nos analyses. Il faut notamment déterminer

— si les données sont enregistrées en format long ;
— si les données sont mesurées à des intervalles réguliers ;
— s'il y a autant de mesures répétées pour chaque individu (échantillon balancé) ;
— lesquelles des variables explicatives varient dans le temps et lesquelles sont fixes.

5.2 Modélisation de la matrice de covariance

Pour un vecteur aléatoire $\boldsymbol{Y}$, on définit la matrice de covariance comme étant la matrice symétrique $n \times n$

$$\mathsf{Co}(\boldsymbol{Y}) = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \cdots & \sigma_{2n} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \ddots & \sigma_{3n} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \sigma_{n3} & \cdots & \sigma_n^2 \end{pmatrix}.$$

Le ie élément de la diagonale de $\mathsf{Co}(\boldsymbol{Y})$ est la variance de Y_i et $\boldsymbol{\Sigma}$ est symmétrique, donc $\sigma_{ij} = \sigma_{ji}$. Une matrice de covariance est positive définie, d'où $\mathbf{v}^\top \boldsymbol{\Sigma} \mathbf{v}$ pour tout n vecteur $\mathbf{v}$ non-nul.

Le point de départ de notre analyse est la dérivation de l'estimateur des moindres carrés ordinaires du modèle de régression linéaire sous l'hypothèse que les valeurs de la variable réponses sont conditionnellement indépendantes, normales et homoscédastiques avec $Y_i \mid \mathbf{x}_i \sim \mathsf{No}(\mathbf{x}_i\boldsymbol{\beta}, \sigma^2)$.

Si les données sont conditionnellement normales, la matrice de covariance $\boldsymbol{\Sigma}$ encode la structure de dépendance entre les observations : les éléments hors-diagonale sont nuls si les observations sont indépendantes, tandis que les éléments de la diagonale encodent la variance de chaque mesure. Si les données sont homoscédastiques, alors tous les éléments de la diagonale sont identiques et de la même mesure tous les éléments hors-diagonale valent zéro si les données sont indépendantes.

Une façon alternative d'écrire la vraisemblance d'un échantillon de taille n du modèle linéaire classique est de considérer qu'il s'agit d'une seule réalisation d'un vecteur aléatoire de dimension n, avec $\boldsymbol{Y} \sim \mathsf{No}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2, \mathbf{I}_n)$ où $\mathbf{I}_n$ est la matrice identité $n \times n$. À partir de là, on peut logiquement

déduire qu'il suffit de modifier la matrice de covariance $\boldsymbol{\Sigma}$ pour relaxer les postulats d'indépendance et d'homoscédasticité.

Un problème demeure : sans contrainte aucune, la matrice de covariance $\boldsymbol{\Sigma}$ possèdera $n(n+1)/2$ éléments uniques puisqu'elle doit être symmétrique : c'est davantage que le nombre d'observations à disposition! Cette réalité nous contraint donc à paramétriser $\boldsymbol{\Sigma}$ à l'aide d'un modèle doté de quelques paramètres $\boldsymbol{\psi}$ qui pourront être estimés conjointement avec les paramètres de la moyenne $\boldsymbol{\beta}$ par le biais de la méthode du maximum de vraisemblance (et des variantes de cette dernière).

En résumé, la généralisation du modèle linéaire pour prendre en compte la dépendance entre observation passe par la modélisation de la matrice de covariance $\boldsymbol{\Sigma}$ et nous verrons plusieurs modèles possibles pour cette dernière.

La première supposition qui nous permettra de réduire le nombre de paramètres au sein de la matrice de covariance est l'hypothèse que les observations de différentes unités sont**indépendantes**. La deuxième supposition sera le postulat que la structure de covariance intra-unité sera identique pour toutes les unités.

Plus précisément, si les données sont ordonnées par unité et qu'on a m groupes, alors la matrice de covariance $\boldsymbol{\Sigma}$ sera diagonale en bloc,

$$\mathsf{Co}(\boldsymbol{Y}) = \begin{pmatrix} \boldsymbol{\Sigma}_1 & \mathbf{O} & \cdots & \mathbf{O} \\ \mathbf{O} & \boldsymbol{\Sigma}_2 & \cdots & \mathbf{O} \\ \vdots & \ddots & \ddots & \vdots \\ \mathbf{O} & \mathbf{O} & \cdots & \boldsymbol{\Sigma}_m \end{pmatrix}.$$

où $\mathbf{O}$ dénote une matrice de zéro. La covariance inter-groupe est nulle parce qu'on suppose que les données d'unités différentes sont indépendantes les unes des autres.

Si on paramétrise les sous-blocs $\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_m$ qui représentent la covariance au sein d'une unité, on supposera que la structure (et les paramètres) seront identiques, en gardant en tête que ces matrices ne sont pas de la même taille si le nombre de réplications temporelles n'est pas identique pour toutes les unités. En particulier, si n_i dénote le nombre de réplications (temporelles) pour l'unité i avec $\sum_{i=1}^{m} n_i = n$ et que $T = n_1 = \cdots = n_m$, alors chaque matrice bloc sera identique et $\boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_m$. Les paramètres $\boldsymbol{\psi}$ de la matrice de covariance seront estimables parce qu'on aura m réplications de la matrice.

Souvent, on ne s'intéresse pas aux paramètres de la matrice de covariance, $\boldsymbol{\psi}$: ces derniers ne servent qu'à guarantir la validité de l'inférence pour les paramètres de la moyenne, $\boldsymbol{\beta}$.

5.2.1 Covariance non structurée

Le modèle le plus général pour la dépendance intra-individu est une matrice **non structurée** de taille $n_{\max} \times n_{\max}$ où $n_{\max} = \max\{n_1, \ldots, n_m\}$, avec

$$\boldsymbol{\Sigma}_i = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots \sigma_{1n_{\max}} & \\ \sigma_{21} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \sigma_{n_{\max}1} & \cdots & \cdots & \sigma_{n_{\max}n_{\max}} \end{pmatrix}.$$

Le modèle non structuré présenté prend en compte à la fois la corrélation entre observations d'une même unité et l'hétéroscédasticité, en postulant en revanche que cette dernière est la même pour toutes les unités pour une répétition donnée. On pourrait aisément contraindre le modèle de manière à spécifier une variance égale pour toutes les observations en fixant $\sigma_{11} = \cdots = \sigma_{n_{\max}n_{\max}}$.

Ce modèle ne sera logique que si les mesures sur chaque unités sont comparables : par exemple, si le temps de réponse à chaque vague d'un sondage est différent, il serait illogique d'associer la réponse d'individus pour des temps différents. Par contre, si on modélise la circonférence d'arbres et que l'on obtient des mesures pour ces derniers à des hauteurs régulières (un mètre, deux mètres et cinq mètres du sol), alors on pourra associer chaque mesure à un paramètre.

Ainsi, il faut savoir si les données sont mesurées à des intervalles réguliers ou sont comparables. Le nombre de paramètres à estimer, $n_{\max}(n_{\max}1)/2$, restreint son utilisation aux cas où le nombre maximum d'observations par unité est petit et le nombre d'unités m est grand. On va considérer deux autres modèles plus simples.

5.2.2 Modèles de covariance autorégressif

La dépendance entre observations consécutives est plus forte qu'entre deux observations plus distantes. Un modèle simple issu de la litérature des séries chronologiques postule que la corrélation entre deux observations ne dépend pas du temps de la mesure, mais uniquement de la distance $h > 0$ entre deux observations. Le modèle de corrélation autorégressive d'ordre 1, pour deux mesures à temps t et $t+h$, est

$$\mathsf{Cov}(Y_t, Y_{t+h}) = \sigma^2 \rho^h, \qquad |\rho| < 1.$$

Ainsi, on postule que la corrélation entre deux observations à distance h décroît comme une série géométrique : plus la distance est grande entre les observations, plus la corrélation est faible.

Ce modèle n'inclut qu'un seul paramètre de corrélation supplémentaire, ρ, en plus de la variance commune des mesures σ^2. Le modèle autorégressif d'ordre 1, dénoté $\mathsf{AR}(1)$, est ainsi parcimonieux comparativement au modèle de covariance non structurée. On peut aussi garder la même structure

de corrélation en ajoutant une variance différente pour chaque temps donné : le modèle résultant est alors dit hétérogène.

Si les pas de temps sont réguliers et chaque observation est mesurée au temps $t = 0, 1, 2, \ldots$, on peut écrire la matrice de covariance intra-groupe de l'unité i, $\boldsymbol{Y}_i$, sous la forme

$$\boldsymbol{\Sigma}_i = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n_i} \\ \rho & 1 & \rho & \cdots & \rho^{n_i-1} \\ \rho^2 & \rho & 1 & \cdots & \rho^{n_i-2} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \rho^{n_i} & \rho^{n_i-1} & \cdots & \rho & 1 \end{pmatrix}.$$

Sans s'attarder pour l'instant à l'estimation de la structure de covariance, on présente la matrice de corrélation estimée les données *Beating the Blues* avec comme modèle pour la moyenne les variables explicatives `mois` (traitée comme une variable catégorielle), un indicateur binaire pour `medicaments` qui indique si la personne a une prescription et finalement un indicateur pour le traitement.

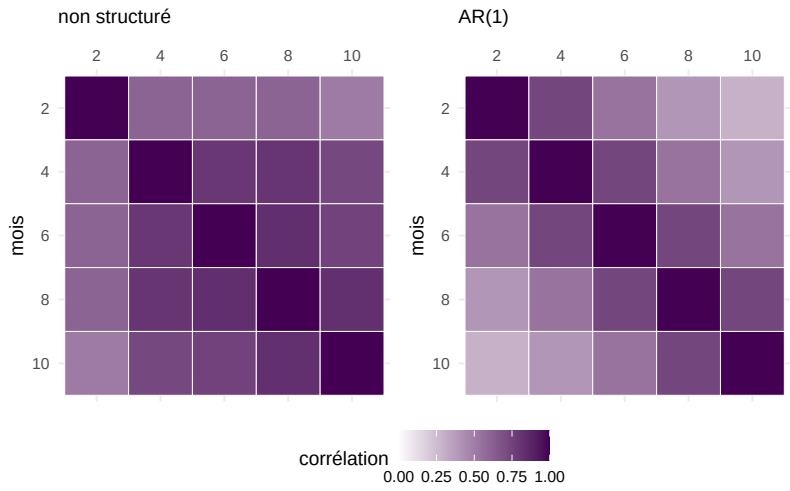

FIGURE 22 – Estimés de la matrice de corrélation pour l'individu 6 avec les données *Beating the Blues*, avec modèle de covariance non structurée (gauche) et modèle autorégressif hétérogène d'ordre 1 (droite).

Les estimés de la corrélation pour **Beating the blues** sont présentés dans la Figure 22 : on voit que les estimés de la corrélation sont très élevés et presque constant pour toutes les observations entre 4 et 10 mois, mais la corrélation est plus faible entre la mesure à 2 mois et les autres estimés.

Par construction, la matrice de corrélation autorégressive d'ordre 1 assume que la corrélation est décroissante ; or, au vu du panneau de gauche, ce modèle ne semble pas adéquat. Cela n'est pas surprenant dans la mesure où les courbes sont systématiquement au dessus ou en dessous de la courbe moyenne ; nous verrons dans le prochain chapitre comment prendre en compte ce décalage individuel de chaque courbe. Pour le modèle non structuré, les estimés de l'écart type ne sont pas très variables : l'estimé va de 12 au temps $t = 6$ mois jusqu'à 9 au temps $t = 10$ mois, mais presque constants pour les premiers 8 mois (4 premières mesures).

Quel est l'impact de la modélisation de la covariance sur l'inférence ? Reportons-nous à nos données *Beating the Blues*. Ici, on s'intéresse à l'efficacité du traitement par rapport au groupe contrôle (traitement usuel) : plus le score pour l'inventaire de dépression de Beck (`idb`) est faible, plus le traitement est efficace. Comme l'évolution chronologique du traitement est très différente, on spécifie un terme différent pour chaque tranche de deux mois en incluant `mois` comme variables catégorielle au modèle pour la moyenne. Outre `mois`, on inclut la variable contrôle `medicaments`, une variable binaire qui indique si la personne prend des médicaments.

Règle générale, les estimés des coefficients ne varient que très peu lorsqu'on met toutes les informations des 100 individus en commun. En revanche, comme l'échantillon est débalancé et que plusieurs personnes arrêtent avant 10 mois, la prise en compte de la corrélation intra-individu a un impact important sur les estimés. Le Tableau 12 montre les coefficients estimés avec leurs erreurs-types et les intervalles de confiance asymptotique symmétrique : si on ignore

Ainsi, la prise en compte de la corrélation entraîne une augmentation des coefficient pour *Beating the blues* (B), possiblement parce que plus de personnes traitées avec *Beating the blues* ont quitté l'étude prématurément et on ignore la structure individuelle. L'information fournie par un individu est partiellement redondante : on a 100 individus, mais 380 inventaires à cause des mesures répétées.

Si notre but est de comparer l'évolution de la maladie, alors cette approche ferait une meilleur utilisation des données que la différence entre la réponse au temps $t = 0$ mois et celle au temps $t = k$ mois pour $k \in \{2, \ldots, 8\}$ mois parce qu'on estimerait la trajectoire avec chacune des personnes même lorsque certaines courbes sont incomplètes. Ici, le modèle est paramétrisé en termes de contrastes : concentrons-nous sur l'effet du traitement pour une personne qui ne prend pas de médicaments (la réponse sera la même pour ces individus). La moyenne modélisée comprends une interaction entre l'indicateur de `traitement` et le mois, ce qui revient à spécifier une moyenne pour chacune de ces instances une fois l'effet de médicaments pris en compte. Pour une personne assignée à *Beating the blues*, la différence d'inventaire après 2 mois est β_1, tandis que celle pour une personne assignée au traitement usuel est $\beta_1 + \beta_6$. L'hypothèse nulle que la différence de différences de traitements entre usuel et *Beating the blues* est nulle revient donc $\mathcal{H}_0 : \beta_6 = 0$ (ou 2 `mois (U)` dans le tableau) et notre modèle est paramétrisé de telle sorte que cette information est directement disponible dans la sortie. Comparativement au modèle linéaire classique qui regroupe toutes les mesures de la variable réponse sans égard pour la structure de groupe, nos intervalles de confiance pour l'effet de traitement pour *Beating the blues* sont fortement décalés. Dans ce cas de

TABLE 12 – Coefficients (erreurs-types) et intervalles de confiance à 95% pour le modèle de régression linéaire classique (Modèle 1) et modèles avec corrélation intra-individu avec covariance non structurée (Modèle 2).

	Modèle 1	Modèle 2
ordonnée à l'origine	21.944 (1.624)	20.769 (1.850)
	[18.762, 25.127]	[17.143, 24.396]
2 mois (B)	-7.827 (2.094)	-7.827 (1.299)
	[-11.931, -3.722]	[-10.373, -5.281]
4 mois (B)	-10.530 (2.297)	-8.948 (1.523)
	[-15.031, -6.028]	[-11.933, -5.962]
6 mois (B)	-13.307 (2.475)	-9.601 (1.597)
	[-18.157, -8.456]	[-12.731, -6.470]
8 mois (B)	-13.703 (2.533)	-10.986 (1.640)
	[-18.667, -8.738]	[-14.199, -7.772]
traitement (usuel)	1.943 (2.163)	2.524 (2.207)
	[-2.296, 6.181]	[-1.802, 6.850]
médicaments (oui)	1.030 (1.155)	3.066 (1.922)
	[-1.234, 3.294]	[-0.700, 6.833]
2 mois (U)	3.132 (3.049)	3.334 (1.901)
	[-2.844, 9.108]	[-0.392, 7.060]
4 mois (U)	4.023 (3.289)	2.900 (2.194)
	[-2.424, 10.470]	[-1.401, 7.200]
6 mois (U)	5.376 (3.526)	1.954 (2.294)
	[-1.535, 12.286]	[-2.541, 6.449]
8 mois (U)	3.086 (3.654)	0.673 (2.367)
	[-4.076, 10.248]	[-3.966, 5.312]
AIC	2859.4	2644.7
BIC	2906.3	2746.3

figure, l'incertitude des estimés qui prennent en compte la corrélation est moindre. Cet exemple illustre que pour les cas de figure où les données sont débalancées (nombre de réponse différent par individu), la conclusion et les estimés sont fortement affectés. Plus généralement, on verra un changement important pour les erreurs-types des coefficients, mais dans une moindre mesure pour les estimés ponctuels de ces derniers.

5.2.3 Données groupées et modèle d'équicorrélation

Dans certains contextes, il n'y aura pas de structure logique pour les données : dans l'exemple précédent, le modèle AR(1) se base sur le nombre de mois séparant les questionnaires, mais qu'en serait-il si on interrogait les employés de différents départements au sein de l'école sur leur expérience durant la pandémie? Selon la pression de leur responsables d'unités, ils ou elles ont pu avoir une expérience très différente. Quelquefois, la dépendance proviendra non pas de mesures répétées, mais plutôt de regroupements logiques.

Un exemple nous est fourni par le sondage *Workplace Employee Relations Survey*, une étude du gouvernement britannique qui a été conduite en 1980, 1984, 1990, 1998, 2004 et 2011. L'étude est multi-niveaux : des entrevues détaillées sont conduites le ou la gestionnaire senior en charge des relations de travail, une personne syndiquée et non-syndiquée oeuvrant au sein de l'entreprise ainsi qu'un questionnaire administré à au plus 25 personnes dans chaque lieu de travail. Les milieux de travail sont sélectionnés par le biais d'un échantillon stratifié et certaines sont suivies d'une vague à l'autre ; voir la méthodologie pour plus de détails à ce sujet.

Ainsi, il serait logique de croire que les réponses des membres d'une même entreprise, qui partagent des politiques communes (convention collective, milieu de travail), etc. sont corrélées. Contrairement aux données longitudinales, le regroupement se fera par milieu de travail et il n'y a pas de relation logique entre les individus qui sont échangeables car l'ordre dans lequel ils apparaissent dans la base de données est arbitraire.

Dans ce contexte, une matrice de covariance non structurée ou une structure autorégressive est illogique, puisque chaque individu est différent d'une entreprise à l'autre. Un modèle simple qui peut être adéquat est le modèle d'équicorrélation, dans lequel la corrélation entre deux réponses au sein d'une unité a une corrélation constante de ρ.

Le modèle d'équicorrélation est d'ordinaire paramétrisé sous la forme

$$\boldsymbol{\Sigma}_i = \begin{pmatrix} \sigma^2+\tau & \tau & \cdots & \tau \\ \tau & \sigma^2+\tau & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \tau & \tau & \cdots & \sigma^2+\tau \end{pmatrix}.$$

et la corrélation intra-groupe pour deux observations est $\rho = \tau/(\sigma^2+\tau)$ avec $|\rho| < 1$; ce modèle est valide si et seulement si $1 + (n_{\max} - 1)\rho > 0$ en dimension $n_{\max}$ et cette inégalité se traduit par la contrainte $\tau > -\sigma^2/n_{\max}$. On recouvre le modèle avec données non corrélées et homoscédastiques si $\tau = 0$: ce constat servira lorsque nous considérerons les tests d'hypothèse.

On peut modéliser la covariance des unités intra-groupes dans le cas de données longitudinales ou groupées si le postulat d'indépendance ne tient pas la route.

— En prenant en compte la corrélation intra-unité, on obtient typiquement des estimés des erreurs-types plus élevés pour les coefficients.
— On suppose que les observations au sein d'une même unité sont corrélées, mais que les observations d'unités différentes sont indépendantes.

On a considéré trois principaux modèles de covariance :

— Le modèle de corrélation AR(1) suppose que la corrélation décroît exponentiellement dans le temps et ne dépend que de la distance (temporelle) entre deux observations au sein d'une même unité.
— Le modèle d'équicorrélation suppose que les données sont échangeable et que leur corrélation ρ est identique pour toutes les unités de groupe. Cette structure est plus logique pour les données groupées.
— Le modèle non structuré permet modéliser de manière générale la matrice de corrélation intra-unité. Le modèle a un nombre élévé de paramètres et n'est logique ou utile que si nous avons suffisamment de réplications à chaque temps donné pour estimer les paramètres.
— Des versions hétérogènes de ces trois modèles existent et impliquent des variances différents à chaque temps t, mais égales pour tous les individus.

5.3 Comparaisons de modèles

5.4 Hétéroscédasticité de groupe

6 Modèles linéaires mixtes

6.1 Comparaison de modèles

Cette brève discussion traite de méthodes de comparaisons de modèles selon différents scénarios d'intérêt. En particulier, puisque la plupart des logiciels promeuvent l'utilisation du critère de vraisemblance restreinte (REML) par défaut pour l'ajustement de modèles mixtes, il convient de porter une attention spéciale aux tests que l'on réalise.

En ajustant un modèle avec la méthode REML, on élimine la contribution de la moyenne de la vraisemblance. Cela permet d'obtenir des estimateurs des paramètres de variance $\boldsymbol{\psi}$ qui sont moins biaisés, mais cette fonction objective ne permet pas de comparer des modèles qui ont une matrice de modèle $\mathbf{X}$ différente.

On préfère si possible les tests d'hypothèse (rapport de vraisemblance) aux critères d'informations pour la sélection de modèle. Les tests d'hypothèse requièrent la comparaison de deux modèles **emboîtés** : c'est le cas si on peut obtenir en imposant des contraintes sur **un** des modèles le second.

— Test du rapport de vraisemblance, méthode du maximum de vraisemblance restreint (REML) : modèles emboîtés, même modèle pour la moyenne. Par exemple, tester si le modèle d'équicorrélation CS est une simplification adéquate du modèle non-structure UN.

— Test du rapport de vraisemblance, méthode du maximum de vraisemblance : modèles emboîtés (pas d'autre contrainte). Par exemple, dans un modèle linéaire avec erreurs autorégressives, tester si l'effet de la variable X_j est nul sachant le reste et si les erreurs sont indépendantes, soit $\mathcal{H}_0 : \beta_j = 0$, $\mathcal{H}_0 : \rho = 0$, ou encore $\mathcal{H}_0 : \beta_j = \rho = 0$.

Si on veut comparer des modèles non-emboîtés, on doit se rabattre sur la performance prédictive ou les critères d'information. Dans ce dernier cas, il faut que les deux modèles aient les même variables réponse. Si on utilise des fonctions de vraisemblance différentes, il faut aussi s'assurer que notre logiciel calcule les constantes de normalisation pour s'assurer que la comparaison soit valide.

— Par exemple, comparer un modèle linéaire avec erreurs autorégressives $\mathsf{AR}(1)$ versus un modèle avec un effet aléatoire sur la pente.

La seule comparaison de modèle emboîtés que je vous déconseille de faire à l'aide de tests d'hypothèse est celle dans lequel la comparaison entre les deux modèles implique de contraindre des paramètres positifs à zéro (éliminer un effet aléatoire revient à fixer sa variance à zéro). Ce faisant, on se trouve avec un cas où la valeur du paramètre est sur la bordure de l'espace des valeurs admissible. Ce cas limite donne une loi nulle différente de la loi χ^2 usuelle. Bien que la statistique de test soit calculable et correcte, l'approximation de la loi de référence est compliquée à dériver et de mauvaise qualité.

— Exemple de scénarios : regarder dans un modèle avec $\boldsymbol{b} \sim \mathsf{No}_2(\mathbf{0}_2, \boldsymbol{\Omega})$, où b_1 est une ordonnée à l'origine aléatoire et b_2 une pente aléatoire. Tester si la pente aléatoire est nécessaire revient à tester $\mathcal{H}_0 : \omega_{22} = 0$, et comme le paramètre de variance est positif, ce test n'est pas régulier.

7 Analyse de survie

A Compléments mathématiques

A.1 Population et échantillons

Ce qui différencie la statistique des autres sciences est la prise en compte de l'incertitude et de la notion d'aléatoire. Règle générale, on cherche à estimer une caractéristique d'une population définie à l'aide d'un échantillon (un sous-groupe de la population) de taille restreinte.

La **population d'intérêt** est une collection d'individus formant la matière première d'une étude statistique. Par exemple, pour l'Enquête sur la population active (EPA) de Statistique Canada, « la population cible comprend la population canadienne civile non institutionnalisée de 15 ans et plus ». Même si on faisait un recensement et qu'on interrogeait tous les membres de la population cible, la caractéristique d'intérêt peut varier selon le moment de la collecte ; une personne peut trouver un emploi, quitter le marché du travail ou encore se retrouver au chômage. Cela explique la variabilité intrinsèque.

En général, on se base sur un **échantillon** pour obtenir de l'information. L'**inférence statistique**

vise à tirer des conclusions, pour toute la population, en utilisant seulement l'information contenue dans l'échantillon et en tenant compte des sources de variabilité. Le sondeur George Gallup (traduction libre) a fait cette merveilleuse analogie entre échantillon et population :

> «Il n'est pas nécessaire de manger un bol complet de soupe pour savoir si elle est trop salé ; pour autant qu'elle ait été bien brassée, une cuillère suffit.»

Un **échantillon** est un sous-groupe d'individus tiré aléatoirement de la population. La création de plans d'enquête est un sujet complexe et des cours entiers d'échantillonnage y sont consacrés. Même si on ne collectera pas de données, il convient de noter la condition essentielle pour pouvoir tirer des conclusions fiables à partir d'un échantillon : ce dernier doit être représentatif de la population étudiée, en ce sens que sa composition doit être similaire à celle de la population. On doit ainsi éviter les biais de sélection, notamment les échantillons de commodité qui consistent en une sélection d'amis et de connaissances.

Si notre échantillon est **aléatoire**, notre mesure d'une caractéristique d'intérêt le sera également et la conclusion de notre procédure de test variera d'un échantillon à l'autre. Plus la taille de ce dernier est grande, plus on obtiendra une mesure précise de la quantité d'intérêt. L'exemple suivant illustre pourquoi le choix de l'échantillon est important.

Exemple A.1. Désireuse de prédire le résultat de l'élection présidentielle américaine de 1936, la revue *Literary Digest* a sondé 10 millions d'électeurs par la poste, dont 2.4 millions ont répondu au sondage en donnant une nette avance au candidat républicain Alf Landon (57%) face au président sortant Franklin D. Roosevelt (43%). Ce dernier a néanmoins remporté l'élection avec 62% des suffrages, une erreur de prédiction de 19%. Le plan d'échantillonnage avait été conçu en utilisant des bottins téléphoniques, des enregistrements d'automobiles et des listes de membres de clubs privés, etc. : la non-réponse différentielle et un échantillon biaisé vers les classes supérieures sont en grande partie responsable de cette erreur.

Gallup avait de son côté correctement prédit la victoire de Roosevelt en utilisant un échantillon aléatoire de (seulement) 50 000 électeurs. L'histoire complète (en anglais).

A.2 Variables aléatoires

Suppsons qu'on cherche à décrire le comportement d'un phénomène aléatoire. Pour ce faire, on cherche à décrire l'ensemble des valeurs possibles et leur probabilité/fréquence relative au sein de la population : ces dernières sont encodées dans la loi de la variable aléatoire.

On fera la distinction entre deux cas de figure : quand le phénomène prend des valeurs finies, comme par exemple un événement binaire (achat/non-achat d'un produit) ou un continuum de valeurs (par exemple, le prix d'un item). On dénote les variables aléatoires par des lettres majuscules : par exemple, $Y \sim \mathrm{No}(\mu, \sigma^2)$ indique que Y suit une loi normale de paramètres μ et σ, qui représentent respectivement l'espérance et l'écart-type de Y.

La fonction de répartition $F(y)$ donne la probabilité cumulative qu'un événement n'excède pas une variable donnée, $F(y) = \Pr(Y \leq y)$.

Si la variable Y prend des valeurs discrètes, alors on utilise la fonction de masse $f(y) = \Pr(Y = y)$ qui donne la probabilité pour chacune des valeurs de y. Si la variable Y est continue, aucune valeur numérique de y n'a de probabilité non-nulle ; la densité sert à estimer la probabilité que la variable Y appartienne à un ensemble B, via $\Pr(Y \in B) = \int_B f(y)\mathrm{d}y$; la fonction de répartition est ainsi $F(y) = \int_{-\infty}^{y} f(x)\mathrm{d}x$.

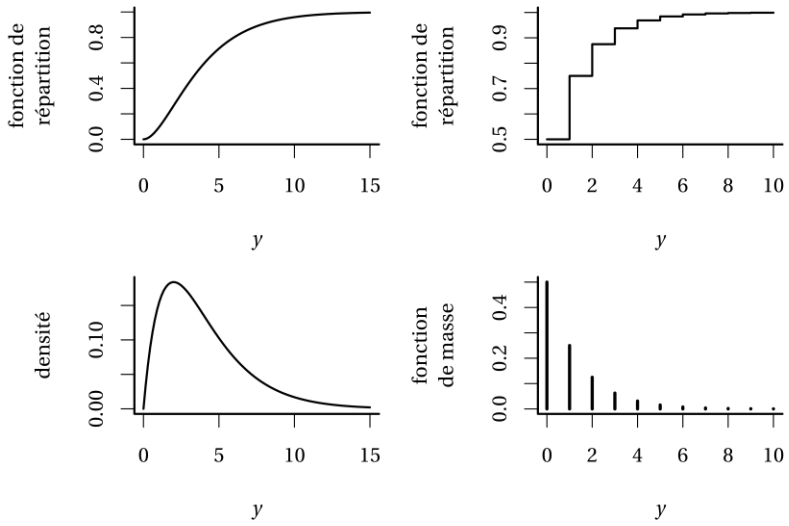

FIGURE 23 – Fonctions de répartition (panneau supérieur) et fonctions de densité et de masse (panneau inférieur) pour une loi continue (gauche) et discrète (droite).

A.2.1 Moments

Un premier cours de statistique débute souvent par la présentation de statistiques descriptives comme la moyenne et l'écart-type. Ce sont des estimateurs des moments (centrés), qui caractérisent la loi du phénomène d'intérêt. Dans le cas de la loi normale unidimensionnelle, qui a deux paramètres, l'espérance et la variance caractérisent complètement le modèle.

Soit Y une variable aléatoire de fonction de densité (ou de masse) $f(x)$. Cette fonction est non-négative et satisfait $\int_{\mathbb{R}} f(x)\mathrm{d}x = 1$: elle décrit la probabilité d'obtenir un résultat dans un ensemble donné des réels $\mathbb{R}$.

On définit l'espérance d'une variable aléatoire Y comme

$$\mathsf{E}(Y) = \int_{\mathbb{R}} x f(x)\mathrm{d}x.$$

L'espérance est la « moyenne théorique» : dans le cas discret, $\mu = \mathsf{E}(Y) = \sum_{x \in \mathscr{X}} x\mathsf{Pr}(X = x)$, où $\mathscr{X}$ représente le support de la loi, à savoir les valeurs qui ont une probabilité non-nulle. Plus généralement, l'espérance d'une fonction $g(x)$ pour une variable aléatoire Y est simplement l'intégrale de $g(x)$ pondérée par la densité $f(x)$. De même, si l'intégrale est convergente, la variance est

$$\mathsf{Va}(Y) = \mathsf{E}\{Y - \mathsf{E}(Y)\}^2 \equiv \int_{\mathbb{R}} (x-\mu)^2 f(x) \mathrm{d}x.$$

Un estimateur $\hat{\theta}$ pour un paramètre θ est sans biais si son biais $\mathsf{biais}(\hat{\theta}) = \mathsf{E}(\hat{\theta}) - \theta$ est nul. L'estimateur sans biais de l'espérance de Y est $\overline{Y}_n = n^{-1}\sum_{i=1}^n Y_i$ et celui de la variance $S_n = (n-1)^{-1}\sum_{i=1}^n (Y_i - \overline{Y})^2$. Un estimateur sans biais est souhaitable, mais pas toujours optimal. Quelquefois, il n'existe pas d'estimateur non-biaisé!

Souvent, on cherche à balancer le biais et la variance : rappelez-vous qu'un estimateur est une variable aléatoire (étant une fonction de variables aléatoires) et qu'il est lui-même variable : même s'il est sans biais, la valeur numérique obtenue fluctuera d'un échantillon à l'autre. On peut chercher un estimateur qui minimise l'erreur moyenne quadratique,

$$\mathsf{EMQ}(\hat{\theta}) = \mathsf{E}\{(\hat{\theta} - \theta)^2\} = \mathsf{Va}(\hat{\theta}) + \{\mathsf{E}(\hat{\theta})\}^2.$$

C'est donc un compromis entre le carré du biais et la variance de l'estimateur. La plupart des estimateurs que nous considérerons dans le cadre du cours sont des estimateurs du maximum de vraisemblance. Ces derniers sont asymptotiquement efficaces, c'est-à-dire qu'ils minimisent l'erreur moyenne quadratique parmi tous les estimateurs possibles quand la taille de l'échantillon est suffisamment grande. Ils ont également d'autre propriétés qui les rendent attractifs comme choix par défaut pour l'estimation.

A.2.2 Distributions

Plusieurs lois aléatoires décrivent des phénomènes physiques simples et ont donc une justification empirique ; on revisite les distributions les plus fréquemment couvertes.

Exemple A.2 (Loi de Bernoulli)**.** On considère un phénomène binaire, comme le lancer d'une pièce de monnaie (pile/face). De manière générale, on associe les deux possibilités à succès/échec et on suppose que la probabilité de succès est π. Par convention, on représente les échecs (non) par des zéros et les réussites (oui) par des uns. Donc, si la variable Y vaut 0 ou 1, alors $\mathsf{Pr}(Y = 1) = \pi$ et $\mathsf{Pr}(Y = 0) = 1 - \pi$ (complémentaire). La fonction de masse de la loi Bernoulli s'écrit de façon plus compacte

$$\mathsf{Pr}(Y = y) = \pi^y (1-\pi)^{1-y}, \quad y = 0, 1.$$

Un calcul rapide montre que $\mathsf{E}(Y) = \pi$ et $\mathsf{Va}(Y) = \pi(1-\pi)$. Voici quelques exemples de questions de recherches comprenant une variable réponse binaire :

— est-ce qu'un client potentiel a répondu favorablement à une offre promotionnelle ?
— est-ce qu'un client est satisfait du service après-vente ?
— est-ce qu'une firme va faire faillite au cours des trois prochaines années ?
— est-ce qu'un participant à une étude réussit une tâche ?

Exemple A.3 (Loi binomiale)**.** Si les données représentent la somme d'événements Bernoulli indépendants, la loi du nombre de réussites Y pour un nombre d'essais donné m est dite binomiale, dénotée $\mathsf{Bin}(m,\pi)$; sa fonction de masse est

$$\mathsf{Pr}(Y=y)=\binom{m}{y}\pi^y(1-\pi)^{1-y},\quad y=0,1.$$

La vraisemblance pour un échantillon de la loi binomiale est (à constante de normalisation près qui ne dépend pas de π) la même que pour un échantillon aléatoire de m variables Bernoulli indépendantes. L'espérance d'une variable binomiale est $\mathsf{E}(Y)=m\pi$ et la variance $\mathsf{Va}(Y)=m\pi(1-\pi)$.

On peut ainsi considérer le nombre de personnes qui ont obtenu leur permis de conduire parmi m candidat(e)s ou le nombre de clients sur m qui ont passé une commande de plus de 10$ dans un magasin.

Plus généralement, on peut considérer des variables de dénombrement qui prennent des valeurs entières. Parmi les exemples de questions de recherches comprenant une variable réponse de dénombrement :

— le nombre de réclamations faites par un client d'une compagnie d'assurance au cours d'une année.
— le nombre d'achats effectués par un client depuis un mois.
— le nombre de tâches réussies par un participant lors d'une étude.

Exemple A.4 (Loi géométrique)**.** La loi géométrique décrit le comportement du nombre d'essais Bernoulli de probabilité de succès π nécessaires avant l'obtention d'un premier succès. La fonction de masse de $Y\sim\mathsf{Geo}(\pi)$ est

$$\mathsf{Pr}(Y=y)=\pi(1-\pi)^{y-1},\quad y=1,2,\ldots$$

Par exemple, on pourrait modéliser le nombre de visites d'une maison en vente avant une première offre d'achat à l'aide d'une variable géométrique.

Exemple A.5 (Loi de Poisson)**.** Si la probabilité d'un événement Bernoulli est petite (succès rare) dans le sens où $m\pi\to\lambda$ quand le nombre d'essais m augmente, alors le nombre de succès suit une loi de Poisson de fonction de masse

$$\mathsf{Pr}(Y=y)=\frac{\exp(-\lambda)\lambda^y}{\Gamma(y+1)},\quad y=0,1,2,\ldots$$

où $\Gamma(\cdot)$ dénote la fonction gamma. Le paramètre λ de la loi de Poisson représente à la fois l'espérance et la variance de la variable, c'est-à-dire que $\mathsf{E}(Y) = \mathsf{Va}(Y) = \lambda$.

Exemple A.6 (Loi binomiale négative)**.** On considère une série d'essais Bernoulli de probabilité de succès π jusqu'à l'obtention de m succès. Soit Y, le nombre d'échecs : puisque la dernière réalisation doit forcément être un succès, mais que l'ordre des succès/échecs précédents n'importe pas, la fonction de masse est

$$\mathsf{Pr}(Y = y) = \binom{m-1+y}{y} \pi^m (1-\pi)^y.$$

La loi binomiale négative apparaît également si on considère la loi non-conditionnelle du modèle hiérarchique gamma-Poisson, dans lequel on suppose que le paramètre de la moyenne de la loi Poisson est aussi aléatoire, c'est-à-dire $Y \mid \Lambda = \lambda \sim \mathsf{Po}(\lambda)$ et Λ suit une loi gamma de paramètre de forme r et de paramètre d'échelle θ, dont la densité est

$$f(x) = \theta^{-r} x^{r-1} \exp(-x/\theta)/\Gamma(r).$$

Le nombre d'événements suit alors une loi binomiale négative.

La paramétrisation la plus courante pour la modélisation est légèrement différente : on utilise la fonction de masse est

$$\mathsf{Pr}(Y = y) = \frac{\Gamma(y+r)}{\Gamma(y+1)\Gamma(r)} \left(\frac{r}{r+\mu}\right)^r \left(\frac{\mu}{r+\mu}\right)^y, y = 0, 1, \ldots, \mu, r > 0,$$

où Γ dénote la fonction gamma. À noter que le paramètre $r > 0$ n'est plus nécessairement entier. La moyenne théorique et la variance sont $\mathsf{E}(Y) = \mu$ et $\mathsf{Va}(Y) = \mu + k\mu^2$, où $k = 1/r$. La variance d'une variable binomiale négative est *supérieure* à sa moyenne et le modèle est utilisé comme alternative à la loi de Poisson pour modéliser la surdispersion.

A.2.3 Diagrammes quantiles-quantiles

Si on ajuste un modèle à des données, il convient de vérifier la qualité de l'ajustement et l'adéquation du modèle, par exemple graphiquement. Le diagramme quantile-quantile sert à vérifier l'adéquation du modèle et découle du constat suivant : si Y est une variable aléatoire continue et F sa fonction de répartition, alors l'application $F(Y) \sim \mathsf{U}(0,1)$. De la même façon, appliquer la fonction quantile à une variable uniforme permet de simuler de la loi F, et donc $F^{-1}(U)$. Supposons un échantillon de taille n de variables uniformes. On peut démontrer que les statistiques d'ordre $U_{(1)} \leq \cdots \leq U_{(n)}$ ont une loi marginale Beta : et $U_{(k)} \sim \mathsf{Beta}(k, n+1-k)$ d'espérance $k/(n+1)$.

Les paramètres de la loi F sont inconnus, mais on peut obtenir un estimateur $\widehat{F}$ et appliquer la transformation inverse pour obtenir une variable approximativement uniforme. Un diagramme quantile-quantile représente les données en fonction des moments des statistiques d'ordre transformées

— sur l'axe des abscisses, les quantiles théoriques $\widehat{F}^{-1}\{\text{rang}(Y_i)/(n+1)\}$
— sur l'axe des ordonnées, les quantiles empiriques Y_i

Si le modèle est adéquat, les valeurs ordonnées devraient suivre une droite de pente unitaire qui passe par l'origine. L'oeil humain a de la difficulté à juger de la qualité de l'adéquation en regardant une droite, aussi est-il préférable de soustraire cette pente pour faciliter l'interprétation (une méthode proposée par Tukey, mais faire attention à l'échelle !) Le diagramme des différences moyennes prend les positions du diagramme quantile-quantile, (x_i, y_i), et représente graphiquement la moyenne des coordonnées sur l'axe des abcisses versus la différence$(\{x_i + y_i\}/2, y_i - x_i)$ sur l'axe des ordonnées. Les données de la Figure 24 montrent ces deux représentations sur des mêmes données simulées d'une loi normale standard.

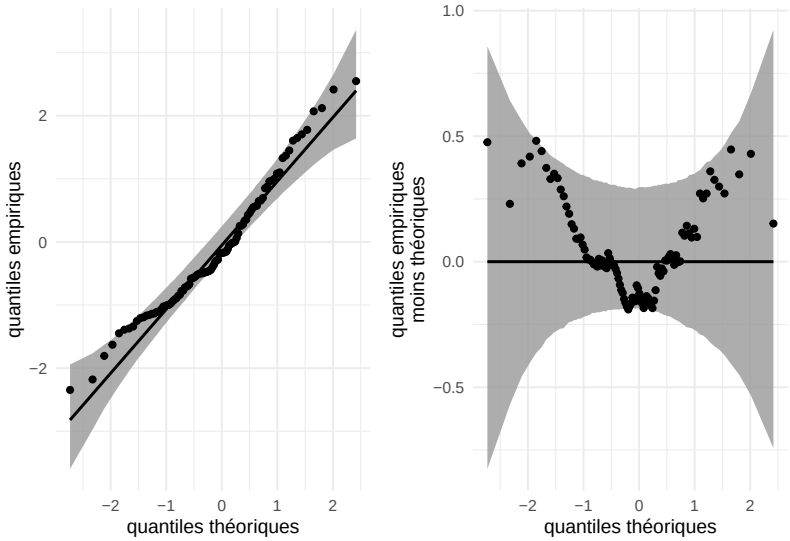

FIGURE 24 – Diagramme quantile-quantile normal (gauche) et représentation de Tukey du même diagramme (en soustrayant la traîne)

Même si on connaissait exactement la loi aléatoire des données, la variabilité intrinsèque à l'échantillon fait en sorte que des déviations qui semblent significatives et anormales à l'oeil de l'analyste sont en fait compatibles avec le modèle : un simple estimé ponctuel sans mesure d'incertitude ne permet donc pas facilement de voir ce qui est plausible ou pas. On va donc idéalement ajouter un intervalle de confiance (approximatif) ponctuel ou conjoint au diagramme.

Pour obtenir l'intervalle de confiance approximatif, la méthode la plus simple est par simulation (autoamorçage paramétrique), en répétant B fois les étapes suivantes

1. simuler un échantillon $\{Y_i^{(b)}\}(i = 1, \ldots, n)$ du modèle $\widehat{F}$
2. estimer les paramètres du modèle F pour obtenir $\widehat{F}_{(b)}$

3. calculer et stocker les positions $\widehat{F}^{-1}_{(b)}\{i/(n+1)\}$.

Le résultat de cette opération sera une matrice $n \times B$ de données simulées ; on obtient un intervalle de confiance symmétrique en conservant le quantile $\alpha/2$ et $1-\alpha/2$ de chaque ligne. Le nombre de simulation B devrait être large (typiquement 999 ou davantage) et être choisi de manière à ce que $B/alpha$ soit un entier.

Pour l'intervalle de confiance ponctuel, chaque valeur représente une statistique et donc individuellement, la probabilité qu'une statistique d'ordre sorte de l'intervalle de confiance est α. En revanche, les statistiques d'ordres ne sont pas indépendantes et sont qui est plus ordonnées, ce qui fait qu'un point hors de l'intervalle risque de n'être pas isolé. [Il est aussi possible d'obtenir par autoamorçage un intervalle de confiance (approximatif) conjoint, pour lequel une valeur sort de l'intervalle $100(1-\alpha)\%$ du temps ; voir à ce sujet mes notes de cours Section 4.4.3 (en anglais). Les intervalles présentés dans la Figure 24 sont ponctuels. La variabilité des statistiques d'ordre uniformes est plus grande à mesure qu'on s'éloigne de $1/2$, mais celles des variables transformées dépend de F.

A.3 Loi des grands nombres

Un estimateur est dit **convergent** si la valeur obtenue à mesure que la taille de l'échantillon augmente s'approche de la vraie valeur que l'on cherche à estimer. Mathématiquement parlant, un estimateur est dit convergent s'il converge en probabilité, ou $\hat{\theta} \xrightarrow{\mathsf{Pr}} \theta$: en langage commun, la probabilité que la différence entre $\hat{\theta}$ et θ diffèrent est négligeable quand n est grand.

La condition *a minima* pour le choix d'un estimateur est donc la convergence : plus on récolte d'information, plus notre estimateur devrait s'approcher de la valeur qu'on tente d'estimer.

La loi des grands nombres établit que la moyenne empirique de n observations indépendantes de même espérance, $\overline{Y}_n$, tend vers l'espérance commune des variables μ, où $\overline{Y}_n \to \mu$. En gros, ce résultat nous dit que l'on réussit à approximer de mieux en mieux la quantité d'intérêt quand la taille de l'échantillon (et donc la quantité d'information disponible sur le paramètre) augmente. La loi des grands nombres est très utile dans les expériences Monte Carlo : on peut ainsi approximer par simulation la moyenne d'une fonction $g(x)$ de variables aléatoires en simulant de façon répétée des variables Y indépendantes et identiquement distribuées et en prenant la moyenne empirique $n^{-1}\sum_{i=1}^{n} g(Y_i)$.

Si la loi des grands nombres nous renseigne sur le comportement limite ponctuel, il ne nous donne aucune information sur la variabilité de notre estimé de la moyenne et la vitesse à laquelle on s'approche de la vraie valeur du paramètre.

A.4 Théorème central limite

Le théorème central limite dit que, pour un échantillon aléatoire de taille n dont les observations sont indépendantes et tirées d'une loi quelconque d'espérance μ et de variance finie σ^2, alors la moyenne empirique tend non seulement vers μ, mais à une vitesse précise :

— l'estimateur $\overline{Y}$ sera centré autour de μ,
— l'erreur-type sera de $\sigma/\sqrt{n}$; le taux de convergence est donc de $\sqrt{n}$. Ainsi, pour un échantillon de taille 100, l'erreur-type de la moyenne empirique sera 10 fois moindre que l'écart-type de la variable aléatoire sous-jacente.
— la loi approximative de la moyenne $\overline{Y}$ sera normale.

Mathématiquement, le théorème central limite dicte que $\sqrt{n}(\overline{Y}-\mu) \xrightarrow{\text{d}} \text{No}(0,\sigma^2)$. Si n est grand (typiquement supérieur à 30, mais cette règle dépend de la loi sous-jacente de Y), alors $\overline{Y} \dot{\sim} \text{No}(\mu,\sigma^2/n)$.

Comment interpréter ce résultat ? On considère comme exemple le temps de trajet moyen de trains à haute vitesse AVE entre Madrid et Barcelone opérés par la Renfe.

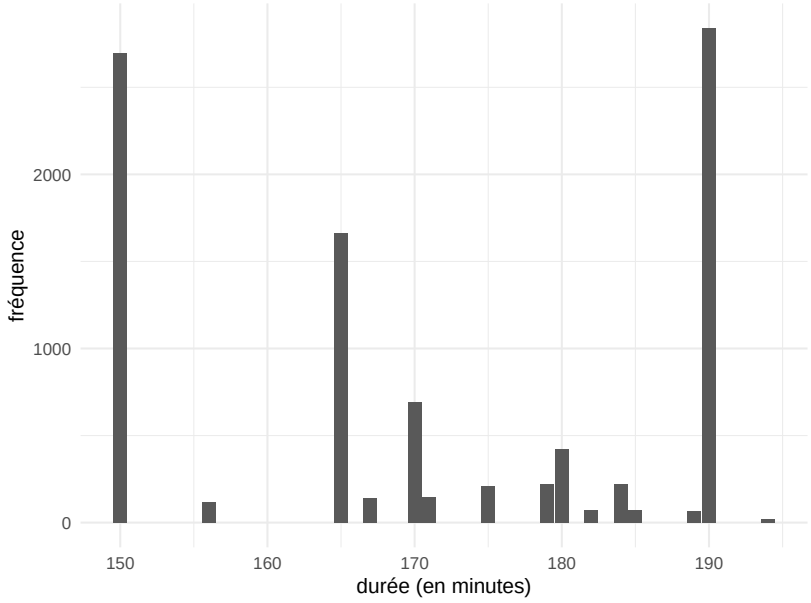

FIGURE 25 – Distribution empirique des temps de trajet en trains à grande vitesse.

Une analyse exploratoire indique que la durée du trajet de la base de données est celle affichée sur le billet (et non le temps réel du parcours). Ainsi, il n'y a ainsi que 15 valeurs possibles. Le temps affiché moyen pour le parcours, estimé sur la base de 9603 observations, est de 170 minutes et 41 secondes. La Figure (voir 25) montre la distribution empirique des données.

Considérons maintenant des échantillons de taille $n = 10$. Dans notre premier échantillon aléatoire, la durée moyenne affichée est 170.9 minutes, elle est de 164.5 minutes dans le deuxième, de 172.3 dans le troisième, et ainsi de suite.

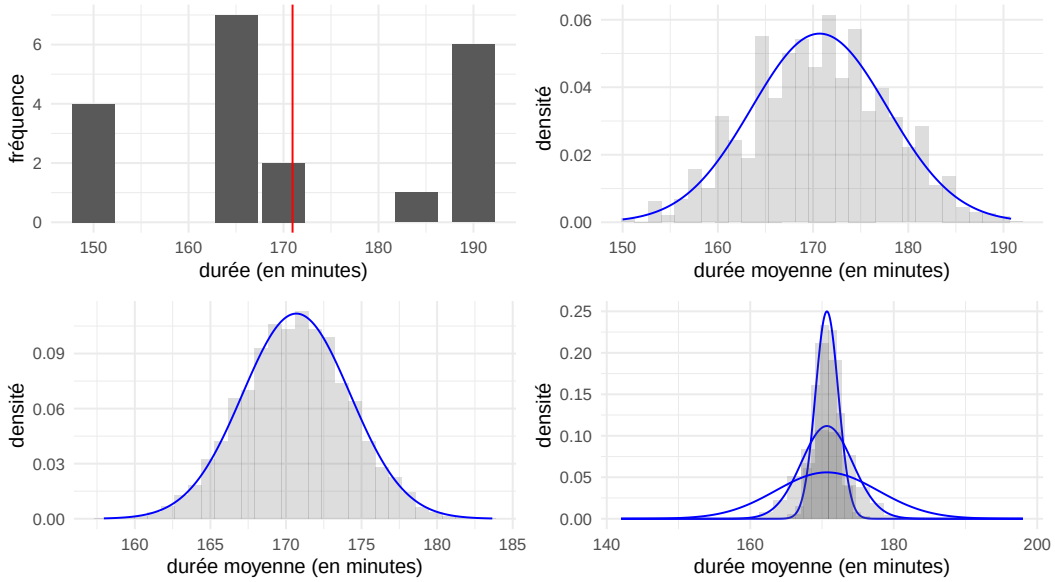

FIGURE 26 – Représentation graphique du théorème central limite : échantillon aléatoire de 20 observations avec leur moyenne empirique (trait vertical rouge) (en haut à gauche). Les trois autres panneaux montrent les histogrammes des moyennes empiriques d'échantillons répétés de taille 5 (en haut à droite), 20 (en bas à gauche) et les histogrammes pour $n = 5, 20, 100$ (en bas à droite) avec courbe de densité de l'approximation normale fournie par le théorème central limite.

Supposons qu'on tire $B = 1000$ échantillons différents, chacun de taille $n = 5$, de notre ensemble, et qu'on calcule la moyenne de chacun d'entre eux. Le graphique supérieur droit 26 montre un de ces 1000 échantillons aléatoire de taille $n = 20$ tiré de notre base de données. Les autres graphiques de la Figure 26 illustrent l'effet de l'augmentation de la taille de l'échantillon : si l'approximation normale est approximative avec $n = 5$, la distribution des moyennes est virtuellement identique à partir de $n = 20$. Plus la moyenne est calculée à partir d'un grand échantillon (c'est-à-dire, plus n augmente), plus la qualité de l'approximation normale est meilleure et plus la courbe se concentre autour de la vraie moyenne ; malgré le fait que nos données sont discrètes, la distribution des moyennes est approximativement normale.

On a considéré une seule loi aléatoire inspirée de l'exemple, mais vous pouvez vous amuser à regarder l'effet de la distribution sous-jacent et de la taille de l'échantillon nécessaire pour que l'effet du théorème central limite prenne effet : il suffit pour cela de simulant des observations d'une loi quelconque de variance finie, en utilisant par exemple cette applette.

Les statistiques de test qui découlent d'une moyenne centrée-réduite (ou d'une quantité équivalente pour laquelle un théorème central limite s'applique) ont souvent une loi nulle standard normale, du moins asymptotiquement (quand n est grand, typiquement $n > 30$ est suffisant). C'est ce qui garantie la validité de notre inférence !

B Dérivations mathématiques

Cette section regroupe les dérivations mathématiques optionnelles qui sont fournies par souci de complétude.

B.1 Dérivation de l'estimateur des moindres carrés ordinaires

L'estimateur des moindres carrés ordinaires résoud le problème d'optimisation non-contraint

$$\widehat{\boldsymbol{\beta}} = \min_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}).$$

On peut calculer la dérivée première par rapport à $\boldsymbol{\beta}$, égaler à zéro et isoler le maximum pour obtenir une formule explicite pour $\widehat{\boldsymbol{\beta}}$,

$$\begin{aligned}
\mathbf{0}_n &= \frac{\partial}{\partial \boldsymbol{\beta}} (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) \\
&= \frac{\partial (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \frac{\partial (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})}{\partial (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})} \\
&= \mathbf{X}^\top (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})
\end{aligned}$$

en utilisant la règle de dérivation en chaîne ; on peut ainsi distribuer les termes pour obtenir l'*équation normale*

$$\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} = \mathbf{X}^\top \boldsymbol{y}.$$

Si $\mathbf{X}$ est une matrice de rang p, alors la forme quadratique $\mathbf{X}^\top \mathbf{X}$ est inversible et l'unique solution du problème d'optimisation est celle fournie dans l'équation (2).

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{Y}.$$

B.2 Dérivation du coefficient de détermination

Le point de départ de cette dérivation est la décomposition orthogonale entre le vecteur de valeurs ajustées et de résidus, $\boldsymbol{y} = \widehat{\boldsymbol{y}} + \boldsymbol{e}$. Pourvu que la matrice du modèle $\mathbf{X}$ contienne l'équivalent de

l'ordonnée à l'origine $\mathbf{1}_n \in \mathscr{S}(\mathbf{X})$, alors la moyenne des résidus ordinaires est nulle, $\overline{\boldsymbol{e}} = 0$ et il en découle que la moyenne empirique des réponses est égale à la moyenne empirique des valeurs ajustées. Puisque $n^{-1}\sum_{i=1}^n \widehat{y}_i = n^{-1}\sum_{i=1}^n (y_i - e_i) = \overline{y}$,

$$\begin{aligned}
\widehat{\mathsf{Cor}}\big(\widehat{\boldsymbol{y}}, \boldsymbol{y}\big) &= \frac{(\boldsymbol{y} - \overline{y}\mathbf{1}_n)^\top (\widehat{\boldsymbol{y}} - \overline{y}\mathbf{1}_n)}{\|\boldsymbol{y} - \overline{y}\mathbf{1}_n\|\|\widehat{\boldsymbol{y}} - \overline{y}\mathbf{1}_n\|} \\
&= \frac{(\widehat{\boldsymbol{y}} - \overline{y}\mathbf{1}_n)^\top (\widehat{\boldsymbol{y}} - \overline{y}\mathbf{1}_n) + \boldsymbol{e}^\top (\widehat{\boldsymbol{y}} - \overline{y}\mathbf{1}_n)}{\|\boldsymbol{y} - \overline{y}\mathbf{1}_n\|\|\widehat{\boldsymbol{y}} - \overline{y}\mathbf{1}_n\|} \\
&= \frac{\|\widehat{\boldsymbol{y}} - \overline{y}\mathbf{1}_n\|}{\|\boldsymbol{y} - \overline{y}\mathbf{1}_n\|} \\
&= \frac{\|\boldsymbol{y} - \overline{y}\mathbf{1}_n\| - \|\boldsymbol{e}\|}{\|\boldsymbol{y} - \overline{y}\mathbf{1}_n\|} \\
&= \sqrt{\frac{\mathsf{SC}_c - \mathsf{SC}_e}{\mathsf{SC}_c}} = \mathrm{R}.
\end{aligned}$$

Cela justifie la proposition de la Section 2.5 voulant que le carré de la corrélation entre les valeurs ajustées et la variable réponse est égal à R^2.

B.2.1 Optimisation pour les modèles linéaires généralisés

Il n'existe règle générale pas de solution explicite pour l'estimateur du maximum de vraisemblance $\widehat{\boldsymbol{\beta}}$ pour les modèles linéaires généralisés ; l'équation du score étant typiquement nonlinéaire en $\boldsymbol{\beta}$, on doit obtenir les estimateurs du maximum de vraisemblance par le biais d'algorithmes d'optimisation numérique.

On dérive la log vraisemblance $\ell = \sum_{i=1}^n \log f(y_i; \theta, \phi)$ par rapport à $\boldsymbol{\beta}$. Pour simplifier, on considère chaque terme et coefficient à tour de rôle. Par la règle du produit,

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{\partial \eta_i}{\partial \beta_j}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \ell_i}{\partial \theta_i}$$

et les dérivations antérieures nous donnent $\partial \ell_i / \partial \theta_i = (y_j - \mu_i)/a_i(\phi)$ et $\partial \mu_i / \partial \theta_i = b''(\theta_i) = \mathsf{Va}(Y_i)/a_i(\phi)$. La dérivée du prédicteur linéaire est $\partial \eta_i / \partial \beta_j = \mathrm{X}_{ij}$. La seule dérivée partielle manquante, $\partial \mu_i / \partial \eta_i$, dépend de la fonction de liaison puisque $\eta_i = g(\mu_i)$; cette dérivée vaut un pour la fonction de liaison canonique.

Soit l'équation de score et la fonction d'information

$$U(\boldsymbol{\beta}) = \frac{\partial \ell}{\partial \boldsymbol{\beta}}, \qquad j(\boldsymbol{\beta}) = -\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top},$$

qui sont le gradient et la hessienne de la fonction de log vraisemblance ; en prenant la somme de tous les termes individuels, le je élément du vecteur de score $\boldsymbol{U}$ est

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)\mathrm{X}_{ij}}{g'(\mu_i)V(\mu_i)a_i(\phi)}, \qquad j = 0, \ldots, p.$$

Puisque le maximum de vraisemblance $\widehat{\boldsymbol{\beta}}$ résoud l'équation du score $U(\widehat{\boldsymbol{\beta}}) = \mathbf{0}_{p+1}$, on peut construire un algorithme de Newton–Raphson pour obtenir ce dernier. Si on fait un développement limité de Taylor du score $U(\widehat{\boldsymbol{\beta}})$ autour de $\boldsymbol{\beta}$,

$$\mathbf{0}_{p+1} = U(\widehat{\boldsymbol{\beta}}) \doteq U(\beta) - j(\boldsymbol{\beta})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}).$$

Pour autant que la matrice $(p+1)$ $j(\boldsymbol{\beta}^{(t)})$ soit invertible, on peut utiliser la procédure itérative suivante : en partant d'une valeur initiale $\boldsymbol{\beta}^{(0)}$, on calcule à l'étape $t+1$

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + j^{-1}(\boldsymbol{\beta}^{(t)})U(\boldsymbol{\beta}^{(t)}).$$

et on itère la formule jusqu'à convergence. La plupart des logiciels implémente une version de cet algorithme, dans lequel le négatif de la hessienne $j(\boldsymbol{\beta})$ est parfois remplacée par son espérance, $i(\boldsymbol{\beta})$: l'algorithme résultant est dénommé score de Fisher. Pour les modèles linéaire généralisés, ces récursions peuvent être effectuées à l'aide d'une variante des moindres carrés connue sous le nom de moindres carrés itérativement pondérés.

R

Je suis partisan de la philosophie Tinyverse ; voir l'introduction sur les fonctionnalités de **R** avec des ressources et des dépendances minimales.